Alexandra I. Cristea · Erin Walker ·
Yu Lu · Olga C. Santos · Seiji Isotani (Eds.)

LNAI 15881

Artificial Intelligence in Education

26th International Conference, AIED 2025
Palermo, Italy, July 22–26, 2025
Proceedings, Part V

5 Part V

AIED PALERMO 2025

Springer

MOREMEDIA

Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence 15881

Founding Editor

Jörg Siekmann

Series Editors

Randy Goebel, *University of Alberta, Edmonton, Canada*
Wolfgang Wahlster, *DFKI, Berlin, Germany*
Zhi-Hua Zhou, *Nanjing University, Nanjing, China*

The series Lecture Notes in Artificial Intelligence (LNAI) was established in 1988 as a topical subseries of LNCS devoted to artificial intelligence.

The series publishes state-of-the-art research results at a high level. As with the LNCS mother series, the mission of the series is to serve the international R & D community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings.

Alexandra I. Cristea · Erin Walker · Yu Lu ·
Olga C. Santos · Seiji Isotani
Editors

Artificial Intelligence in Education

26th International Conference, AIED 2025
Palermo, Italy, July 22–26, 2025
Proceedings, Part V

Springer

Editors
Alexandra I. Cristea
Durham University
Durham, UK

Yu Lu
Beijing Normal University
Beijing, China

Seiji Isotani
University of São Paulo
São Carlos, Brazil

Erin Walker
University of Pittsburgh
Pittsburgh, PA, USA

Olga C. Santos
UNED
Madrid, Spain

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-031-98461-7 ISBN 978-3-031-98462-4 (eBook)
https://doi.org/10.1007/978-3-031-98462-4

LNCS Sublibrary: SL7 – Artificial Intelligence

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

Welcome to the 26th International Conference on Artificial Intelligence in Education, taking place from July 22–26, 2025 in the vibrant city of Palermo, Italy. AIED 2025 is ranked A in ICORE[1], in the top 14.92% of all 784 ranked venues, and clearly growing and attracting more submissions, this year with over 800 initial submissions. Thus, as we celebrate over a quarter of a century of advancing the science and engineering of intelligent human-technology ecosystems that support learning, this edition marked a significant evolution in our conference structure, with the introduction of a dedicated track-based system. This new format was designed to capture the growth of the area, as well as the dynamic and multifaceted nature of research in educational AI, reflecting the rapid pace and expanding impact of the field.

The theme for AIED 2025—"AI as a Catalyst for Inclusive, Personalised, and Ethical Education: Empowering Teachers and Students for an Equitable Future"—embodies our vision for a transformative future in education. Artificial Intelligence is progressing at an unprecedented rate, with generative AI and related technologies spearheading this transformative shift. As these innovations reshape both teaching and learning processes, they unlock novel opportunities for creating personalised and dynamic learning experiences. Even as traditional AI techniques continue to evolve and enrich our understanding, true change lies in leveraging these advancements to provide educators with powerful tools for deeper insights, tailored instruction, and increased student engagement.

At the heart of this transformation is the commitment to equity and ethical practice. By harnessing AI's capacity to personalise learning experiences, we can address the diverse needs of students and reduce educational disparities, especially among underserved and marginalised communities. Equally important is the need to integrate AI responsibly, ensuring that fairness, transparency, and privacy remain paramount in our educational practices. This dual commitment—to innovation and to ethical inclusion—was the cornerstone of our conference theme.

The conference program comprised seven thematic tracks:

- **Track 1 – AIED Architectures and Tools:** This track highlighted innovative engineering advances, from multi-agent systems and tangible interfaces to scalable, domain-specific applications across a variety of educational contexts.
- **Track 2 – Machine Learning and Generative AI:** Emphasising data-driven techniques and generative models, this track explored novel algorithms and computational approaches that enhance predictive and adaptive functionalities in educational environments.
- **Track 3 – Learning, Teaching, and Pedagogy:** Focusing on empirical and theoretical work, this track examined the impact of AI on teaching practices, learner engagement, and the evolution of both formal and informal learning experiences.

[1] https://portal.core.edu.au/conf-ranks.

- **Track 4 – Human-Centred Design and Design-Based Research:** This track addressed the design principles and interactive strategies that place educators and learners at the centre, fostering collaborative human-AI partnerships to promote agency and effective decision-making.
- **Track 5 – Teaching AI:** Dedicated to advancing AI literacy, the research in this track explored methods and technologies that demystify AI, empowering individuals to understand, utilise, and critically evaluate its capabilities.
- **Track 6 – Ethics, Equity, and AIED in Society:** This track delved into the societal dimensions of AI in education, with an emphasis on fairness, transparency, privacy, and strategies to mitigate educational disparities.
- **Track 7 – Theoretical Aspects of AIED and AI-Based Modeling for Education:** This track emphasised conceptual and theoretical contributions providing deeper insights into cognitive, motivational, and socio-cognitive processes, underpinning rigorous models of learning and teaching.

AIED 2025 attracted an exceptionally broad range of submissions. We received over 800 submissions across all tracks, of which 711 met the eligibility criteria and were accepted for review. From the full paper submissions, 130 were accepted as full papers, and an additional 129 were accepted as short papers. The acceptance rates for both full and short papers were 18%, respectively.

Every paper and presentation at AIED 2025 underwent a rigorous, double-blind peer-review process designed to ensure the highest scholarly standards and fair evaluation, with an average of three reviews per submission. This meticulous process—drawing on the expertise of leading researchers, seasoned reviewers, and senior committee members—culminated in an outstanding program that spanned the full spectrum of AIED research and practice.

Beyond the scholarly sessions and paper presentations compiled in these LNCS volumes, AIED 2025 offered a rich program of keynote addresses, panel discussions, interactive demonstrations, and specialised sessions which are compiled in the CCIS volumes of the conference. Specifically, this year had both a general AIED panel as well as a joint panel with sister conferences, Educational Data Mining (EDM 2025) and ACM Learning @ Scale (L@S 2025), which were co-located in Palermo with AIED 2025. Attendees had the opportunity to participate in the Doctoral Consortium, Workshops and Tutorials, Late-Breaking Results sessions, and dedicated tracks for practitioners, industry, and policy. Notably, the BlueSky track invited visionary reflections on past progress and future trajectories in AIED, while the WideAIED track amplified perspectives from under-represented regions and communities. Additionally, the conference offered presentations on the latest publication in IJAIED[2], the journal associated with the conference. As a member of the International Alliance to Advance Learning in the Digital Era (IAALDE)[3] alliance, AIED 2025 hosted, last but not least, presentations of the best papers from other IAALDE conferences (in particular, ICCE 2024 and LAK 2025).

[2] https://link.springer.com/journal/40593.
[3] https://alliancelss.com/.

This conference would not have been possible without the unwavering commitment of the International Artificial Intelligence in Education Society, the Organising Committee, the Program and Senior Program Committees, and the sponsors and partners who support our mission. Their passion and dedication to advancing intelligent educational systems is the foundation of AIED's continued success and innovation.

We invite you to engage wholeheartedly in the exchange of ideas, to challenge conventional paradigms, and to forge new collaborations that will shape the future of education. Together, we will explore how artificial intelligence can act as a true catalyst for creating more inclusive, personalised, and ethical learning environments for teachers and students around the world. Welcome to AIED 2025 — where the future of education is actively being reimagined, one innovative idea at a time.

July 2025

Alexandra I. Cristea
Erin Walker
Yu Lu
Olga C. Santos
Seiji Isotani

Organization

Conference General Co-chairs

Olga C. Santos	UNED, Spain
Seiji Isotani	University of São Paulo, Brazil

Program Co-chairs

Alexandra I. Cristea	Durham University, UK
Erin Walker	Arizona State University, USA
Yu Lu	Beijing Normal University, China

Local Organising Co-chairs

Davide Taibi	Institute for Educational Technology, National Research Council, Italy
Giusuè Lo Bosco	University of Palermo, Italy

Doctoral Consortium Co-chairs

Zitao Liu	Jinan University, China
Carla Limongelli	University of Roma Tre, Italy
Sergey Sosnovsky	Utrecht University, Netherlands

Workshop and Tutorials Co-chairs

Gabriella Casalino	University of Bari, Italy
Tanja Mitrovic	University of Canterbury, New Zealand
Sebastiane Lalle	Sorbonne University, France

Practitioners, Industry, and Policy Co-chairs

Sreecharan Sankaranarayanan	Amazon, USA
Penghe Chen	Beijing Normal University, China
Emmanuel G. Blanchard	Le Mans University, France
Christian M. Stracke	University of Bonn, Germany

Posters and Late Breaking Results Co-chairs

Jingyun Wang	Durham University, UK
Ilona Buchem	Berliner Hochschule für Technik, Germany

AIED Mentoring Fellowship Co-chairs

Vania Dimitrova	University of Leeds, UK
Ig Ibert Bittencourt	Universidade Federal de Alagoas, Brazil

Panel Co-chairs

Ryan Baker	University of Pennsylvania, USA
Marta Cimitile	UnitelmaSapienza, Italy

Blue Sky Co-chairs

Stamos Katsigiannis	Durham University, UK
Bruce M. McLaren	Carnegie Mellon University, USA

WideAIED Co-chairs

Isabel Hilliger	Pontificia Universidad Católica, Chile
Giovanni Fulantelli	Institute for Educational Technology, National Research Council, Italy
Ilaria Torre	University of Genoa, Italy

Diversity and Inclusion Co-chairs

Valery Psyché	TÉLUQ University, Canada
Kaska Porayska-Pomsta	UCL Institute of Education, UK

Interactive Events Co-chairs

Leonardo B. Marques	Federal University of Alagoas, Brazil
Emmanuel G. Blanchard	LIUM, Le Mans Université, France

Proceedings Co-chairs

Paraskevi Topali	Radboud University, Netherlands
Rafael D. Araújo	Universidade Federal de Uberlândia, Brazil
Elaine Harada	Universidade Federal do Amazonas, Brazil

International Journal of AIED Chair

Conrad Borchers	Human-Computer Interaction Institute, USA

Virtual Experiences Co-chairs

Teng Guo	Jinan University, China
Salvatore Calderaro	University of Palermo, Italy

Publicity Co-chairs

Miguel Portaz	UNED, Spain
Xinyuan Hao	University College London, UK
Mariella Farella	Institute for Educational Technology, National Research Council, Italy

Volunteer Co-chairs

Lele Sha	Monash University, Australia
Domenico Amato	University of Palermo, Italy

Sponsorship Chair

Pietro Paolo Corso	University of Palermo, Italy

Scholarship Chair

Michele Tumminello	University of Palermo, Italy

Website Chair

Andrea Nelson Mauro	Institute for Educational Technology, National Research Council, Italy

Research Track Junior Advisory Board

Conrad Borchers	Human-Computer Interaction Institute, USA
Amine Hatun Ataş	Galatasaray University, Turkey
Krishna Chaitanya Rao Kathala	University of Massachusetts, USA
Guher Gorgun	IPN – Leibniz Institute for Science and Mathematics Education, Germany
Isaac Alpizar-Chacon	Utrecht University, Netherlands

Research Track 1 – AIED Architectures and Tools

Senior Program Committee Members

Ig Ibert Bittencourt	Federal University of Alagoas, Brazil
Emmmanuel G. Blanchard	LIUM, Le Mans Université, France
François Bouchet	Sorbonne Université - LIP6, France
Julien Broisin	IRIT, Université Toulouse III - Paul Sabatier, France

Vania Dimitrova	University of Leeds, UK
Ralph Ewerth	L3S Research Center, Leibniz Universität Hannover, Germany
Sébastien George	LIUM, Le Mans Université, France
Srecko Joksimovic	University of South Australia, Australia
Mirko Marras	University of Cagliari, Italy
Riichiro Mizoguchi	Japan Advanced Institute of Science and Technology, Japan
Andrew Olney	University of Memphis, USA
Craig Stewart	Durham University, UK

Program Committee Members

Pablo Arnau-González	Universidad de Valencia, Spain
Juan I. Asensio-Pérez	Universidad de Valladolid, Spain
Aneesha Bakharia	University of Queensland, Australia
Beata Beigman Klebanov	Educational Testing Service, USA
Anis Bey	Higher School of Management Sciences - Annaba, Algeria
Anthony F. Botelho	University of Florida, USA
Anis Boubaker	École de Technologie Supérieure (ETS), Canada
German Capdehourat	Ceibal, Uruguay
Alberto Casas-Ortiz	Universidad Nacional de Educación a Distancia, Spain
Eason Chen	Carnegie Mellon University, Taiwan
Jiahao Chen	TAL, China
Penghe Chen	Beijing Normal University, China
Pei Hua Cher	National University of Singapore, Singapore
Sergio Chevtchenko	Western Sydney University, Australia
Maria González	Tecnológico de Monterrey Campus Laguna, Mexico
Rafael D. Araújo	Universidade Federal de Uberlândia, Brazil
Barbara Ericson	University of Michigan, USA
Effat Farhana	Vanderbilt University, USA
Brendan Flanagan	Kyoto University, Japan
Kaushik Gopalan	FLAME University, India
Cyril Goutte	National Research Council Canada, Canada
Jiangang Hao	Educational Testing Service, USA
Elaine Harada	Universidade Federal do Amazonas, Brazil
Yann Hicke	Cornell University, USA
Ruikun Hou	University of Tübingen, Germany
Wanqing Hu	Beijing Normal University, China

Eleni Ilkou	Leibniz University Hannover, Germany
Patricia Jaques	PPGC/UFPEL, Brazil
Yeo Jin Kim	North Carolina State University, USA
Vasily Konovalov	AIRI, Russia
Salima Lamsiyah	Université Sidi Mohamed Ben Abdellah Fès, Morocco
Juho Leinonen	Aalto University, Finland
Zachary Levonian	University of Minnesota, USA
Hang Li	Michigan State University, USA
Zhaohui Li	Penn State University, USA
Siaw Ling Lo	Singapore Management University, Singapore
James Martin	University of Colorado Boulder, USA
Sungjin Nam	ACT, Inc., USA
Seyed Parsa Neshaei	EPFL, Switzerland
Hiran Nonato Macedo Ferreira	Universidade Federal de Uberlândia, Brazil
Narges Norouzi	UC Berkeley, USA
Pedro Orvalho	University of Oxford, UK
Tadachika Ozono	Nagoya Institute of Technology, Japan
Ulrike Pado	HFT Stuttgart, Germany
Eduard Pogorskiy	Open Files Ltd., UK
Valéry Psyché	Université TÉLUQ, Canada
Mitra Purandare	Eastern Switzerland University of Applied Sciences, Switzerland
Manikandan Ravikiran	Hitachi India Pvt Ltd, India
Luiz Rodrigues	Federal University of Alagoas, Brazil
Hannes Rosenbusch	University of Amsterdam, Netherlands
Anirban Roy	SRI International, USA
José A. Ruipérez Valiente	University of Murcia, Spain
Stefan Ruseti	University “Politehnica” of Bucharest, Romania
Ana Serrano-Mamolar	Universidad de Burgos, Spain
Mahsa Sheikhi Karizaki	Penn State University, USA
Atsushi Shimada	Kyushu University, Japan
Siddharth Siddharth	Plaksha University, India
Aakash Soni	ECE Paris, France
Vinitra Swamy	EPFL, Switzerland
Liang Zhang	University of Memphis, USA
Qi Zhou	University College London, UK
Xiaofei Zhou	University of Rochester, USA

Research Track 2 – Machine Learning and Generative AI

Senior Program Committee Members

Ayan Banerjee	Arizona State University, USA
Abhinava Barthakur	University of South Australia, Australia
Gautam Biswas	Vanderbilt University, USA
Nigel Bosch	University of Illinois Urbana-Champaign, USA
Dan Carpenter	North Carolina State University, USA
Guanliang Chen	Monash University, Australia
Ricardo Conejo	Universidad de Málaga, Spain
Mingyu Feng	WestEd, USA
Dragan Gasevic	Monash University, Australia
Johan Jeuring	Utrecht University, Netherlands
Yang Jiang	Columbia University, USA
Irena Koprinska	University of Sydney, Australia
Vitomir Kovanovic	University of South Australia, Australia
Sébastien Lallé	Sorbonne University, France
Andrew Lan	University of Massachusetts Amherst, USA
Mikel Larrañaga	University of the Basque Country UPV/EHU, Spain
Hady Lauw	Singapore Management University, Singapore
Nguyen-Thinh Le	Humboldt Universität zu Berlin, Germany
James Lester	North Carolina State University, USA
Fuhua Lin	Athabasca University, Canada
Zitao Liu	Jinan University, China
Jeffrey Matayoshi	McGraw Hill ALEKS, USA
Noboru Matsuda	North Carolina State University, USA
Wookhee Min	North Carolina State University, USA
Antonija Mitrovic	University of Canterbury, New Zealand
Huy Nguyen	Carnegie Mellon University, USA
Roger Nkambou	Université du Québec à Montréal, Canada
Radek Pelánek	Masaryk University, Czechia
Rinat B. Rosenberg-Kima	Technion - Israel Institute of Technology, Israel
Sreecharan Sankaranarayanan	Carnegie Mellon University, USA
Filippo Sciarrone	Universitas Mercatorum, Italy
Maomi Ueno	University of Electro-Communications, Japan

Program Committee Members

Seth Adjei	Northern Kentucky University, USA
Itziar Aldabe	University of the Basque Country (UPV/EHU), Spain
Manar Alkhatib	British University in Dubai, United Arab Emirates
Samah Alkhuzaey	University of Liverpool, UK
Isaac Alpizar-Chacon	Utrecht University, Netherlands
Antonio R. Anaya	Universidad Nacional de Educacion a Distancia, Spain
Eduardo Aranha	UFRN, Brazil
Pasquale Ardimento	University of Bari Aldo Moro, Italy
Miguel Arevalillo-Herráez	University of Valencia, Spain
Vedant Bahel	University of British Columbia, Canada
Aneesha Bakharia	University of Queensland, Australia
Tiffany Barnes	North Carolina State University, USA
Alex Barrett	Florida State University, USA
Idir Benouaret	Epita Research Laboratory, France
Kamal Berahmand	Queensland University of Technology, Australia
Yoav Bergner	New York University, USA
Mario Luca Bernardi	University of Sannio, Italy
Maria Bolsinova	Tilburg University, Netherlands
Rex Bringula	University of the East, Philippines
Okan Bulut	University of Alberta, Canada
Tyler Burleigh	Khan Academy, USA
Jie Cao	University of Colorado Boulder, USA
Paulo Carvalho	Carnegie Mellon University, USA
Alberto Casas-Ortiz	Universidad Nacional de Educación a Distancia, Spain
Binglin Chen	University of Illinois Urbana-Champaign, USA
Jiahao Chen	TAL, China
Li Chen	Kyushu University, Japan
Wei-Peng Chen	Fujitsu Labs of America, USA
Keith Cochran	DePaul University, USA
Jade Mai Cock	EPFL, Switzerland
Lea Cohausz	University of Mannheim, Germany
Maria González	Tecnológico de Monterrey Campus Laguna, Mexico
Mark G. Core	University of Southern California, USA
Jeffrey Cross	Tokyo Institute of Technology, Japan
Yuan Cui	Northwestern University, USA
Sidney D'Mello	University of Colorado Boulder, USA

Marcos Da Silveira	Luxembourg Institute of Science and Technology, Luxembourg
Mihai Dascalu	University Politehnica of Bucharest, Romania
Pasquale De Meo	University of Messina, Italy
Hannah Deininger	Eberhard-Karls-Universität Tübingen, Germany
Michel Desmarais	Polytechnique Montréal, Canada
Emmanuel Dorley	University of Florida, USA
Junlei Du	Beijing Normal University, Beijing, China
Jamie Duell	Nanyang Technological University, Singapore
Cristina Dumdumaya	University of Southeastern Philippines, Philippines
Andrew Emerson	National Board of Medical Examiners, USA
Wicaksono Febriantoro	University College London, UK
Shihui Feng	University of Hong Kong, China
Nigel Fernandez	University of Massachusetts Amherst, USA
Alessio Ferrato	Roma Tre University, Italy
Rafael Ferreira Mello	Cesar School, Brazil
Joyce Fonteles	Vanderbilt University, USA
Cristiano Galafassi	Universidade Federal do Rio Grande do Sul, Brazil
Wenbin Gan	National Institute of Information and Communications Technology, Japan
Ananya Ganesh	University of Colorado, Boulder, USA
Oscar Garces	University of Technology Sydney, Australia
Alireza Gharahighehi	imec - KU Leuven, Belgium
Guher Gorgun	University of Alberta, Canada
Sabine Graf	Athabasca University, Canada
Wolfgang Gritz	TIB – Leibniz Information Centre for Science and Technology, Germany
Quanlong Guan	Jinan University, China
Haoyu Han	Michigan State University, USA
Jieun Han	KAIST, South Korea
Elaine Harada	Universidade Federal do Amazonas, Brazil
Peng He	Washington State University, USA
Hasnain Heickal	University of Massachusetts Amherst, USA
Anett Hoppe	Leibniz Universität Hannover, Germany
Shuyan Huang	TAL Education Group, China
Yu-Kai Huang	Carnegie Mellon University, USA
Corina Ilinca	University of Bucharest, Romania
Tsunenori Ishioka	National Center for University Entrance Examinations, Japan
Hamid Karimi	Utah State University, USA

Matthew C. Kelley	George Mason University, USA
Rene Kizilcec	Cornell University, USA
Ali Koksal	A*STAR, Singapore
Sotiris Kotsiantis	University of Patras, Greece
Charles Koutcheme	Aalto University, Finland
Udo Kruschwitz	University of Regensburg, Germany
Andrey Kuznetsov	AIRI, Russia
Chahrazed Labba	Université de Lorraine, France
Chandra Shekar Lakshminarayanan	Indian Institute of Technology Madras, India
Ehsan Latif	University of Georgia, USA
Tai Le Quy	University of Koblenz, Germany
Gayoung Lee	Baekseok University, South Korea
Wee Leong Lee	Singapore University of Social Sciences, Singapore
Haiying Li	University of Pennsylvania, USA
Hongming Li	University of Florida, USA
Huiyong Li	Kyushu University, Japan
Juanhui Li	Michigan State University, USA
Shan Li	Lehigh University, USA
Xinyu Li	Monash University, Australia
Zhi Li	University of California, Berkeley, USA
Jionghao Lin	University of Hong Kong, China
Siyuan Liu	Nanyang Technological University, Singapore
Feng Liu	Drexel University, USA
Guimei Liu	A*STAR, Singapore
Naiming Liu	Rice University, USA
Ninghao Liu	University of Georgia, USA
Ziyuan Liu	Singapore Management University, Singapore
Rose Luckin	University College London, UK
Jakub Macina	ETH Zurich, Switzerland
Christopher MacLellan	Georgia Institute of Technology, USA
George Magoulas	University of London, UK
Subhankar Maity	IIT Kharagpur, India
Talles Medeiros	Federal University of Ouro Preto, Brazil
Guilherme Medeiros Machado	ECE Paris, France
Paolo Mengoni	Hong Kong Baptist University, China
Eva Millan	Universidad de Málaga, Spain
Tsunenori Mine	Kyushu University, Japan
Sein Minn	Inria, France
Yoshimitsu Miyazawa	National Center for University Entrance Examinations, Japan

Mehrnoush Mohammadi Bolbanabad	University of Queensland, Australia
Mukesh Mohania	IIIT Delhi, India
Pedro J. Muñoz-Merino	Universidad Carlos III de Madrid, Spain
Takashi Nagai	Institute of Technologists, Japan
Vinay Namboodiri	University of Bath, UK
Hung Tuan Nguyen	Tokyo University of Agriculture and Technology, Japan
Ange Adrienne Nyamen Tato	Université Laval, Canada
Hilário Oliveira	Instituto Federal do Espírito Santo - Campus Serra, Brazil
Luc Paquette	University of Illinois Urbana-Champaign, USA
Zach Pardos	University of California, Berkeley, USA
Rebecca Passonneau	Pennsylvania State University, USA
Tung Phung	MPI-SWS, Germany
Yang Pian	Beijing Normal University, China
Juan Pinto	University of Illinois Urbana-Champaign, USA
Valentina Poggioni	Università di Perugia, Italy
Marjana Prifti Skenduli	University of New York Tirana, Albania
Ratnavel Rajalakshmi	VIT University, Chennai Campus, India
Mladen Rakovic	Monash University, Australia
Erik Rodner	Hochschule für Technik und Wirtschaft, Germany
Oleg Rogov	AIRI, Russia
José Raúl Romero	University of Córdoba, Spain
Lu Rong	Tianjin University, China
Leo Sylvio Rüdian	Humboldt-Universität zu Berlin, Germany
Rohit Saluja	IIT Mandi, India
Batuhan Sayis	Universitat Pompeu Fabra, Spain
Alexander Scarlatos	University of Massachusetts, USA
Kshitij Sharma	Norwegian University of Science and Technology, Norway
Mahsa Sheikhi Karizaki	Penn State University, USA
Leixian Shen	Tsinghua University, China
Bruce Sherin	Northwestern University, USA
Andy Smith	North Carolina State University, USA
Hyun-Je Song	Jeonbuk National University, South Korea
Francesco Sovrano	ETH Zurich, Switzerland
Sebastian Strauß	Ruhr University Bochum, Germany
Man Su	Saarland University, Germany
Zhongtian Sun	University of Kent, UK
Abhijit Suresh	University of Colorado Boulder, USA
Cheng Tang	Kyushu University, Japan

Patricia Tedesco	Universidade Federal de Pernambuco, Brazil
Khushboo Thaker	University of Pittsburgh, USA
Danielle R Thomas	Carnegie Mellon University, USA
Becky Tomaso	University of South Alabama, USA
Rodothea Myrsini Tsoupidi	Independent Researcher, Sweden
Emiko Tsutsumi	University of Electro-Communications, Japan
Miriam Ugarte Querejeta	Mondragon Unibertsitatea, Spain
Josh Underwood	Independent, Spain
Masaki Uto	University of Electro-Communications, Japan
Kirk Vanacore	Worcester Polytechnic Institute, USA
Maureen Villamor	University of Southeastern Philippines, Philippines
Sree Aurovindh Viswanathan	Infosys Technologies Limited, India
Alessandro Vivas	UFVJM, Brazil
Marco Volpe	University of Leicester, UK
Christabel Wayllace	New Mexico State University, USA
Daniel Weitekamp	Carnegie Mellon University, USA
Jinta Weng	University of Chinese Academy of Sciences, China
Chris Wong	University of Technology Sydney, Australia
Tat Wing Wong	Chinese University of Hong Kong, China
Anna Wróblewska	Warsaw University of Technology, Poland
Bo Wu	Colorado School of Mines, USA
Jinran Wu	University of Queensland, Australia
Zihan Wu	University of Michigan, USA
Jinjun Xiong	University at Buffalo, USA
Yaping Xu	Beijing Normal University, China
Takayoshi Yamashita	Chubu University, Japan
Xi Yang	IBM Research, USA
Calvin Yeung	Nagoya University, Japan
Guoxian Yu	Shandong University, China
Jifan Yu	Tsinghua University, China
Rajamanickam Yuvaraj	Anna University, India
Andres Felipe Zambrano	University of Pennsylvania, USA
Jiayi Zhang	University of Pennsylvania, USA
Rui Zhang	Tongji University, China
Shugang Zhang	Ocean University of China, China
Yupei Zhang	Northwestern Polytechnical University, China
Nina Zhou	A*STAR, Singapore

Research Track 3 – Learning, Teaching, and Pedagogy

Senior Program Committee Members

Bita Akram	North Carolina State University, USA
Laura Allen	University of Minnesota, USA
Ivon Arroyo	University of Massachusetts Amherst, USA
Ryan Baker	University of Pennsylvania, USA
Ig Ibert Bittencourt	Federal University of Alagoas, Brazil
Jesus G. Boticario	UNED, Spain
Irene-Angelica Chounta	University of Duisburg-Essen, Germany
Floriana Grasso	University of Liverpool, UK
Beate Grawemeyer	Coventry University, UK
Jason Harley	McGill University, Canada
Srecko Joksimovic	University of South Australia, Australia
Akihiro Kashihara	University of Electro-Communications, Japan
Amruth Kumar	Ramapo College of New Jersey, USA
Tanja Käser	EPFL, Switzerland
Roberto Martinez-Maldonado	Monash University, Australia
Gordon McCalla	University of Saskatchewan, Canada
Kathryn McCarthy	Georgia State University, USA
Bruce McLaren	Carnegie Mellon University, USA
Tomohiro Nagashima	Saarland University, Germany
Ido Roll	Technion - Israel Institute of Technology, Israel
Nikol Rummel	Ruhr-Universität Bochum, Germany
Craig Stewart	Durham University, UK
Marco Temperini	Università di Roma, Italy
Diego Zapata-Rivera	Educational Testing Service, USA

Program Committee Members

Adetunji Adeniran	Carnegie Mellon University, USA
Daniele Agostini	University of Trento, Italy
Azza Abdullah S. Alghamdi	KAU, Saudi Arabia
Manal Alhathli	Princess Nourah bint Abdulrahman University, Saudi Arabia
Ioannis Anastasopoulos	University of California, Berkeley, USA
Ella Anghel	Rutgers University, USA
David Arnau	University of Valencia, Spain
Burcu Arslan	Educational Testing Service, USA

Lili Aunimo	Haaga-Helia University of Applied Sciences, Finland
Miriam Celia Bergue Alves	Naval Postgraduate School, USA
Nirav Bhatt	IIT Madras, India
Conrad Borchers	Carnegie Mellon University, USA
Kristy Elizabeth Boyer	University of Florida, USA
Maria Concetta Carruba	Università Telematica Pegaso, Italy
Omar D. Castrillon	Universidad Nacional de Colombia, Colombia
Silvio César Cazella	UFCSPA, Brazil
Li Chen	Kyushu University, Japan
Si Chen	University of Notre Dame, USA
Jennifer Cromley	University of Illinois Urbana-Champaign, USA
Sidney D'Mello	University of Colorado Boulder, USA
Rafael D. Araújo	Universidade Federal de Uberlândia, Brazil
Cesar Delgado	North Carolina State University, USA
Diego Dermeval	Federal University of Alagoas, Brazil
Blaženka Divjak	University of Zagreb, Croatia
Isabela Dramnesc	West University of Timisoara, Romania
Benedict du Boulay	University of Sussex, UK
Nghia Duong-Trung	German Research Centre for Artificial Intelligence, Germany
Barbara Ericson	University of Michigan, USA
Cassia Fernandez	USP, Brazil
Carol Forsyth	Educational Testing Service, USA
Selen Galiç	Hacettepe University, Turkey
Andrea Garavaglia	Università degli Studi di Milano, Italy
Francisco José García-Peñalvo	University of Salamanca, Spain
Jan-Martin Geiger	University of Münster, Germany
Ralucca Gera	Naval Postgraduate School, USA
Lucia Giraffa	Pontificia Universidade Católica do Rio Grande do sul, Brazil
Jiesi Guo	Australian Catholic University, Australia
Shivang Gupta	Carnegie Mellon University, USA
Ashish Gurung	Carnegie Mellon University, USA
Songhee Han	Florida State University, USA
Elaine Harada	Universidade Federal do Amazonas, Brazil
Yugo Hayashi	Ritsumeikan University, Japan
Yusuke Hayashi	Hiroshima University, Japan
Davinia Hernández-Leo	Universitat Pompeu Fabra, Barcelona, Spain
László Horváth	Eötvös Loránd University, Hungary
Xiangen Hu	Hong Kong Polytechnic University, China
Stephen Hutt	University of Denver, USA

Paul Inventado	California State University Fullerton, USA
Jennifer Jacobs	University of Colorado Boulder, USA
Enkelejda Kasneci	Technical University of Munich, Germany
Hassan Khosravi	University of Queensland, Australia
Elizabeth Koh	Nanyang Technological University, Singapore
Kazuaki Kojima	Teikyo University, Japan
Harsh Kumar	University of Toronto, Canada
Eleni Kyza	Cyprus University of Technology, Cyprus
Élise Lavoué	Université Jean Moulin Lyon 3, LIRIS, France
Tobias Ley	University for Continuing Education Krems, Austria
Liang-Yi Li	National Taiwan Normal University, Taiwan
Shan Li	Lehigh University, USA
Tingting Li	Washington State University, USA
Chang-Yen Liao	National Central University, Taiwan
Amna Liaqat	University of Toronto, Canada
Boxuan Ma	Kyushu University, Japan
Uwe Maier	University of Education Schwäbisch Gmünd, Germany
Vasco Manquinho	Universidade de Lisboa, Portugal
Michele Marangi	eCampus University, Italy
Leonardo Brandão Marques	Universidade Federal de Alagoas, Brazil
Ruben Martins	Carnegie Mellon University, USA
Alejandra Martínez-Monés	Universidad de Valladolid, Spain
D. Moritz Marutschke	Kyoto University of Advanced Science, Japan
Ana Mouta	Universidad de Salamanca, Portugal
Michael Mrazek	University of Texas Health Science Center, USA
Nidhi Nasiar	University of Pennsylvania, USA
Allen Nie	Stanford University, USA
Abbie Olszewski	University of Nevada, Reno, USA
Ranilson Paiva	Universidade Federal de Alagoas, Brazil
Abelardo Pardo	University of Adelaide, Australia
Terry Payne	University of Liverpool, UK
Livia Petti	University of Molise, Italy
Federica Picasso	University of Trento, Italy
Oleksandra Poquet	University of South Australia, Australia
Liu Qiang	Sichuan Chengdu Shude High School, China
Ilana Ram	Technion Israel Institute of Technology, Israel
Sowmya Ramachandran	Stottler Henke Associates Inc., USA
J. Elizabeth Richey	University of Pittsburgh, USA
Maren Scheffel	Ruhr University Bochum, Germany
Tami Seifert	Kibbutzim College of Education, Israel

Lele Sha	Monash University, Australia
Hajime Shirouzu	National Institute for Educational Policy Research, Japan
Katharina Simbeck	HTW Berlin, Germany
Stephen Sireci	University of Massachusetts Amherst, USA
Krista Sorri	Tampere University, Finland
Angela Stewart	Carnegie Mellon University, USA
Qing Sun	Beihang University, China
Jakub Swacha	University of Szczecin, Poland
Tamara Tate	University of California, Irvine, USA
Paraskevi Topali	Radboud University, Netherlands
Yi-Shan Tsai	Monash University, Australia
Manuel Valle Torre	Delft University of Technology, Netherlands
Alexandra Vassar	UNSW, Sydney, Australia
Esteban Vazquez-Cano	UNED, Spain
Deliang Wang	University of Hong Kong, China
Ning Wang	University of Southern California, USA
Stephan Weibelzahl	Private University of Applied Sciences Göttingen, Germany
Adam Wierzbicki	Polish-Japanese Academy of Information Technology, Poland
Ruiwei Xiao	Carnegie Mellon University, USA
Dong Yang	Beijing Normal University, China
Franck Zenasni	Université Paris Cité, France
Lechen Zhang	University of Helsinki, Finland
Yi Zhang	Beijing Normal University, China
Gaoxia Zhu	Nanyang Technological University, Singapore
Craig Zilles	University of Illinois Urbana-Champaign, USA
Mia Čarapina	University of Applied Sciences, Zagreb, Croatia
Piotr Olaf Żylicz	WSB Merito University in Poznań, Poland

Research Track 4 – Human-Centered Design and Design-Based Research

Senior Program Committee Members

Laia Albó	Universitat de Vic, Spain
Jesus G. Boticario	UNED, Spain
Cesar Collazos	Universidad del Cauca, Colombia
Mutlu Cukurova	University College London, UK
Tomoo Inoue	University of Tsukuba, Japan

David Joyner	Georgia Institute of Technology, USA
Simon Knight	University of Technology Sydney, Australia
Maria Mercedes T. Rodrigo	Ateneo de Manila University, Philippines
Stefan Trausan-Matu	University Politehnica of Bucharest, Romania

Program Committee Members

Yusuf Akemoğlu	Duzce University, Turkey
Sihem Amer-Yahia	CNRS - LIG, France
Yuya Asano	University of Pittsburgh, USA
Jordan Barria-Pineda	University of Pittsburgh, USA
Sarah Bichler	Ludwig Maximilian University of Munich, Germany
Carmen Bisogni	Università degli Studi di Salerno, Italy
Minghao Cai	University of Alberta, Canada
Hyunjin Cha	Soonchunhyang University, South Korea
Eason Chen	Carnegie Mellon University, USA
Elaine Chen	Nottingham Trent University, UK
Thomas K. F. Chiu	Chinese University of Hong Kong, China
M. Ali Akber Dewan	Athabasca University, Canada
Rahul Divekar	Rensselaer Polytechnic Institute, USA
Martin Ebner	Graz University of Technology, Austria
Ann-Christin Falhs	Ruhr-Universität Bochum, Germany
Gloria Fernandez	Universidad Distrital Francisco José de Caldas, Colombia
Reva Freedman	Northern Illinois University, USA
Nikki G. Lobczowski	University of Pittsburgh, USA
Kobi Gal	University of Edinburgh, UK
Shivang Gupta	Carnegie Mellon University, USA
Ella Haig	University of Portsmouth, UK
Songhee Han	Florida State University, USA
Peng He	Washington State University, USA
John Hollander	Arkansas State University, USA
Xinying Hou	University of Michigan, USA
Xiaoshan Huang	McGill University, Canada
Xinyue Jiao	New York University, USA
Dorota Kamińska	Instytut Mechatroniki i Systemów Informatycznych Politechniki Łódzkiej, Poland
Yoon Jeon Kim	University of Wisconsin-Madison, USA
Harsh Kumar	University of Toronto, Canada
Seiyon Lee	University of Florida, USA

Bibeg Limbu	University of Duisburg-Essen, Germany
Xinyi Lu	University of Michigan, USA
Linda Mannila	University of Helsinki, Finland
Leonardo Brandão Marques	University of São Paulo, Brazil
Wannisa Matcha	Prince of Songkla University, Thailand
Jennifer Olsen	University of San Diego, USA
Fatiha Sadat	UQAM, Canada
Jerry Schnepp	Roosevelt University, USA
Ratan Sebastian	TIB Leibniz Information Centre for Science and Technology, Germany
Lei Shi	Newcastle University, UK
Valerie Shute	Florida State University, USA
Sri Yash Tadimalla	University of North Carolina Charlotte, USA
Marwa Trabelsi	L3i Laboratory, La Rochelle University, France
Alyssa Van Camp	TeachFX, USA
Alexandra Sasha Vassar	UNSW, Sydney, Australia
Jayakrishnan Warriem	IIT Madras, India
Beverly Woolf	University of Massachusetts, USA
Grzegorz Zwoliński	Lodz University of Technology, Poland

Research Track 5 – Teaching AI

Senior Program Committee Members

Bert Bredeweg	University of Amsterdam, Netherlands
Floriana Grasso	University of Liverpool, UK
Carla Limongelli	Università Roma Tre, Italy
Benjamin Nye	University of Southern California, USA
Mohammed Saqr	University of Eastern Finland, Finland
Craig Stewart	Durham University, UK
Rosa Vicari	Universidade Federal do Rio Grande do Sul, Brazil

Program Committee Members

Azza Abdullah S. Alghamdi	King Abdulaziz University, Saudi Arabia
Mario Luca Bernardi	University of Sannio, Italy
Alka Bhushan	Indian Institute of Technology Madras, India
Ilona Buchem	Berlin University of Applied Sciences, Germany
William Cain	University of Wyoming, USA

Gaowei Chen	University of Hong Kong, China
Penghe Chen	Beijing Normal University, China
Si Chen	University of Notre Dame, USA
Chih-Yueh Chou	Yuan Ze University, Taiwan
Yo Ehara	Tokyo Gakugei University, Japan
Manuel Freire Morán	Universidad Complutense de Madrid, Spain
Bruno Galmar	National Central University, Taiwan
Sai Gattupalli	University of Massachusetts Amherst, USA
Jeff Ginger	University of Illinois Urbana-Champaign, USA
Arto Hellas	Aalto University, Finland
Iris Howley	Williams College, USA
Xinyue Jiao	New York University, USA
Judy Kay	University of Sydney, Australia
Marie Lefevre	LIRIS - Université Lyon 1, France
Nicholas Lytle	Georgia Institute of Technology, USA
Michael Mrazek	University of Texas Health Science Center, USA
Kasia Muldner	Carleton University, Canada
Andrew Olney	University of Memphis, USA
Livia Petti	University of Molise, Italy
Marcela Pianucci	COGNA, Brazil
Liu Qiang	Sichuan Chengdu Shude High School, China
Angelo Rega	Pegaso Telematic University, Italy
Daniela Rotelli	Sorbonne Université, France
Mika Saari	Tampere University of Technology, Finland
Isabella Saccardi	Utrecht University, Netherlands
Sabine Seufert	University of St. Gallen, Switzerland
Christian M. Stracke	University of Bonn, Germany
Vidya Sudarshan	Nanyang Technological University, Singapore
Lars-Jochen Thoms	Thurgau University of Teacher Education, Switzerland
Xiaoyi Tian	North Carolina State University, USA
Maya Usher	Technion, Israel
Ruiwei Xiao	Carnegie Mellon University, USA
Pingjing Yang	University of Illinois Urbana-Champaign, USA
Jianwei Zhang	University at Albany, USA
Yingbin Zhang	South China Normal University, China
Jia Zhu	Zhejiang Normal University, China
Meina Zhu	Wayne State University, USA

Research Track 6 – Ethics, Equity, and AIED in Society

Senior Program Committee Members

Armelle Brun	LORIA - Université de Lorraine, France
Isabela Gasparini	Santa Catarina State University, Brazil
Monique Grandbastien	LORIA, Université de Lorraine, France
Peter Hastings	DePaul University, USA
Vanda Luengo	Sorbonne Université - LIP6, France
Collin Lynch	North Carolina State University, USA
Roberto Martinez-Maldonado	Monash University, Australia
Craig Thompson	University of British Columbia, Canada

Program Committee Members

Meital Amzalag	Holon Institute of Technology, Israel
Clara Belitz	University of Illinois Urbana-Champaign, USA
Susan Beudt	Deutsches Forschungszentrum für Künstliche Intelligenz, Germany
William Cain	University of Wyoming, USA
Francisco Enrique Vicente Castro	New York University, USA
Chelsea Chandler	University of Colorado Boulder, USA
Irene-Angelica Chounta	University of Duisburg-Essen, Germany
Markus Christen	University of Zurich, Switzerland
Lea Cohausz	University of Mannheim, Germany
Jeanine DeFalco	University of New Haven, USA
Diego Dermeval	Federal University of Alagoas, Brazil
Aniket Deroy	IIT Kharagpur, India
Yannis Dimitriadis	University of Valladolid, Spain
Konomu Dobashi	Aichi University, Japan
Juan Manuel Dodero	Universidad de Cádiz, Spain
Deborah Fields	Utah State University, USA
Kobi Gal	University of Edinburgh, UK
Andrew Gibson	Queensland University of Technology, Australia
Daniela Inclezan	Miami University, USA
Kirsty Kitto	University of Technology, Sydney, Australia
Rene Kizilcec	Cornell University, USA
Gayoung Lee	Baekseok University, South Korea
Haejin Lee	University of Illinois Urbana-Champaign, USA
Seiyon Lee	University of Florida, USA
Chenglu Li	University of Utah, USA

Tingting Li	Washington State University, USA
Elisa Marengo	University of Torino, Italy
Goran Martinovic	J.J. Strossmayer University of Osijek, Croatia
Jon Mason	Charles Darwin University, Australia
Hagit Meishar-Tal	Holon Institute of Technology, Israel
Roberto Micalizio	Università di Torino, Italy
Audrey Michal	University of Michigan, USA
Iulia Mihalache	Université du Québec en Outaouais, Canada
Phaedra Mohammed	University of the West Indies, Trinidad and Tobago
Patrick Ocheja	Kyoto University, Japan
Martha Palmer	University of Colorado Boulder, USA
Viktoria Pammer-Schindler	Graz University of Technology, Austria
Dimitra Pappa	NCSR Demokritos, Greece
Giorgio Pedrazzi	University of Brescia, Italy
Bruno Pimentel	Universidade Federal de Alagoas, Brazil
Niels Pinkwart	Humboldt-Universität zu Berlin, Germany
Priya Sharma	Pennsylvania State University, USA
Amy Stornaiuolo	University of Pennsylvania, USA
Hiroki Tanioka	Tokushima University, Japan
Marcelo Tibau	Universidade Federal do Estado do Rio de Janeiro, Brazil
Adrienn Toth	ETH Zurich, Switzerland
Alistair Windsor	Memphis University, USA
Feiwen Xiao	Penn State University, USA
Piotr Olaf Żylicz	WSB Merito University in Poznań, Poland

Research Track 7 – Theoretical Aspects of AIED and AI-Based Modeling for Education

Senior Program Committee Members

Mohsen Dorodchi	University of North Carolina Charlotte, USA
H. Chad Lane	University of Illinois Urbana-Champaign, USA
Sergey Sosnovsky	Utrecht University, Netherlands

Program Committee Members

Stephen Aguilar	University of Southern California, USA
Yusuf Akemoglu	Duzce University, Turkey

Ahmed Al Khateeb	University of Southampton, UK
Ainhoa Alvarez	University of the Basque Country, UPV/EHU, Spain
Francisco Bellas	Universidade da Coruna, Spain
Ching-sing Chai	Chinese University of Hong Kong, China
Sidney D'Mello	University of Colorado Boulder, USA
Fabiano Dorça	Universidade Federal de Uberlândia, Brazil
Yo Ehara	Tokyo Gakugei University, Japan
Mohamed Ez-Zaouia	Université de Rennes, France
Yael Feldman-Maggor	KTH Royal Institute of Technology, Sweden
Sebastian Gombert	DIPF | Leibniz Institute for Research and Information in Education, Germany
Art Graesser	University of Memphis, USA
Jiesi Guo	Australian Catholic University, Australia
Yusuke Hayashi	Hiroshima University, Japan
Corina Ilinca	University of Bucharest, Romania
Mizue Kayama	Shinshu University, Japan
Chanmin Kim	Pennsylvania State University, USA
Armanda Lewis	New York University, USA
Jiaju Lin	Pennsylvania State University, USA
Frieder Loch	Ostschweizer Fachhochschule, Switzerland
Abbie Olszewski	University of Nevada, Reno, USA
Sebastian Strauß	Ruhr University Bochum, Germany
Mikel Villamañe	University of the Basque Country, Spain
Peter Wulff	Heidelberg University of Education, Germany
Sho Yamamoto	Kindai University, Japan
Zhiwei Yang	Jinan University, China

International Artificial Intelligence in Education (IAIED) Society

IAIED Management Board

President

Olga C. Santos	UNED, Spain

President-Elect

Seiji Isotani	University of São Paulo, Brazil

Secretary

Vania Dimitrova	University of Leeds, UK

Journal Editors

Vincent Aleven	Carnegie Mellon University, USA
Cristina Conati	University of British Columbia
Marcus Specht	Delft University of Technology

Finance Chair

Benedict du Boulay	University of Sussex, UK

Membership Chair

Benjamin D. Nye	University of Southern California, USA

IAIED Officers

Conrad Borchers	Carnegie Mellon University, USA
Xinyuan Hao	University College London, UK
Jionghao Lin	University of Hong Kong, China
Miguel Portaz	UNED, Spain

IAIED Executive Committee

Akihiro Kashihara	University of Electro-Communications, Japan
Amruth Kumar	Ramapo College of New Jersey, USA
Christothea Herodotou	Open University, UK
Jeanine A. Defalco	CCDC-STTC, USA
Judith Masthoff	Utrecht University, Netherlands
Maria Mercedes T. Rodrigo	Ateneo de Manila University, Philippines
Ning Wang	University of Southern California, USA
Olga C. Santos	UNED, Spain
Rawad Hammad	University of East London, UK
Zitao Liu	Jinan University, China
Bruce M. McLaren	Carnegie Mellon University, USA
Cristina Conati	University of British Columbia, Canada
Diego Zapata-Rivera	Educational Testing Service, Princeton, USA
Erin Walker	University of Pittsburgh, USA
Seiji Isotani	University of São Paulo, Brazil
Tanja Mitrovic	University of Canterbury, New Zealand
Noboru Matsuda	North Carolina State University, USA
Min Chi	North Carolina State University, USA

Alexandra I. Cristea	Durham University, UK
Neil Heffernan	Worcester Polytechnic Institute, USA
Andrew M. Olney	University of Memphis, USA
Irene-Angelica Chounta	University of Duisburg-Essen, Germany
Beverly Park Woolf	University of Massachusetts, USA
Ig Ibert Bittencourt	Universidade Federal de Alagoas, Brazil

International Alliance to Advance Learning in the Digital Era (IAALDE)

The IAIED Society is a member of the International Alliance to Advance Learning in the Digital Era (IAALDE) and belongs to its board. IAIED is currently represented at IAALDE by Olga C. Santos, as president of IAIED.

AIED 2025 Sponsors

Platinum



Gold



Silver



Bronze







Best Paper Award Sponsorship



Contents – Part V

Short Papers

When GPT-4 Goes to Class: Benchmarking MOOCs and Enhancing Course Design with AI 3
Mohamed Elhayany and Christoph Meinel

CogGen: A Learner-Centered Generative AI Architecture for Intelligent Tutoring with Programming Videos 11
Wengxi Li, Roy Pea, Nick Haber, and Hariharan Subramonyam

Utilizing Log-Based and Neurophysiological Measures to Understand Engagement and Learning with Intelligent Tutoring Systems 19
Yushuang Liu, Ido Davidesco, Bruce McLaren, J. Elizabeth Richey, Xiaorui Xue, Leah Teffera, Hayden Stec, Hyosun Lee, Jiayi Zhang, Suyi Liu, and Elana Zion-Golumbic

"My Grade is Wrong!": A Contestable AI Framework for Interactive Feedback in Evaluating Student Essays 27
Shengxin Hong, Chang Cai, Sixuan Du, Haiyue Feng, Siyuan Liu, and Xiuyi Fan

WILLM: A System for Academic Writing Improvement Based on Large Language Models 36
Yongli Mou, Fatemeh Fathi, Ben Thillen, and Stefan Decker

What Makes Teamwork Work? A Multimodal Case Study on Emotions and Diagnostic Expertise in an Intelligent Tutoring System 44
Xiaoshan Huang, Haolun Wu, Xue Liu, and Susanne P. Lajoie

Beyond Code: Leveraging ChatGPT to Enhance (Not Replace) Traditional Programming Education 53
Adetunji Adeniran and Manal Alhathli

One Code to Predict Them All: Universal Encoding for Inquiry Modeling 60
Jade Mai Cock, Valentine Delevaux, Ido Roll, Richard Davis, and Tanja Käser

"I Would Share It, But..." Exploring Ways to Optimize the Privacy-Personalization Trade-Off in Intelligent Tutoring Systems 68
Adrienn Toth, Neele Roch, Noé Zufferey, and Verena Zimmermann

A Participatory Strategy for AI Ethics in Education and Rehabilitation Grounded in the Capability Approach 77
Valeria Cesaroni, Eleonora Pasqua, Piercosma Bisconti, and Martina Galletti

Teacher Perceptions on AI Support for Bi/multilingual Learners: Superpowers and Constraints 85
Alina Guha, Shamya Karumbaiah, Cynthia Baeza, Mariana Castro, Kiyo White, and Diego Roman

Towards a Multimodal Document-Grounded Conversational AI System for Education 92
Karan Taneja, Anjali Singh, and Ashok K. Goel

Generative AI in K-12 Education Policy: A Preliminary Global Review 100
Camila Hidalgo and Noha Lea Halim

Where's the F***ing Filter? A Survey on Handling Problematic and Sensitive Questions in Open Educational Chatbots 108
Thilo I. Dieing and Lea Cohausz

Advancing AI Literacy in Medical Education: A Medical AI Competency Framework Development 116
Chang Cai, Jamie Duell, Daisy Minghui Chen, Weng Kin Ho, Bernett Teck Kwong Lee, Fang Li, Siyuan Liu, Olivia Ng, Vidya Sudarshan, Shang-Ming Zhou, Gaoxia Zhu, and Xiuyi Fan

SMARE—Structure Matching and Recognition Engine for Hand-Drawn Chemical Formulas 124
Mitra Purandare, Tobias Rothlin, Frieder Loch, Johannes Huwer, and Lars-Jochen Thoms

An Emergent Bottom-Up Categorization of Students' LLMs Usage in an Undergraduate Research Course 133
Ivan Orozco Vasquez, Romina Mahinpei, Noureddine Elouazizi, and Cristina Conati

Large Language Models for Zero-Shot Exercise Recommendation in Adaptive Learning 140
Tengju Li, Cunling Bian, Kaiquan Chen, and Weigang Lu

Learning by Teaching: Enhancing Music Learning Through LLM-Based Teachable Agents 148
Lingxi Jin, Baicheng Lin, Mengze Hong, Hyo-Jeong So, and Kun Zhang

Searching for Transactivity in a Collaborative Chat: Analysis of Novelty and Reference with GenAI 156
David M. Otten, Pantelis M. Papadopoulos, Maryam A. Haeri, and Maaike D. Endedijk

Risk or Promise: Educators' Perspectives on Student Use of ChatGPT 164
Tami Seifert and Taly Tzamir-Levi

Bridging the AI Adoption Gap: Designing an Interactive Pedagogical Agent for Higher Education Instructors 171
Si Chen, Reid Metoyer, Khiem Le, Adam Acunin, Izzy Molnar, Alex Ambrose, James Lang, Nitesh Chawla, and Ronald Metoyer

LipMOOC: A Modular Synthetic MOOC for Lipreading 179
Snigdha Agarwal, Vinay Namboodiri, and C. V. Jawahar

LLM-Driven Personalized Answer Generation and Evaluation 187
Mohammadreza Molavi, Mohammadreza Tavakoli, Mohammad Moein, Abdolali Faraji, and Gábor Kismihók

Predicting User Actions in Algebra Intelligent Tutoring Systems with Markov Models 195
Pablo Arnau-González, Yuyan Wu, Sergi Solera-Monforte, David Arnau, and Miguel Arevalillo-Herráez

Many-Shot Regurgitation Prompting 203
Shashank Sonkar, Naiming Liu, and Richard Baraniuk

Distinguishing AI-Generated and Human-Written Text Through Psycholinguistic Analysis 212
Chidimma Opara

Improving Open-Response Assessment with LearnLM 220
Danielle R. Thomas, Conrad Borchers, Shambhavi Bhushan, Sanjit Kakarla, Alex Houk, Ralph Abboud, Shivang Gupta, Erin Gatz, and Kenneth R. Koedinger

Multimodal Classroom Diarization with GPT Re-scoring: Teacher or Student? 228
Matthew Perez, Berk Coker, Kemal Berk Kocabagli, Jessica Vitale, and Alyssa Van Camp

How Generative AI Impact Student Emotion and Engagement in Programming Tasks? 236
Boxuan Ma, Liyuan Guo, Tianyuan Yang, and Jihong Ding

Student Perceptions of Adaptive Goal Setting Recommendations: A Design Prototyping Study 244
Conrad Borchers, Cindy Peng, Qianru Lyu, Paulo F. Carvalho, Kenneth R. Koedinger, and Vincent Aleven

Ordinal Classification for Transformer-Based Automated Essay Scoring Models 252
Sungjin Nam

Modeling Intrinsic Load: Can Item Difficulty Be a Reasonable Proxy? 260
Guher Gorgun, Minghao Cai, and Carrie Demmans Epp

Artificial Intelligence Bias on English Language Learners in Automatic Scoring 268
Shuchen Guo, Yun Wang, Jichao Yu, Xuansheng Wu, Bilgehan Ayik, Field M. Watts, Ehsan Latif, Ninghao Liu, Lei Liu, and Xiaoming Zhai

Using Self-regulated Learning Theory to Inform the Design of Educational Recommender Systems for Introductory Programming 276
Jordan Barria-Pineda, Deniz Sonmez Unal, Kamil Akhuseyinoglu, Peter Brusilovsky, and Erin Walker

StoryLab: Empowering Personalized Learning for Children Through Teacher-Guided Multimodal Story Generation 285
Zhaohui Li, Feiwen Xiao, Jiaju Lin, Xiaohan Zou, Qingxiao Zheng, and Jinjun Xiong

How Do Learners Read the Content in a Multi-source Reading-to-Write Task? – A Multimodal Study 293
Debarshi Nath, Dragan Gašević, Yizhou Fan, and Ramkumar Rajendran

Efficacy of a Computer Tutor that Models Expert Human Tutors 301
Andrew M. Olney, Sidney K. D'Mello, Natalie Person, Whitney Cade, Patrick Hays, Claire W. Dempsey, Blair Lehman, Betsy Williams, and Art Graesser

An Agentic Framework for Real-Time Pedagogical Plot Generation 309
S. Thomas Christie, Anna N. Rafferty, Zack Lee, Ella Cutler, Yu Tian, and Husni Almoubayyed

Predicting Self-regulated Learning Support Needs During Learning 317
Rick Dijkstra, Max Hinne, Eliane Segers, Josh Ring, and Inge Molenaar

Examining the Influence of Students' Personality on Self-regulatory Behaviors and Learning Outcomes During Complex Learning with MetaTutor 325
Cameron Marano, Megan Wiedbusch, Annamarie Brosnihan, Matthew Moreno, Milla Sherman, Maral Karimi, Tara Delgado, and Roger Azevedo

Leveraging a Multi-agent LLM-Based System to Educate Teachers in Hate Incidents Management 332
Ewelina Gajewska, Michal Wawer, Katarzyna Budzynska, and Jaroslaw A. Chudziak

Leveraging Lecture and Student Knowledge for Improved AI Chatbot Responses in Education 340
Atsushi Shimada

Note-Driven RAG for Learner Performance Estimation via Controlling LLM Knowledge 348
Tsubasa Minematsu and Atsushi Shimada

Exploring Student Identity in Adaptive Learning Systems Through Qualitative Data 356
Clara Belitz, HaeJin Lee, Nidhi Nasiar, Stephen E. Fancsali, Frank Stinar, Husni Almoubayyed, Steve Ritter, Ryan S. Baker, Jaclyn Ocumpaugh, and Nigel Bosch

On the Fairness of Ensemble Learning Methods in Student Dropout Prediction 364
Abdelghafour Aboukacem, Loubna Mekouar, El Houcine Bergou, Youssef Iraqi, and Ismail Berrada

Embedding Conversational Safety in AI for Education 372
Venkata Nagaraju Buddarapu and Collin Lynch

Beyond Programming Assistance Towards the Integration of Generative AI into Computer Science Education: A Scoping Review 380
Sonia Proust-Androwkha, Valéry Psyché, Revekka Kyriakoglou, Anis Boubaker, Sébastien Lallé, Lamprini Chartofylaka, Guilherme Medeiros Machado, and Nour El Mawas

Where Is the Environment in AI Ethics? Integrating Ecoliteracy Into AI Literacy 390
Daniela Inclezan and Luis I. Prádanos

Integrating Large Language Models and Machine Learning to Detect Struggle in Educational Games 398
Xiner Liu, Zhanlan Wei, Ryan S. Baker, Shari J. Metcalf, Jiayi Zhang, Amanda Barany, Stefan Slater, Luke Swanson, and David J. Gagnon

The Impact of Oversampling Techniques on the Detection of Cognitive Presence 406
Vitor Rolim, Cleon Xavier, Luiz Rodrigues, Newarney Costa, Rafael Lins, Dragan Gašević, and Rafael Mello

From Recall to Reasoning: Automated Question Generation for Deeper Math Learning Through Large Language Models 414
Yongan Yu, Alexandre Krantz, and Nikki G. Lobczowski

Enhanced Interpretable Knowledge Tracing *for* Students' Performance Prediction with Human-understandable Feature Space 423
Sein Minn and Roger Nkambou

So What? Unpacking the Complexities in Collaborative Problem Solving with AI-Augmented Sense-Making 431
Seiyon M. Lee, Hongming Li, Shan Zhang, Zirui Zhong, Ji-Eun Lee, and Anthony F. Botelho

Towards A Student-Facing Dashboard to Support Learning of Inquiry Competencies in Interactive Simulations 439
Eman Ganaiem, Tanja Käser, and Ido Roll

HuTCH: Human Teachable Concept Highlighter for Post-hoc Visual Explanations 446
Erfan Mirhaji, Nikhil Krishnaswamy, Jill Zarestky, Lisa Mason, Sarath Sreedharan, and Nathaniel Blanchard

Understanding Public School Enrollment Trends in the Northeast of Brazil: Unraveling Features Through AI Related to the Learning Process for Low-Income Students 454
Artur Cavalcante de Jesus, Leonardo Marques, Diego Carvalho do Nascimento, and Bruno Almeida Pimentel

A Study of the Influence of the Attention Mechanism on Knowledge Tracing .. 462
Yuyan Wu, Miguel Arevalillo-Herráez, and David Arnau

Similarity-Based Scoring Model for Handwritten Answers in Japanese Workbooks 470
Takahiro Saito, Hung Tuan Nguyen, Kanako Komiya, Tsunenori Ishioka, and Masaki Nakagawa

Exploring Generative AI in Learning: Neurodivergent and Disabled Students' Voices in Higher Education 478
Yun He, Elaine Chen, and Richard Fletcher

Author Index 487

Short Papers

When GPT-4 Goes to Class: Benchmarking MOOCs and Enhancing Course Design with AI

Mohamed Elhayany(✉) and Christoph Meinel

Hasso Plattner Institute, Potsdam, Germany
{mohamed.elhayany,christoph.meinel}@hpi.de

Abstract. Massive Open Online Courses offer scalable learning opportunities, yet ensuring robust assessment remains challenging. Advances in Artificial Intelligence, especially large language models like GPT-4, present new possibilities for benchmarking and refining course materials. This paper investigates how GPT-4 compares to human learners in ten introductory programming MOOCs—six Python and four Java courses—on multiple-choice and coding exercises. We analyze the model's performance at an exercise level compared to average student scores. Results show GPT-4 frequently outperforms average learners, particularly in Python courses, with statistically significant advantages observed. However, the model's advantages are less pronounced or insignificant in Java courses. Crucially, we highlight scenarios where GPT-4 struggles: exercises involving specialized libraries, multimedia content, or external contextual knowledge. These insights assist course designers in identifying ambiguous instructions and context-dependent tasks, enabling targeted course design improvements. Such refinements ensure materials better meet human learners' needs while anticipating AI advancements, thus sustaining MOOC quality and fairness.

Keywords: MOOCs · Digital Learning · GPT-4 · Course Design · Benchmarking

1 Introduction

Massive Open Online Courses (MOOCs) have become integral to today's educational landscape, offering scalable learning opportunities for global audiences [3]. As MOOCs expand, educators face challenges in maintaining course quality and effectively supporting diverse learners [19]. Artificial Intelligence in Education (AIED) offers promising solutions to enhance learner experiences, analyze behavior at scale, and offer adaptive support [12,30].

Large language models (LLMs) have gained attention for their potential to support or evaluate student learning in online courses [13,17]. These models can generate responses to tasks, serving as benchmarks to reveal instructional ambiguities or content difficulty [2,6]. By comparing LLM performance with

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 3–10, 2025.
https://doi.org/10.1007/978-3-031-98462-4_1

human learners, educators can better understand strengths and limitations in course designs.

While LLMs excel in structured reasoning tasks, they struggle with contextual understanding, visual cues, or external libraries [31]. Identifying these limitations helps educators recognize areas needing instructional improvement [1]. Continuous benchmarking with updated AI models ensures courses remain robust against AI-based shortcuts [17].

This paper compares GPT-4 performance with human learners in introductory Python and Java MOOCs, specifically addressing these questions:

RQ1: How does GPT-4 performance compare to average learners across introductory MOOCs?
RQ2: On which exercise types does GPT-4 underperform, and why?
RQ3: How can insights from GPT-4 guide improvements in course materials?

Addressing these questions contributes to AI-assisted course design literature, highlighting areas where AI struggles and guiding the creation of clearer, more effective MOOCs.

2 Related Work

The integration of Artificial Intelligence (AI) into educational contexts has deep historical roots, originating with early intelligent tutoring systems (ITS) designed to personalize learner guidance [33,35]. These systems evolved from simple rule-based approaches to sophisticated frameworks leveraging statistical models, machine learning, and natural language processing [1,11]. Recent advancements, particularly large language models (LLMs) and deep learning, have enabled richer, human-like interactions in both synchronous and asynchronous learning environments [12].

AI in education now encompasses capabilities such as identifying learner misconceptions, predicting performance, detecting engagement patterns, and personalizing learning interventions [7,20–22,29,30,32]. Additionally, AI automates administrative tasks [25], supports language learning [36], and enables just-in-time interventions at scale [8,24].

MOOCs represent a significant domain for AI application, generating vast amounts of large-scale learner interaction data [15,27]. Analyzing these data streams with AI and educational data mining techniques allows educators to understand learner behaviors, refine course materials, and optimize content delivery [34].

LLMs can simulate various educational roles, including tutors and peer support [13,17], demonstrating impressive capabilities in natural language understanding, code generation, and problem-solving [4,5]. However, these models primarily excel at pattern recognition and statistical associations, often struggling with tasks involving visual interpretation or external contextual knowledge [16,26]. Specifically, they may falter on tasks requiring specialized libraries, ambiguous instructions, or applying course-specific meta-knowledge [18].

Traditionally, improvements in educational design have relied on manual analysis of quiz results, forums, and surveys [10,14]. While valuable, these approaches can be time-consuming and limited. AI-driven analytics streamline and enhance this process, quickly identifying issues in assessment design or content clarity [6,28].

The evolution from ITS to sophisticated LLMs presents unique opportunities for educational improvement. Examining AI performance alongside human learners reveals instructional ambiguities and subtle design flaws, guiding educators to create clearer, more effective course materials and assessments, thus advancing the field of AIED.

3 Methodology

The courses analyzed in this study were hosted on openHPI, an online platform offering interactive educational content through weekly-released modules. These introductory MOOCs focused on fundamental programming concepts in Python and Java, featuring multiple-choice questions (MCQs), coding tasks, and debugging or logic-based problems. Learners completed weekly assignments and practical exercises, earning points toward a Record of Achievement or Certificate of Participation upon meeting course requirements. The platform's built-in programming environment further simplified participation by eliminating complex setup procedures [9].

Data were collected from ten introductory programming MOOCs—six Python and four Java courses—including enrollment figures and average exercise scores representing aggregate learner performance. To benchmark human learners against AI, we employed GPT-4 [23], specifically the *gpt-4-0613* snapshot, a state-of-the-art large language model recognized for advanced reasoning and coding capabilities [4,5,13,17]. GPT-4 received identical instructions and prompts as students, submitting code solutions directly to the platform's automated grader. Where multiple attempts were allowed, GPT-4 was given the same opportunities and standard feedback provided to human learners, without additional human intervention or clarifications.

Although GPT-4 typically responded quickly, any solution taking significantly longer than the average student completion time was marked as a failure. Performance analysis occurred at the exercise level, comparing mean scores, standard deviations, and correlations between GPT-4 and human learners. Cohen's d was calculated to interpret effect sizes, and paired t-tests (¡.05) assessed statistical significance. This analysis identified scenarios where GPT-4 either underperformed or excelled compared to human participants.

By employing these methods, we aimed to understand GPT-4's relative performance and identify specific task types—such as those requiring contextual understanding, visual interpretation, or specialized libraries—that posed challenges to the model. These insights can guide MOOC designers in refining course materials and assessments, ensuring meaningful challenges and genuine skill development for both human learners and evolving AI models.

4 Results

We present a comparative analysis of GPT-4 and human learner performance across Python and Java MOOCs, focusing on average scores, statistical significance, and effect sizes.

4.1 Python Courses

Figure 1 (left) compares GPT-4 scores with average student scores on Python course exercises. Points above the dashed line indicate superior GPT-4 performance. Overall, GPT-4 frequently outperformed average learners, showing substantial advantages. However, GPT-4 struggled notably on exercises involving image interpretation or external libraries. Table 1 (left) summarizes average differences, effect sizes, and statistical significance.

Most Python courses showed significant GPT-4 advantages with medium to large effect sizes. Only one course (Python 3) did not reach statistical significance.

4.2 Java Courses

Figure 1 (right) shows GPT-4 versus student performance in Java courses. The results appear mixed, with GPT-4 performing similarly or slightly above the student average. Table 1 (right) summarizes statistical outcomes.

GPT-4's performance advantages in Java courses were generally modest or statistically insignificant, with smaller effect sizes.

4.3 Summary of Findings

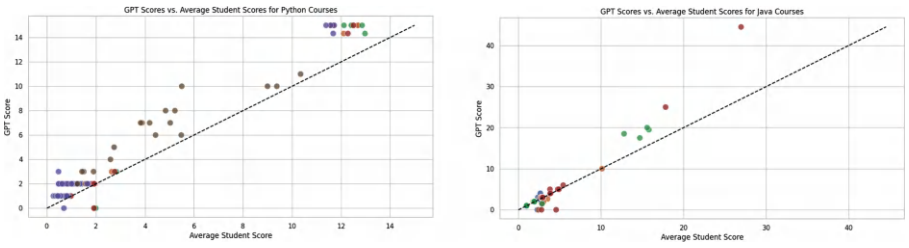


Fig. 1. GPT-4 vs. Average Student Scores. Left: Python courses; Right: Java courses. Each point represents an exercise; dashed lines indicate equal performance.

GPT-4 generally outperformed human learners significantly in Python courses, indicating strong proficiency in standard coding tasks. Java courses

showed more modest differences, often without statistical significance. Exercises requiring visual interpretation, external context, or specialized libraries posed consistent challenges for GPT-4. These results highlight GPT-4's potential and limitations as a benchmarking tool, emphasizing its weaknesses in contextual understanding and multimodal tasks compared to human learners.

Table 1. GPT-4 vs. Students: Python and Java MOOCs

Python Courses

Course	Diff.	d	p-val.	Avg. learners
Python 1	+0.18	1.66	2.9×10^{-15}	2340
Python 2	+0.21	0.32	.00063	3082
Python 3	+0.16	0.24	.067	606
Python 4	+0.21	0.27	.040	1655
Python 5	+1.01	1.27	8.4×10^{-14}	1377
Python 6	+1.92	1.69	1.5×10^{-6}	2457

Java Courses

Course	Diff.	d	p-val.	Avg. learners
Java 1	+0.06	0.12	.394	3674
Java 2	0.00	0.00	.985	2653
Java 3	+0.27	0.27	.022	3485
Java 4	+1.29	0.26	.312	766

5 Discussion

This discussion revisits our research questions, interpreting the comparative performance of GPT-4 and human learners, exploring implications for course design, acknowledging limitations, and suggesting directions for future research.

5.1 Addressing RQ1: Comparative Performance Across MOOCs

Our results indicate GPT-4 generally outperformed average learners, especially in Python courses, often significantly and with notable effect sizes. In contrast, the model's advantages in Java courses were more modest and sometimes not significant. This disparity might reflect differences in course design, exercise complexity, or GPT-4's stronger internal representation of Python tasks. Although GPT-4 effectively tackled standard coding problems, our findings caution against broadly interpreting AI's success without considering its specific weaknesses.

5.2 Addressing RQ2: Identifying Exercise Types that Challenge GPT-4

GPT-4 struggled notably with tasks requiring contextual knowledge not explicitly provided in textual prompts, such as exercises involving external libraries or image-based code interpretation. Unlike human learners who benefit from multimodal resources (videos, diagrams, and contextual explanations), GPT-4's reliance solely on text limited its performance. These observations highlight AI's ongoing limitations in higher-order cognitive and multimodal reasoning, distinguishing clearly between statistical language modeling strengths and uniquely human interpretive capabilities.

5.3 Addressing RQ3: Using Insights to Improve Course Materials

Our analysis provides three actionable insights for course designers:

1. **Clarifying Ambiguous Instructions:** Exercises problematic for GPT-4 may also confuse students. Identifying these allows instructors to improve clarity, add context, and enhance explanatory materials.
2. **Encouraging Skill Transfer Over Pattern Matching:** GPT-4's proficiency in common coding patterns suggests assessments should demand deeper conceptual reasoning and adaptability to novel scenarios, discouraging superficial pattern reproduction.
3. **Leveraging AI as a Diagnostic Tool:** GPT-4's failures and successes can inform instructors about whether student challenges stem from instructional clarity or deeper conceptual misunderstandings, facilitating iterative refinement of course materials.

5.4 Limitations and Future Work

Our study has several limitations. GPT-4 was evaluated without access to the multimodal resources available to human learners, reflecting only how an isolated AI performs against tasks designed for richer instructional contexts. As AI capabilities evolve, identified challenges may not persist, necessitating ongoing retesting and iterative refinement of educational content.

Future research should explore more advanced or specialized MOOCs to determine if similar performance patterns persist across more complex educational contexts.

6 Conclusion

This study compared GPT-4's performance with human learners across introductory Python and Java MOOCs. GPT-4 frequently outperformed average learners in Python courses and performed comparably in Java, though it struggled with exercises involving contextual or multimodal knowledge and specialized libraries. These findings emphasize GPT-4's strengths in standard coding tasks and highlight its limitations in more context-dependent, cognitively demanding activities.

GPT-4's performance serves as a valuable diagnostic tool for course designers, pinpointing ambiguous instructions or areas overly dependent on external contexts. Conversely, exercises where GPT-4 excels can help identify topics requiring clearer instructional scaffolding for students. Continuous, iterative refinement based on these insights will ensure educational content remains effective and equitable as AI capabilities advance.

References

1. Aleven, V.A., Koedinger, K.R.: An effective metacognitive strategy: learning by doing and explaining with a computer-based cognitive tutor. Cogn. Sci. **26**(2), 147–179 (2002)

2. Ani, A., Khor, E.T.: Development and evaluation of predictive models for predicting students performance in MOOCs. Educ. Inf. Technol. **29**(11), 13905–13928 (2024)
3. Baturay, M.H.: An overview of the world of MOOCs. Procedia. Soc. Behav. Sci. **174**, 427–433 (2015). https://doi.org/10.1016/j.sbspro.2015.01.685, https://www.sciencedirect.com/science/article/pii/S1877042815007363
4. Bubeck, S., et al.: Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712 (2023)
5. Chen, M., et al.: Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021)
6. Crossley, S., Paquette, L., Dascalu, M., McNamara, D.S., Baker, R.S.: Combining click-stream data with NLP tools to better understand MOOC completion. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pp. 6–14 (2016)
7. Dewan, M., Murshed, M., Lin, F.: Engagement detection in online learning: a review. Smart Learn. Environ. **6**(1), 1–20 (2019). https://doi.org/10.1186/s40561-018-0080-z
8. Dutt, A., Ismail, M.A., Herawan, T.: A systematic review on educational data mining. IEEE Access **5**, 15991–16005 (2017)
9. Elhayany, M., Meinel, C.: Towards automated code assessment with open Jupyter in MOOCs. In: Proceedings of the Tenth ACM Conference on Learning@ Scale, pp. 321–325 (2023)
10. Gašević, D., Dawson, S., Siemens, G.: Let's not forget: learning analytics are about learning. TechTrends **59**, 64–71 (2015)
11. Heffernan, N.T., Heffernan, C.L.: The assistments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. Int. J. Artif. Intell. Educ. **24**, 470–497 (2014)
12. Holmes, W., Bialik, M., Fadel, C.: Artificial intelligence in education promises and implications for teaching and learning. Center for Curriculum Redesign (2019)
13. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
14. Kay, J., Reimann, P., Diebold, E., Kummerfeld, B.: MOOCs: so many learners, so much potential. IEEE Intell. Syst. **28**(3), 70–77 (2013)
15. Kizilcec, R.F., Piech, C., Schneider, E.: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 170–179 (2013)
16. Krishnamurthy, A., Harris, K., Foster, D.J., Zhang, C., Slivkins, A.: Can large language models explore in-context? arXiv preprint arXiv:2403.15371 (2024)
17. Leiter, C., et al.: ChatGPT: a meta-analysis after 2.5 months. Mach. Learn. with Applications **16**, 100541 (2024)
18. Ma, L.: Construction of steam personalized online learning system for artificial intelligence. In: 2024 International Conference on Integrated Circuits and Communication Systems (ICICACS), pp. 1–5. IEEE (2024)
19. Macleod, H., Haywood, J., Woodgate, A., Alkhatnai, M.: Emerging patterns in MOOCs: learners, course designs and directions. TechTrends **59**(1), 56–63 (2014). https://doi.org/10.1007/s11528-014-0821-y
20. Mohamed, A., Ali, M., Ahmed, S., Hani, N., Hisham, M., Mahmoud, M.: Detecting student disengagement in online classes using deep learning: a review. arXiv preprint arXiv:2411.10464 (2024)

21. Nabizadeh, A.H., Leal, J.P., Rafsanjani, H.N., Shah, R.R.: Learning path personalization and recommendation methods: a survey of the state-of-the-art. Expert Syst. Appl. **159**, 113596 (2020)
22. Ni, L., et al.: Enhancing student performance prediction on learner sourced questions with SGNN-LLM synergy. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 23232–23240 (2024)
23. OpenAI: Gpt-4 technical report (2023). https://openai.com/research/gpt-4
24. Papadakis, S., Kalogiannakis, M.: Education, development and intervention (2024)
25. Parycek, P., Schmid, V., Novak, A.S.: Artificial intelligence (AI) and automation in administrative procedures: Potentials, limitations, and framework conditions. J. Knowl. Econ. **15**(2), 8390–8415 (2024)
26. Pi, R., Yao, L., Gao, J., Zhang, J., Zhang, T.: PerceptionGPT: effectively fusing visual perception into LLM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27124–27133 (2024)
27. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D.: Tuned models of peer assessment in moocs. arXiv preprint arXiv:1307.2579 (2013)
28. Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. Artif. Intell. Rev. **55**(3), 2495–2527 (2022)
29. Rincón-Flores, E.G., López-Camacho, E., Mena, J., López, O.O.: Predicting academic performance with artificial intelligence (AI), a new tool for teachers and students. In: 2020 IEEE Global Engineering Education Conference (EDUCON), pp. 1049–1054. IEEE (2020)
30. Roll, I., Wylie, R.: Evolution and revolution in artificial intelligence in education. Int. J. Artif. Intell. Educ. **26**, 582–599 (2016)
31. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson (2016)
32. Smart, F., Bos, N.D., Bos, J.T.: Can large language models recognize and respond to student misconceptions? In: Sottilare, R.A., Schwarz, J. (eds) Adaptive Instructional Systems. HCII 2024. LNCS, vol. 14727, pp. 288–299. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-60609-0_21
33. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. **46**(4), 197–221 (2011)
34. Wen, M., Yang, D., Rose, C.: Sentiment analysis in MOOC discussion forums: what does it tell us? In: Educational data mining 2014. Citeseer (2014)
35. Woolf, B.P.: Building Intelligent Interactive Tutors: Student-centered Strategies For Revolutionizing E-learning. Morgan Kaufmann (2010)
36. Wu, Y., et al.: See what i'm saying? Comparing intelligent personal assistant use for native and non-native language speakers. In: 22nd International Conference on Human-computer Interaction with Mobile Devices and Services, pp. 1–9 (2020)

CogGen: A Learner-Centered Generative AI Architecture for Intelligent Tutoring with Programming Videos

Wengxi Li[1], Roy Pea[2], Nick Haber[2], and Hariharan Subramonyam[2(✉)]

[1] City University of Hong Kong, Hong Kong SAR, China
wengxili@cityu.edu.hk
[2] Stanford University, Stanford, CA, USA
{roypea,nhaber,harihars}@stanford.edu

Abstract. We introduce CogGen, a learner-centered AI architecture that transforms programming videos into interactive, adaptive learning experiences by integrating student modeling with generative AI tutoring based on the Cognitive Apprenticeship framework. The architecture consists of three components: (1) video segmentation by learning goals, (2) a conversational tutoring engine applying Cognitive Apprenticeship strategies, and (3) a student model using Bayesian Knowledge Tracing to adapt instruction. Our technical evaluation demonstrates effective video segmentation accuracy and strong pedagogical alignment across knowledge, method, action, and interaction layers. Ablation studies confirm the necessity of each component in generating effective guidance. This work advances AI-powered tutoring by bridging structured student modeling with interactive AI conversations, offering a scalable approach to enhancing video-based programming education.

Keywords: Cognitive Apprenticeship · Large Langauge Models · Student Modeling · Conversational Agents

1 Introduction

Video has emerged as a preferred medium for learning programming. According to a recent Stack Overflow survey, developers who learned through "How-to-videos" and "Video-based E-Courses" accounted for 54.2% and 49.9%, respectively[1]. Regardless, the most effective learning paradigm for programming videos is *learning by doing* [4]. However, practicing coding tasks while watching videos is difficult for novices, requiring clear instructions and immediate feedback [1].

Learning to program requires integrating multiple skills: understanding syntax, grasping programming concepts, and applying problem-solving strategies [9]. *Instructional Design Principles* guide how these learning experiences should be structured [7]. The Cognitive Apprenticeship (*CogApp*) framework

[1] https://survey.stackoverflow.co/2024/developer-profile#learning-to-code.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 11–18, 2025.
https://doi.org/10.1007/978-3-031-98462-4_2

addresses these needs by making expert programmers' thought processes visible to novices through six teaching moves: Modeling, Coaching, Scaffolding, Articulation, Reflection, and Exploration [6]. This framework emphasizes guided practice with gradual transfer of responsibility to learners, effectively managing cognitive load through step-by-step instruction [2,14,19].

While large language models (LLMs) have shown promise in programming education [10,11], they present several challenges: they tend toward verbose responses [5,21], struggle to maintain focus in multi-turn conversations [16,18], and lack structured approaches for step-by-step teaching [15,17].

To address these gaps, we developed *CogGen*, a systematic architecture that ensures controlled and relevant message generation tailored to students' specific learning needs based on the *CogApp* framework. *CogGen* operates by first segmenting programming videos into distinct learning goals, then selecting appropriate teaching methods from the *CogGen* framework based on an internal student model for each learning goal. In summary, our key contributions include:

1. *CogGen*: A pipeline for generating learning conversations, using a domain-specific language to organize *CogApp* methods, enabling LLMs to deliver guidance that is both consistent and personalized.
2. Technical evaluation demonstrating *CogGen*'s effectiveness in generating structured and controlled instruction via LLMs, with evidence that each component plays a necessary role in achieving this performance.

2 *CogGen* Architecture

CogGen implements a set of LLM prompting techniques that yield conversational utterances grounded in the *CogApp* framework. The architecture consists of (1) a video segmentation module that *slices* long videos based on learning goals using a prompt chaining approach, (2) a DSL generator that produces the *context* for the LLM to generate utterances in interaction with the learner, and (3) a student model to capture the student's learning progress and inform appropriate mentor *moves*. Here, we provide technical details about each module and data representations to make the generated conversations *consistent* and *controllable*. We use few-shot prompting with GPT-4 (temperature=0.3) for all the modules.

2.1 Video Segmentation by Learning Goals

When segmenting video transcripts, small references to unrelated goals can cause misclassification. Our approach minimizes this through a three-step process: (1) *Summarize* key points for each learning goal; (2) *Retrieve* transcript sentences aligning with these summaries; (3) *Rearrange* segments based on timestamps.

2.2 Instructional Prompt Generator

For each video segment, we generate a teaching plan that aligns with the *CogApp* framework. Our prompting pipeline follows three consecutive steps:

Table 1. Representations of the two types of knowledge in *CogGen*. The bold parts are later used in the student model.

Knowledge	Representation	Example
	Concept Related	
Declarative	[Subject] + [verb phrase] + that + [independent clause].	*The median income by college major shows that majors earn a median income of over $30K right out of college.*
Procedural	To achieve/understand + [specific goal/outcome] + one must + **[actions/processes]** + [additional details] + considering/using + [relevant factors/tools].	*To understand the distribution of earnings by college major, one must **examine the histogram and identify overall trend or extreme values**, considering whether high earnings are due to the field's financial reward.*
	Programming Related	
Declarative	The task is + [final goal] + using + [general method/tool] + and + [additional method/technique for enhancement].	*The task is comparing the distribution of median earnings across different major categories using a box plot and adjusting the visualization for better readability and interpretation.*
Procedural	To achieve + **[specific goal]** + one must + **[action/verb] + [specific tool/method]** + on + [object/target] + because + [reason/purpose].	*To **achieve an ordered factor level** based on the 'Median', one must **use 'fct_reorder'** on 'Major_category', making it easier to compare distributions.*

First, we use few-shot prompts to extract procedural and declarative knowledge from the segmented transcripts using standardized formats (Fig. 1A and Table 1). For example, procedural knowledge follows patterns like "To achieve [specific goal] one must [action/verb] [specific tool/method] on [object/target] because [reason/purpose]." This structured approach enables mapping to the student model and ensures consistent generation of conversational messages.

Second, we apply three principles to select proper teaching methods [6]: (1) global before local skills, prioritizing conceptual overviews before detailed tasks; (2) increasing complexity, adapting instruction to student mastery levels; and (3) increasing diversity, varying task types to maintain engagement (Fig. 1B&C).

Finally, we generate a domain-specific language (DSL) that integrates knowledge summaries, pedagogical moves, and context to produce structured interactions (Fig. 1D&E). Each move (e.g., Modeling, Scaffolding) is mapped to specific actions and interaction types, which are used to generate *prompts* that guide the conversation. The prompts are stored in a queue, the head of which is used to produce the next message. This approach ensures interactions remain focused on the video content while following the intended pedagogical sequence.

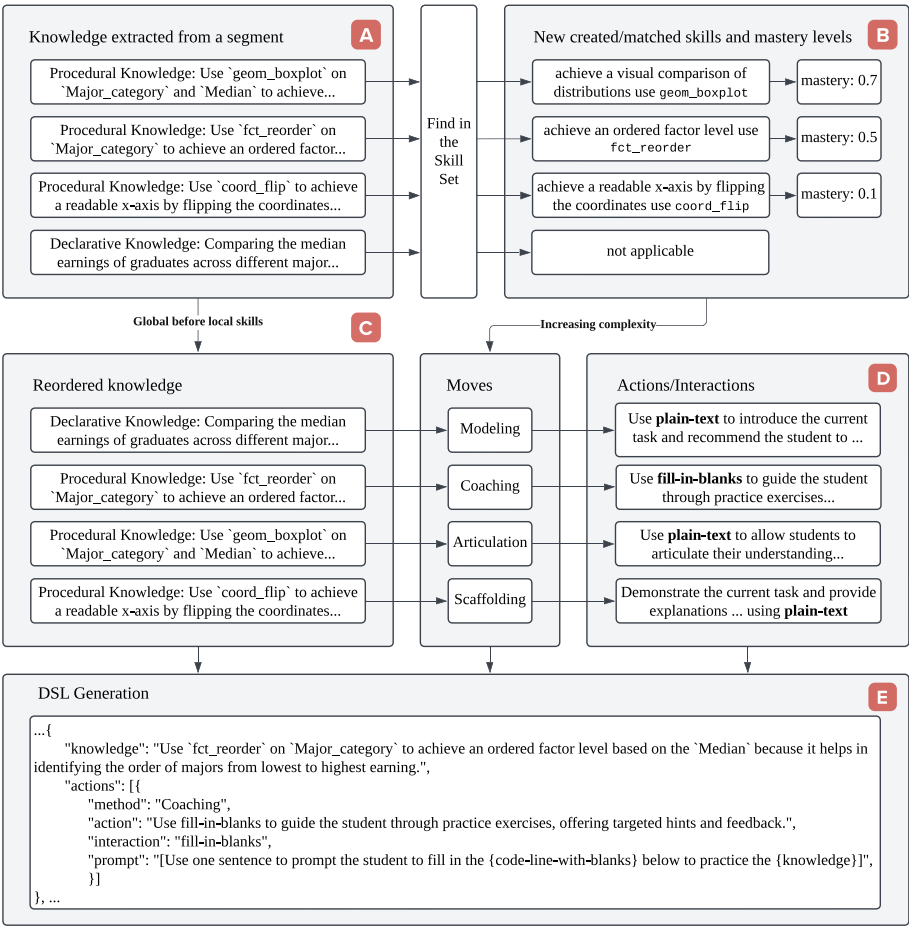


Fig. 1. Details of the *CogGen* architecture: (A) extracting procedural knowledge and declarative knowledge from the video segment, (B) the knowledge is mapped to skills and mastery levels, (C) using the principles to reorder the knowledge and decide moves for each knowledge, (D) moves are mapped to actions and interactions, (E) knowledge, moves, actions, interactions, and prompts are integrated to form the DSL.

2.3 Student Performance Monitoring

To adapt instruction based on student performance, we implement a Bayesian Knowledge Tracing (BKT) framework that tracks skill proficiency [3,13]. Each knowledge component extracted from video segments is translated into a corresponding student skill (e.g., "achieve an ordered factor level use `fct_reorder`").

Skill proficiency levels are initialized based on the student's estimated mastery of the current learning goal (default 0.1). As students interact with the system, BKT parameters are dynamically updated to reflect evolving proficiency. When students encounter previously practiced knowledge, we use semantic simi-

larity [12] to identify relevant skills and update their parameters. The parameters are stored between learning sessions to maintain continuity across videos.

This integrated approach allows CogGen to select appropriate teaching methods based on current student mastery, providing scaffolding for low-mastery skills and encouraging articulation for higher-mastery skills, creating a personalized learning experience that adapts to individual student needs.

3 Evaluation

To evaluate *CogGen*'s effectiveness across diverse learning scenarios, we selected three distinct programming video lesson topics: Exploratory Data Analysis (EDA), Machine Learning (ML), and Game Development. We evaluated the architecture using three metrics: (1) segmentation accuracy, (2) controllability of generated content, and (3) component importance through ablation studies.

3.1 Procedure

Two experts with experience in educational video analysis served as annotators for segmenting videos according to learning goals. For segmentation evaluation, we compared timestamps between manually labeled segments and those generated by LLM, using a five-second error threshold to account for differences between continuous video viewing and discrete transcript processing.

For the controllability evaluation, we created a hierarchical classification scheme based on the prompt pipeline. The system-generated data were extracted from DSL files, which defined expert-specified teaching methods. Three domain experts independently labeled 277 dialogue messages in six videos (124 EDA, 89 ML, and 64 Game Development utterances). We compared labeled data with system-generated data using precision, recall, and F1-score metrics.

To evaluate the component importance of *CogGen*, we compared four conditions: (1) *Baseline*, unconstrained generation using only video segments; (2) *Knowledge-Only*, which summarizes segment knowledge without *CogApp* theory; (3) *Method-Only*, which uses *CogApp* theory without specific knowledge awareness; and (4) *Full* condition with the complete pipeline. Expert evaluators ranked the results in the dimensions of credibility, validity, and interactivity. The agreement between the evaluators was moderate (Spearman $\rho = 0.71$), and we calculated a TrueSkill score [8] for each ablation, performed Kruskal-Wallis tests to determine the general differences between ablations, and applied Dunn's post hoc test [20] to isolate which specific condition differed from another.

3.2 Results

Segmentation Performance: Video transcript segmentation achieved 76.9% accuracy within the five-second threshold, providing acceptable performance for educational content. The accuracy decreased with longer videos, suggesting pre-segmentation into shorter clips (10–12 minutes) could improve results.

Table 2. Comparison of the intent performance across three video topics: exploratory data analysis (EDA), machine learning (ML), and game development (Game).

Topic	Knowledge			Method			Action			Interaction		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Total	0.791	0.787	0.789	0.814	0.809	0.807	0.902	0.895	0.896	0.970	0.968	0.968
EDA	0.810	0.806	0.808	0.849	0.815	0.818	0.896	0.871	0.872	0.984	0.984	0.984
ML	0.795	0.798	0.796	0.825	0.809	0.807	0.900	0.899	0.899	0.956	0.944	0.947
Game	0.827	0.781	0.792	0.808	0.813	0.798	0.954	0.953	0.953	0.973	0.969	0.969

Controllability: As shown in Table 2, *CogGen* demonstrated strong alignment between intended and generated content across all dimensions. Knowledge extraction precision and recall were 0.791 and 0.787 respectively, indicating *CogGen* accurately generates messages according to specified knowledge without missing important content. Method alignment showed consistent performance across topics, with EDA videos performing slightly better. Action and Interaction components showed progressively stronger performance, demonstrating *CogGen*'s ability to reliably convert specified interactions into conversational utterances.

The ascending performance pattern (Knowledge → Method → Action → Interaction) reflects the natural refinement process as the system moves from general knowledge to specific interactions, leveraging work done in previous phases.

Component Importance: The full condition has the highest TrueSkill score ($\mu = 30.14, \sigma = 1.60$), suggesting that combining both knowledge induction and method planning is the most effective strategy. Knowledge-Only ($\mu = 26.11, \sigma = 1.40$) outperformed Method-Only ($\mu = 23.89, \sigma = 1.40$). This suggests that without knowledge induction, the method-only approach may fail to provide precise guidance tailored to the video content, potentially leading to over-scaffolding simple concepts or under-scaffolding more complex ones. In contrast, while the knowledge-only condition has a higher score than the method-only approach, it lacks the adaptive capability provided by method planning, making it less responsive to the student's mastery level. The baseline condition performed worst ($\mu = 19.86, \sigma = 1.60$).

Kruskal-Wallis test indicates significant differences between conditions ($p < 0.001; H = 32.86$), and Dunn's post-hoc test confirmed significant pairwise differences between all ablations ($p < 0.05$). These results demonstrate that each component in *CogGen* contributes significantly to generating effective and coherent guidance, and the full pipeline has the best results than other conditions.

4 Discussion and Conclusion

Our research demonstrates *CogGen*'s effectiveness in enhancing programming education through interactive video-based tutoring. The evaluation results

confirm that *CogGen* successfully integrates learning science principles from the *CogApp* framework to generate targeted pedagogical interventions. The architecture's ability to segment videos by learning goals, maintain controllability across knowledge and pedagogical dimensions, and adapt instruction based on student performance addresses key challenges in LLM-based tutoring.

CogGen shows particular promise in formal classroom settings, where instructors can leverage its controllable parameters to create customized learning tasks within video-based lessons. The architecture's strengths lie in its structured prompting approach that maintains pedagogical coherence while allowing for personalized guidance. By integrating BKT-based student modeling with generative AI, *CogGen* represents a significant step toward scalable, adaptive programming education.

However, our current implementation has limitations that future work should address. While *CogGen* performs well with the three video topics in our evaluation, its effectiveness across more diverse programming tutorials requires further investigation. The architecture's current design assumes videos have a modular structure where segments align with discrete learning goals—an assumption that may not hold for tutorials with overlapping content. Future versions could implement adaptive segmentation using temporal coherence and multimodal cues to better align with instructional goals.

References

1. Alario-Hoyos, C., et al.: Interactive activities: the key to learning programming with MOOCs. In: Proceedings of the European Stakeholder Summit on Experiences and Best Practices in and Around MOOCs, EMOOCS, vol. 319 (2016)
2. Anderson, L.W., Krathwohl, D.R.: A taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Complete Inc, Addison Wesley Longman (2001)
3. Badrinath, A., Wang, F., Pardos, Z.: pybkt: an accessible python library of Bayesian knowledge tracing models. International Educational Data Mining Society (2021)
4. Brown, J.S., Collins, A.M., Duguid, P.: Situated cognition and the culture of learning. Educ. Res. **18**, 32–42 (1989). https://doi.org/10.3102/0013189X018001032
5. Cheng, A.Y., et al.: Scientific and fantastical: creating immersive, culturally relevant learning experiences with augmented reality and large language models. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI 2024, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3613904.3642041
6. Collins, A., Brown, J.S., Holum, A., et al.: Cognitive apprenticeship: making thinking visible. Am. Educ. **15**(3), 6–11 (1991)
7. Gagne, R.M., Wager, W.W., Golas, K.C., Keller, J.M., Russell, J.D.: Principles of instructional design (2005)
8. Herbrich, R., Minka, T., Graepel, T.: Trueskill(TM): a Bayesian skill rating system. In: Advances in Neural Information Processing Systems 20, pp. 569–576. MIT Press (2007), https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system/

9. Linn, M.C.: Designing computer learning environments for engineering and computer science: the scaffolded knowledge integration framework. J. Sci. Educ. Technol. **4**, 103–126 (1995)
10. Liu, R., Zenke, C., Liu, C., Holmes, A., Thornton, P., Malan, D.J.: Teaching cs50 with AI: Leveraging generative artificial intelligence in computer science education. In: Proceedings of the 55th ACM Technical Symposium on Computer Science Education, vol. 1, pp. 750–756. SIGCSE 2024, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3626252.3630938
11. Ma, Q., Shen, H., Koedinger, K., Wu, S.T.: How to teach programming in the AI era? using LLMs as a teachable agent for debugging. In: Artificial Intelligence in Education: 25th International Conference, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part I, pp. 265–279. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-64302-6_19
12. NetEase Youdao, I.: Bcembedding: Bilingual and crosslingual embedding for rag (2023). https://github.com/netease-youdao/BCEmbedding
13. Pardos, Z.A., Tang, M., Anastasopoulos, I., Sheel, S.K., Zhang, E.: Oatutor: an open-source adaptive tutoring system and curated content library for learning sciences research. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI 2023, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581574
14. Pea, R.D.: The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. J. Learn. Sci. **13**(3), 423–451 (2004). https://doi.org/10.1207/s15327809jls1303_6
15. Ren, C., Pardos, Z.A., Li, Z.: Human-AI collaboration increases skill tagging speed but degrades accuracy. CoRR **abs/2403.02259** (2024)
16. Sheng, Z., Finzel, R., Lucke, M., Dufresne, S., Gini, M., Pakhomov, S.: A dialogue system for assessing activities of daily living: Improving consistency with grounded knowledge. In: Muresan, S., et al. (eds.) Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pp. 68–79. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.dialdoc-1.8
17. Sonkar, S., Liu, N., Mallick, D., Baraniuk, R.: CLASS: a design framework for building intelligent tutoring systems based on learning science principles. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 1941–1961. Association for Computational Linguistics, Singapore (2023)
18. Sreedhar, M.N., Rebedea, T., Ghosh, S., Zeng, J., Parisien, C.: CantTalkAboutThis: aligning language models to stay on topic in dialogues. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 12232–12252. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.713
19. Sweller, J.: Instructional design. In: Encyclopedia of Evolutionary Psychological Science, pp. 4159–4163. Springer (2021)
20. Upton, G., Cook, I.: A Dictionary of Statistics 3E. Oxford University Press, USA (2014)
21. Zhao, W.X., et al.: A survey of large language models (2024). https://arxiv.org/abs/2303.18223

Utilizing Log-Based and Neurophysiological Measures to Understand Engagement and Learning with Intelligent Tutoring Systems

Yushuang Liu[1(✉)], Ido Davidesco[1], Bruce McLaren[2], J. Elizabeth Richey[3], Xiaorui Xue[1], Leah Teffera[2], Hayden Stec[2], Hyosun Lee[4], Jiayi Zhang[5], Suyi Liu[1], and Elana Zion-Golumbic[6]

[1] Boston College, Chestnut Hill, MA 02467, USA
yushuang.liu@bc.edu
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] University of Pittsburgh, Pittsburgh, PA 15260, USA
[4] University of Connecticut, Storrs, CT 06269, USA
[5] University of Pennsylvania, Philadelphia, PA 19104, USA
[6] Bar-Ilan University, 5290002 Ramat Gan, Israel

Abstract. Computer-based intelligent tutoring systems (ITSs) are a key to learning in many educational scenarios, but not all students engage effectively with them. Student engagement with ITSs is typically assessed via log data (e.g., problem-solving time), which only partially captures its multidimensional nature. In the current project, we studied how students engage with an intelligent tutor using a combination of log-based, electroencephalography (EEG), and eye-tracking measures. A total of 56 high school students participated in a school-based study with a pretest-intervention-posttest design. During the intervention, students watched short videos and solved chemistry problems using an ITS. Preliminary analysis shows that log-based measures substantially improved model fit, explaining 26% more variance in posttest scores than a model using only pretest scores. Within the model, error rate and hint request rate were significant predictors, while self-reported effort and time spent on problems were not. Hint requests were negatively associated with learning outcomes, emphasizing the need to assess how students use hints and other assistive features of ITSs. Adding EEG and eye-tracking measures did not significantly improve the overall performance of the model. However, a more fine-grained, problem-by-problem analysis revealed that frontal EEG alpha power significantly predicted the number of errors on individual intervention problems. In the future, this finding may support the development of EEG-informed ITSs that can provide more personalized assistance. More generally, these results highlight the value of multimodal measures of engagement in educational technology research.

Keywords: Intelligent tutoring systems · engagement · EEG · eye-tracking

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 19–26, 2025.
https://doi.org/10.1007/978-3-031-98462-4_3

1 Introduction

Computer-based intelligent tutoring systems (ITSs) are a critical component in many educational settings, particularly in K-12, including in-person and online learning. ITSs are designed to complement traditional classroom learning by providing students with personalized guidance and support that are tailored to their learning progress and individual needs [1–3]. ITSs are designed to be engaging, but not all students learn effectively with them, possibly due to differences in engagement [4, 5].

The current study builds on the Interactive-Constructive-Active-Passive (ICAP) theoretical framework, which proposes a direct link between the level of engagement in an instructional activity and its corresponding learning outcomes [6]. According to ICAP, passive activities, such as listening to a lecture, lead only to superficial learning (e.g., recall). Conversely, constructive and interactive activities that require high levels of engagement, such as solving science problems, ultimately lead to robust learning [7]. This framework suggests that examining student engagement with ITSs may help explain differences in learning outcomes.

There is a growing consensus that engagement should be viewed as a multidimensional construct [8, 9]. Fredricks and colleagues [10] proposed three components of engagement: (1) Behavioral engagement: learners' participation in the learning process, including effort and persistence (e.g., note-taking); (2) Cognitive engagement: learners' mental investment in the learning task, such as allocation of attention; and (3) Affective engagement: feelings and attitudes about the learning task or learning context (e.g., interest, boredom). Typically, students' engagement with ITSs is measured via log-based student interaction data (e.g., time to complete a problem), but these data do not fully capture all three dimensions of engagement. For example, it is challenging to determine how students cognitively engage with assistive features of ITS - such as hints - based on log-data alone (do students read hints?).

In the current study, we incorporated a set of neurophysiological measures (Electroencephalography and eye-tracking) to capture a more comprehensive picture of how students engage with ITSs. Specifically, we focus on EEG activity in the alpha band (8–12 Hz), which has been associated with lapses of attention [11, 12] and mind-wandering [13, 14]. We also focus on gaze position as a proxy for the locus of attention [15–17]. These two neurophysiological measures could provide information on the cognitive engagement of learners [9].

The current study addresses the following research questions: RQ1: What are the relations between tutor log-based and neurophysiological measures? RQ2: How much of the variability in learning outcomes with ITS can be attributed to log-based and neurophysiological measures?

2 Methods

2.1 Participants

56 students participated voluntarily ($Mean_{age} = 15.89$ years, 52% female, 66% white), currently taking a chemistry course, from two high schools in the USA. All received permission from their caregivers, and signed an assent form, as required by our Institutional Review Board. We excluded 3 students from EEG and 12 from eye-tracking

due to poor data quality (EEG: >50% artifacts; Eye-tracking: > 30% missing or >1° calibration error).

2.2 ITS

The specific ITS platform that was used in this project is a web-based chemistry tutor that has been validated extensively in previous studies (Fig. 1) [18–20]. During tutored problems, students could request context-sensitive and on-demand hints, as well as access an on-screen calculator. As students worked through each problem, the tutor also provided automatic error messages, adaptive support, and correctness feedback.

2.3 Procedure

Students participated in a self-paced study (~3 h) in a quiet school office. They first completed a pretest (~ 50 min), consisting of four stoichiometry problems (isomorphic to the tutored problems) and three conceptual knowledge questions. Within one week, students participated in the EEG and eye-tracking session (preparation: ~ 30 min), including watching content-related videos and solving six tutored problems, each followed by a self-reported effort rating (~ 1 h). A posttest concluded the session (~ 40 min). Pre/posttests had two isomorphic forms (A and B), counterbalanced across students.

2.4 Log-Based Measures

The current paper focuses on three measures that were extracted from the log file for each tutored problem – duration on problem, number of hint requests, number of errors—as well as self-reported efforts.

2.5 Eye-Tracking

Eye-tracking data were collected using a Tobii Pro Fusion screen-based eye tracker. The ITS display was divided into eight Areas of Interest (AOIs): problem statement, hint window, calculator, four problem steps, and the final result (see Fig. 1). The current paper focuses on the proportion of time looking at each AOI.

2.6 EEG

EEG was recorded using a 32-channel mbt Smarting Pro system. Data were band-pass filtered (1–30 Hz), cleaned of noisy channels/periods, and re-referenced. Ocular/muscle artifacts were removed via ICA. Preprocessed data were epoched into 1-s intervals, tapered with a Hanning window, and transformed (4–30 Hz) via FFT. Alpha power (8–12 Hz) was normalized by total 4–30 Hz power [21], then averaged across frontal, central, and posterior ROIs. Analyses was carried using Matlab R2024b, with EEGLAB 2024.0 [22] and FieldTrip [23].

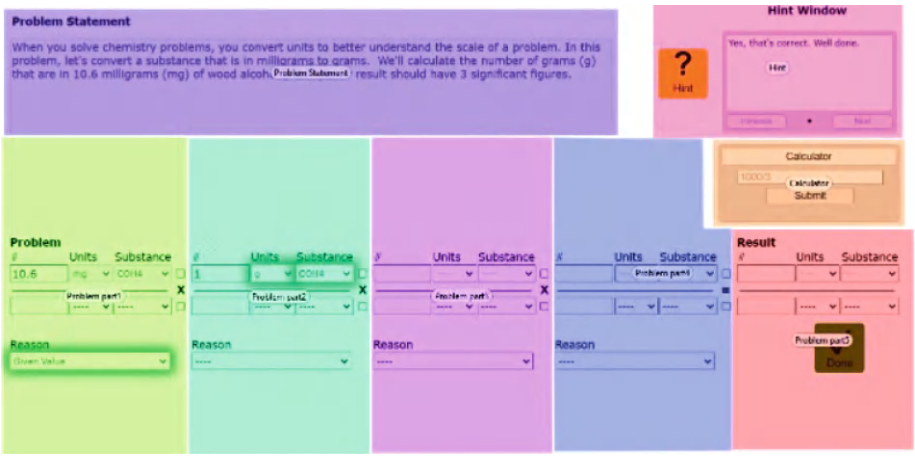


Fig. 1. A screenshot of the Chemistry Tutor with areas of interest for eye gaze analysis.

3 Results

Students' averaged 3.54 (SD = 1.22) on the pretest and averaged 4.37 (SD = 1.08) on the posttest, out of a total possible score of seven. A paired-sample t-test indicated that students' performance significantly improved ($t(55) = 5.55$, $p < .001$).

To address RQ1 ("What are the relations between tutor log-based and neurophysiological measures?"), we first assessed the correlations among log-based measures for each tutored problem. We also averaged the measures across problems and computed the correlations between the aggregate measures. Overall, self-reported effort, problem duration, number of errors, and number of hint requests were positively, significantly correlated with each other (Table 1).

Next, a series of linear regressions were used to test whether the neurophysiological measures significantly predicted each one of the log-based measures. This preliminary analysis indicated that in four out of the six tutored problems, frontal EEG activity in the alpha band significantly predicted the number of errors on problems (βs = -.91 to -.52 and ps =.001 to.04). The regression coefficients were negative, indicating that higher alpha power was associated with fewer errors. Alpha power did not significantly predict self-reported effort, duration, or number of hint requests. Among the eye-tracking measures, the only significant association across problems was between the proportion of time looking at the hint window and number of hint requests.

To address RQ2 ("How much of the variability in learning outcomes with ITS can be attributed to log-based and neurophysiological measures?"), a regression model was used to predict total posttest performance based on pretest performance. Adding the log-based measures to the model explained significantly more variance than pretest alone ($R^2_{\text{Change}} = .27$, $p < .001$). Specifically, the model predicting posttest performance with pretest explained 27.6% of the variance ($F(1, 54) = 20.58$, $p < .001$, $R^2 = .28$, $R^2_{\text{Adjusted}} = .26$), while the model with the log-based measures predicted 53.6% of the variance ($F(5, 50) = 11.56$, $p < .001$, $R^2 = .54$, $R^2_{\text{Adjusted}} = .49$). Within the extended model, pretest score ($\beta = .28$, $p = .017$) was a significant, positive predictor and total number of

errors (β = -.45, p =.018) and total number of hint requests (β = -.29, p =.034) were significant, negative predictors of posttest score. Duration on problem (β =.15, p =.41) and self-reported effort (β = -.04, p =.66) were not significant predictors of posttest performance. A preliminary analysis that included the neurophysiological measures did not improve the model (R^2_{Change} =.11, p =.47).

Table 1. Correlations among test scores, log-based measures, and neurophysiological measures

	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
1. Pretest	.53*	-.51*	-.31*	-.50*	-.03	-.11	-.05	-.04	.02	.11	.03	-.25	.02	-.12	.08
2. Posttest		-.67*	-.58*	-.55*	-.04	-.09	-.15	-.08	.01	-.24	.22	-.04	.004	-.14	.09
3. Errors			.67*	.82*	.03	.27	.27*	.20	-.08	-.03	-.34*	.13	.15	.18	-.02
4. Hints				.63*	-.05	.31*	.37*	.29*	-.08	-.01	-.43*	.05	.03	.25	.06
5. Duration					.03	.34*	.28*	.26	.04	-.11	-.31*	.19	.16	.07	-.07
6. Effort						-.15	-.03	-.06	.45*	-.14	-.35*	-.32*	.11	.22	-.17
7. Alpha frontal							.85*	.84*	-.11	.07	.21	.04	-.12	-.34*	.03
8. Alpha central								.85*	-.04	.14	.13	-.05	-.12	-.37*	.04
9. Alpha posterior									-.002	.16	.14	-.04	.05	-.38*	-.02
10. Gaze hint										-.11	-.23	-.22	.06	-.07	-.61*
11. Gaze area 1											.14	-.33*	-.08	-.25	-.07
12. Gaze area 2												.04	.04	-.39*	-.25
13. Gaze area 3													.02	-.09	-.04
14. Gaze area 4														.14	-.32*
15. Gaze solution															-.15
16. Gaze stat.															

Note. Gaze = proportion of time spent looking at each AOI. Stat. = Problem Statement. (*) $p < .05$.

4 Discussion

In this study, we examined how log-based ITS measures relate to neurophysiological measures of engagement (RQ1) and how each set of measures predict learning outcomes (RQ2). Consistent with prior ITS research, our analysis indicates that log-based measures are strongly correlated among themselves [24]. Among the EEG measures, our preliminary analysis indicates that frontal EEG activity in the alpha band was significantly associated with numbers of errors. Previous studies reported that increases in alpha power are associated with lapses of attention and disengagement [21, 25]. However, most of these studies have focused on alpha activity measured in posterior areas of the brain. Our results suggest that increased frontal alpha activity is associated with solving problems with fewer errors. This aligns with recent findings linking frontal alpha activity to internally demanding processes and deeper understanding [26, 27].

Whereas EEG activity in the alpha band significantly predicted error numbers, it was not predictive of self-reported effort, duration, or number of hint requests. This is

surprising, considering prior research indicates that alpha activity is a marker of task difficulty and cognitive load [28]. It is possible, however, that when students report their effort level after a problem has been solved, their reports do not accurately reflect their cognitive load during problem solving [29]. Similarly, hint requests and problem duration might be influenced by 'gaming the system', not just by cognitive load [30]. Therefore, future analyses will incorporate gaming detectors.

Our analysis of eye-tracking data has, so far, yielded just one significant result: The proportion of time students spent looking at the hint window was predictive of hint requests. Based on prior research, other eye-tracking metrics (e.g., transitions, fixations, gaze dispersion) may yield more insight [31–33].

With regard to predicting learning outcomes with ITS, our results indicate that other than pretest performance, the only significant predictors of posttest performance were hint requests and number of errors. This replicates a common finding in ITS research: hint requests and error rates during learning predict posttest performance when accounting for pretest [24]. Both log-based measures were negatively associated with posttest performance: students who made more errors and requested more hints performed more poorly on the posttest. Indeed, numerous hint requests are known to be an indicator of 'gaming the system' [30]. Interestingly, self-reported effort did not significantly predict posttest performance. This is surprising given prior research on cognitive load and learning outcomes [34]. However, as previously mentioned, it is possible that self-reported effort does not accurately capture cognitive load during problem solving.

Finally, our analysis suggests that incorporating EEG and eye-tracking measures did not improve the prediction of posttest performance. This analysis is preliminary as it only included one EEG measure (alpha power) and one eye-tracking measure (proportion of looking time per AOI). Consistent with our findings, recent research suggests that alpha power is a sensitive measure of moment-to-moment changes in engagement, not overall learning [21]. It is very possible that other EEG measures would be more predictive of learning outcomes. One such measure might be theta power, which has been associated with memory retention [28]. Similarly, more complex eye-tracking measures might have better predictive value.

Taken together, our results indicate that EEG measures, especially frontal alpha activity, relate to how students engage with ITSs. These findings suggest that neurophysiological measures could uncover aspects of ITS engagement that are not easily detectable using log-based data. This is important because according to the ICAP framework, there is a direct link between the level of cognitive engagement and learning outcomes. While it is assumed that the interactive design of ITSs promotes cognitive engagement, engagement with ITSs varies within and across students [35]. EEG measures could potentially help assess student cognitive engagement more accurately and ultimately contribute to the design of more effective ITSs. Subsequent analyses within the current study as well as future studies should examine a wider range of EEG and eye-tracking measures and investigate the unique contribution of these measures to our understanding of student engagement and learning with ITS.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. 2141139. We thank the participating schools, students, and families, as well as M. Coburn, B. Cary, A. Pradeep, and M. Salha for their help with data collection.

References

1. Mousavinasab, E., Zarifsanaiey, N., Niakan Kalhori, S.R., Rakhshan, M., Keikha, L., Ghazi Saeedi, M.: Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. Interact. Learn. Environ. **29**, 142–163 (2021)
2. VanLehn, K.: The behavior of tutoring systems. Int. J. Artif. Intell. Educ. **16**, 227–265 (2006)
3. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. **46**, 197–221 (2011)
4. D'Mello, S., Dieterle, E., Duckworth, A.: Advanced, analytic, automated (AAA) measurement of engagement during learning. Educ. Psychol. **52**, 104–123 (2017)
5. Koedinger, K.R., Aleven, V.: Exploring the assistance dilemma in experiments with cognitive tutors. Educ. Psychol. Rev. **19**, 239–264 (2007)
6. Chi, M.T.H., Wylie, R.: The ICAP framework: Linking cognitive engagement to active learning outcomes. Educ. Psychol. **49**, 219–243 (2014)
7. Chi, M.T.H., et al.: Translating the ICAP theory of cognitive engagement into practice. Cogn. Sci. **42**, 1777–1832 (2018)
8. Schwartz, R.N., Plass, J.L.: Types of engagement in learning with games. In: Plass, J.L., Mayer, R.E. (eds.) Handbook of Game-Based Learning, pp. 53–80. MIT Press (2020)
9. D'Mello, S.K.: Improving student engagement in and with digital learning technologies. In: Pushing the Frontiers with Artificial Intelligence. Blockchain and Robots, pp. 79–104. OECD Publishing, Paris (2021)
10. Fredricks, J.A., Blumenfeld, P.C., Paris, A.H.: School engagement: potential of the concept, state of the evidence. Rev. Educ. Res. **74**, 59–109 (2004)
11. Clayton, M.S., Yeung, N., Cohen Kadosh, R.: The roles of cortical oscillations in sustained attention. Trends Cogn. Sci. **19**, 188–195 (2015)
12. Jensen, O., Mazaheri, A.: Shaping functional architecture by oscillatory alpha activity: gating by inhibition. Front. Hum. Neurosci. **4**, 186 (2010)
13. Baldwin, C.L., et al.: Detecting and quantifying mind wandering during simulated driving. Front. Hum. Neurosci. **11**, 406 (2017)
14. Compton, R.J., Gearinger, D., Wild, H.: The wandering mind oscillates: EEG alpha power is enhanced during moments of mind-wandering. Cogn. Affect. Behav. Neurosci. **19**, 1184–1191 (2019)
15. Casarotti, M., et al.: Paying attention through eye movements: A computational investigation of the premotor theory of spatial attention. J. Cogn. Neurosci. **24**, 1519–1531 (2012)
16. Risko, E.F., Kingstone, A.: Eyes wide shut: implied social presence, eye tracking and attention. Atten. Percept. Psychophys. **73**, 291–296 (2011)
17. Geisler, W.S., Cormack, L.K.: Models of overt attention. In: Liversedge, S., Gilchrist, I.D., Everling, S. (eds.) The Oxford Handbook of Eye Movements, pp. 439–454. Oxford University Press, Oxford (2011)
18. McLaren, B.M., et al.: The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments. Comput. Hum. Behav. **55**, 87–99 (2016)
19. McLaren, B.M., Lim, S.-J., Koedinger, K.R.: When is assistance helpful to learning? In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) Intelligent Tutoring Systems, pp. 677–680. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-69132-7_75
20. McLaren, B.M., Isotani, S.: When is it best to learn with all worked examples? In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) Artificial Intelligence in Education. Springer, Heidelberg, pp. 222–229 (2011). https://doi.org/10.1007/978-3-642-21869-9_30
21. Davidesco, I., et al.: Detecting fluctuations in student engagement and retention during video lectures using electroencephalography. Br. J. Educ. Technol. **54**, 1895–1916 (2023)

22. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods **134**, 9–21 (2004)
23. Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M.: FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. **2011**, Article ID 156869 (2011)
24. Aleven, V., Koedinger, K.R.: Investigations into help seeking and learning with a cognitive tutor. In: Papers of the AIED-2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments, pp. 47–58 (2001)
25. Boudewyn, M.A., Long, D.L., Swaab, T.Y.: Effects of working memory span on processing of lexical associations and congruence in spoken discourse. Front. Psychol. **4**, 60 (2013)
26. Farkish, A., et al.: Evaluating the effects of educational multimedia design principles on cognitive load using EEG signal analysis. Educ. Inf. Technol. **28**, 2827–2843 (2023)
27. Pi, Z., et al.: Learning by explaining to oneself and a peer enhances learners' theta and alpha oscillations while watching video lectures. Br. J. Educ. Technol. **52**, 659–679 (2021)
28. Antonenko, P.D.: Educational neuroscience: exploring cognitive processes that underlie learning. In: Parsons, T.D., Lin, L., Cockerham, D. (eds.) Mind, Brain and Technology. Springer, Cham, pp 27–46 (2019). https://doi.org/10.1007/978-3-030-02631-8_3
29. Davidesco, I., et al.: Neuroscience research in the classroom: Portable brain technologies in education research. Educ. Res. **50**, 649–656 (2021)
30. Paquette, L., et al.: Cross-system transfer of machine learned and knowledge engineered models of gaming the system. In: Ricci, F., Bontcheva, K., Conlan, O., Lawless, S. (eds.) User Modeling, Adaptation and Personalization. Springer, Cham, pp 183–194 (2015). https://doi.org/10.1007/978-3-319-20267-9_15
31. Kardan, S., Conati, C.: Comparing and combining eye gaze and interface actions for determining user learning. In: Carberry, S., Weibelzahl, S., Micarelli, A., Semeraro, G. (eds.) User Modeling, Adaptation, and Personalization. Springer, Berlin, Heidelberg, pp 215–227 (2013). https://doi.org/10.1007/978-3-642-38844-6_18
32. Lallé, S., et al.: The impact of student individual differences and visual attention to pedagogical agents during learning with MetaTutor. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., Du Boulay, B. (eds.) Artificial Intelligence in Education. Springer, Cham, pp 149–161 (2017). https://doi.org/10.1007/978-3-319-61425-0_13
33. Zhang, H., et al.: Wandering eyes: eye movements during mind wandering in video lectures. Appl. Cogn. Psychol. **34**, 449–464 (2020)
34. Sweller, J.: Cognitive load theory: recent theoretical advances. In: Plass, J.L., Moreno, R., Brünken, R. (eds.) Cognitive Load Theory, pp. 29–47. Cambridge University Press, Cambridge (2010)
35. Mills, C., et al.: Mind wandering during learning with an intelligent tutoring system. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) Artificial Intelligence in Education, vol 9112. LNCS. Springer, Cham, pp 267–276 (2015) . https://doi.org/10.1007/978-3-319-19773-9_27



"My Grade is Wrong!": A Contestable AI Framework for Interactive Feedback in Evaluating Student Essays

Shengxin Hong[2], Chang Cai[1], Sixuan Du[3], Haiyue Feng[4], Siyuan Liu[1], and Xiuyi Fan[1(✉)]

[1] Nanyang Technological University, Singapore, Singapore
xyfan@ntu.edu.sg
[2] Hubei University of Technology, Wuhan, China
[3] Hohai University, Nanjing, China
[4] Shenzhen university, Shenzhen, China

Abstract. Interactive feedback, where feedback flows in both directions between teacher and student, is more effective than traditional one-way feedback. However, it is often too time-consuming for widespread use in educational practice. While Large Language Models (LLMs) have potential for automating feedback, they struggle with reasoning and interaction in an interactive setting. This paper introduces CAELF, a Contestable AI Empowered LLM Framework for automating interactive feedback. CAELF allows students to query, challenge, and clarify their feedback by integrating a multi-agent system with computational argumentation. Essays are first assessed by multiple Teaching-Assistant Agents (TA Agents), and then a Teacher Agent aggregates the evaluations through formal reasoning to generate feedback and grades. A case study on 500 critical thinking essays with user studies demonstrates that CAELF significantly improves interactive feedback, enhancing the reasoning and interaction capabilities of LLMs. This approach offers a promising solution to overcoming the time and resource barriers that have limited the adoption of interactive feedback in educational settings.

Keywords: Interactive Feedback · Human-AI Interaction · LLM

1 Introduction

Feedback is one of the most powerful influences on learning and achievement [9,18]. More recently, Nicol [17] has proposed considering **Interactive Feedback** as a more timely and process-focused feedback format [21]. Nicol suggests that feedback should not merely be a one-way transmission of information from instructor to student, it should involve students actively engaging with the feedback, interpreting it, and using it as a basis for further learning and improvement. However, despite the benefits of interactive feedback, it is not commonly used

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 27–35, 2025.
https://doi.org/10.1007/978-3-031-98462-4_4

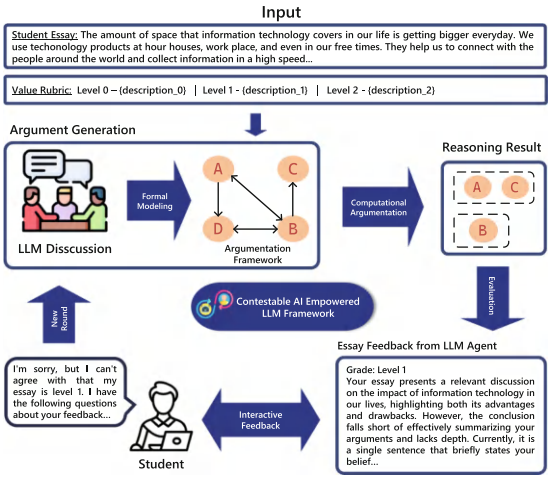


Fig. 1. Diagram of our contestable AI empowered LLM framework for interactive feedback generation (CAELF).

in practice, as it is time-consuming for teachers and requires significantly more investment in preparation and teaching delivery [11,24].

The advent of Large Language Models (LLMs) has emerged as a promising technological foundation for next generation interactive AI educational tool. However, a review of past efforts to explore LLM capabilities [16,23,25] reveals that existing LLM techniques, such as prompt engineering, fall short in meeting two critical requirements for effective interactive feedback: **Reasoning** and **Interaction**. LLMs are prone to issues of over-reliance on user input, which undermines their ability to provide reliable interactive feedback. For instance, while LLMs may initially generate correct responses, they often falter when challenged by illogical or invalid arguments from users, failing to maintain consistency.

In this work, we propose a **Contestable AI Empowered LLM Framework for Interactive Feedback Generation (CAELF)** as illustrated in Fig. 1. CAELF takes students' essays and an assessment rubric as its inputs and produces feedback as its outputs. CAELF supports interactive feedback as it allows students to enquiry and challenge its feedback and provide additional justifications as needed. To evaluate the effectiveness of CAELF, we conducted a case study on the evaluation of critical thinking essays. Using a dataset of 500 essays and a four-dimensional evaluation rubric, CAELF demonstrated significantly better performance in interaction grading accuracy and maintaining consistent evaluations despite user challenges. Additionally, in a separate human user study, we found that CAELF's feedback outperformed the baselines in terms of "factual accuracy," "self-regulation," and "suggestions for future improvement."

2 Related Work and Background

LLMs for Essay Evaluation and Feedback. LLMs have become increasingly popular in automating essay evaluation and feedback generation, reducing the manual effort traditionally required [12]. They have shown promise in automating scoring, cutting down on time and labor [4,5]. For example, [26] demonstrated that GPT-4 can evaluate short English essays with near-equal performance to modern Automatic Writing Evaluation (AWE) methods, without specific training. However, [12] highlighted LLM limitations in evaluating complex academic texts, showing a gap between LLM capabilities and the nuanced requirements of student essay evaluation.

Contestable AI and Computational Argumentation. Contestable AI asserts that models used in critical tasks like decision-making or evaluation should enable users to question, contest, and review their outputs [1,10]. [13] argue that contestable AI requires computational argumentation, allowing for dynamic explainability and the ability to adjust decisions in response to valid challenges. Computational Argumentation (CA), a branch of AI, the core idea and appeal of CA stems from the fact that solutions to controversial arguments can be computed automatically. The central concept in CA is the abstract argumentation framework (AF) [7], represented as a directed graph $\langle \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A}$ is set of arguments and $\mathcal{R}$ a set of binary attack relations over $\mathcal{A}$. With an AF, we can determine the set of arguments to be *accepted* using semantics such as the *admissibility* and *completeness*. Intuitively, an admissible set of arguments is a consistent set that can defend itself, and a complete set is the largest admissible set with respect to set inclusion. If you are interested in more technical details about CA, please read this handbook [3] of formal argumentation.

3 Framework Design and Implementation

LLM Discussion. Discussions and debates between multiple LLMs can enhance factual accuracy and reasoning skills in textual evaluation [6,14]. This debate process enables LLMs to detect inconsistencies in their analysis and effectively presents arguments and counterarguments [22]. Building on this capability, we apply role-playing techniques to extend this approach to essay evaluation. In CAELF, multiple TA agents are used to generate arguments and counterarguments through dialogue. The TA agents engage in several rounds of discussion, where they exchange responses to each other's feedback. Each agent autonomously contributes by either supporting or rebutting the others' points, continuing the debate until the set number of rounds is completed. Importantly, each TA agent is equipped with a memory function to store all previous responses in chat transcripts and assigned a specific role based on an assessment rubric, so the entire process operates without human intervention. An example of CAELF execution is illustrated in online Appendix A.

Formal Reasoning for Feedback Generation. Once the TA agents complete their discussion, the teacher agent analyzes their arguments and produces both assessment scores and feedback. To this end, the teacher agent aggregates the evaluations from the TA agents, forming a set of arguments that are then analyzed for semantic relationships (attacks). These relationships are used to construct an argumentation framework, within which formal reasoning is applied to identify coherent and non-conflicting arguments. The *complete* semantics is used, which provides criteria for consistency and comprehensiveness when evaluating arguments. In this way, the teacher agent can determine the most valid positions from the TA agents, represented by the selected set of arguments. These positions serve as knowledge-enhancing prompts that assist the teacher agent in assigning essay grades and generating summary feedback.

Interaction with User. To realize interactive feedback and AI contestability, CAELF allows students to challenge the summary feedback generated by the teacher agent through an argumentative process. When a student submits a challenge, the TA agent initiates a new round of discussion focused on the student's argument. Any new arguments raised during this discussion are incorporated into the argumentation framework, refining the logic chain of the formal reasoning process. This helps CAELF generate feedback that is both logical and human-centered. In this process, computational argumentation supports both LLM reasoning and student learning. For student learning, the dialectical argumentation process transparently illustrates the relationships between different arguments, making the feedback generated by the LLM easier to understand. This clarity allows users to see how their rebuttals influence the LLM's reasoning. As a result, interactive feedback involves learners and participants collaboratively constructing new knowledge through dialogue, promoting reflection and working toward consensus in achieving educational goals.

4 Experiment and Result

Settings. We use the critical thinking essay assessment as a case study to evaluate the general effectiveness of CAELF. Previous research has demonstrated that critical thinking skills can be developed through writing critical thinking essays [19,20]. We compiled a dataset of 500 critical thinking essays sourced from Hugging Face [8]. Based on prior research [2], we developed evaluation rubrics with four dimensions: *issues*, *evidence*, *position*, and *conclusions*. For baselines, we focus on comparing models that are publicly accessible via API. The details for each dimension and implementation are provided in the online Appendix.

To assess the performance of interactive feedback, we introduce four key metrics: 1) **Initial Accuracy (IA)**: The rate of correct initial grading before any interaction; 2) **Interaction Accuracy (Acc)**: The rate of correct grading after one round of interaction with the student; 3) **Maintain Truth (MT)** [23]: The rate of maintaining initial correct grades after being rebutted; 4) **Admit Mistake (AM)**: The rate of correcting initial errors grades after being refuted.

Table 1. Experiment results of evalution task for four dimension. Results in bold are the best performances. We also list the standard errors for each result.

Dimension	Method	IA (%)	Acc (%)	MT (%)	AM (%)
Issue	CAELF	48.40 ± 2.23	**51.00 ± 2.24**	**80.17 ± 1.78**	**57.55 ± 2.21**
	GPT-4o-mini	**55.00 ± 2.22**	43.20 ± 2.21	39.27 ± 2.18	35.18 ± 2.14
	GPT-4o	53.80 ± 2.23	47.20 ± 2.23	49.07 ± 2.23	42.45 ± 2.21
	Llama	53.20 ± 2.23	42.20 ± 2.21	31.58 ± 2.08	36.49 ± 2.15
Evidence	CAELF	**79.00 ± 1.82**	**77.00 ± 1.88**	**91.90 ± 1.22**	**39.29 ± 2.18**
	GPT-4o-mini	66.20 ± 2.11	32.40 ± 2.09	33.23 ± 2.11	18.37 ± 1.73
	GPT-4o	78.60 ± 1.83	44.20 ± 2.22	47.58 ± 2.23	14.41 ± 1.57
	Llama	55.40 ± 2.22	32.60 ± 2.10	23.10 ± 1.88	27.37 ± 1.99
Position	CAELF	67.20 ± 2.09	**68.20 ± 2.08**	**88.10 ± 1.44**	**51.14 ± 2.23**
	GPT-4o-mini	63.40 ± 2.15	43.80 ± 2.22	20.50 ± 1.81	41.62 ± 2.20
	GPT-4o	**69.60 ± 2.06**	55.20 ± 2.22	61.78 ± 2.17	31.28 ± 2.07
	Llama	47.40 ± 2.23	42.20 ± 2.21	14.77 ± 1.59	40.65 ± 2.20
Conclusion	CAELF	75.80 ± 1.92	**62.80 ± 2.16**	**75.72 ± 1.92**	22.88 ± 1.88
	GPT-4o-mini	69.60 ± 2.06	25.00 ± 1.94	13.21 ± 1.51	20.31 ± 1.80
	GPT-4o	**79.80 ± 1.80**	40.20 ± 2.19	29.07 ± 2.03	23.35 ± 1.89
	Llama	36.00 ± 2.15	28.60 ± 2.02	20.56 ± 1.81	**29.28 ± 2.04**

To assess the interactive aspect, we employ an independent ChatGPT instance to simulate a human user by presenting counterarguments based on the model's initial grade. We limit the number of interaction rounds to one to maintain natural and realistic responses.

Initial and Interaction Accuracy. Table 1 presents the accuracy results for the critical thinking essay dataset. We compared CAELF to the three baseline models under the same setup. It is important to note that we do not pursue CAELF to be SOTA in non-interactive scenarios. As conclusive findings of the initial accuracy results, although CAELF generates a lot of context in reasoning, it does not reduce its accuracy of initial grading (without interaction) in language models. However, interaction accuracy shows a catastrophic drop in the accuracy of the baseline models after one round of interactions (30% drop on average). In contrast, CAELF is minimally affected by the interaction. After interaction, CAELF achieves far better performance than the baseline models, achieving the best performance in each dimension, where CAELF's interaction accuracy is on average 40% higher than the three baseline model. This suggests that the basic LLM with direct prompts suffers from a fundamental reasoning flaw on interactive feedback task [23,25]. In contrast, we effectively mitigate this shortcoming by introducing formal reasoning and multi-agent argumenta-

tion, thus highlighting the potential of CAELF as an application in educational environments.

Maintain Truth and Admit Mistake. To evaluate the correctness of the model's responses and the effectiveness of interactive feedback, we measured the maintain truth rate and admit mistakes rate, as shown in Table 1. When assessing the model's ability to maintain consistency, CAELF achieved a success rate of 80%-90%, while baseline model had success rates below 40% on most dimensions. These results suggest that basic LLMs are highly susceptible to user interference during interactive feedback. We also assessed the models' ability to admit mistakes, where CAELF outperformed the baseline models by 10%-20% in most cases. This improvement indicates that CAELF's strong performance in maintaining accuracy is not due to over-defending its responses but rather its ability to correctly identify errors in previous feedback. In contrast, the baseline models using direct prompts relied on surface-level patterns from initial grades and user responses. These results show that CAELF is more suitable for handling human interaction in interactive feedback.

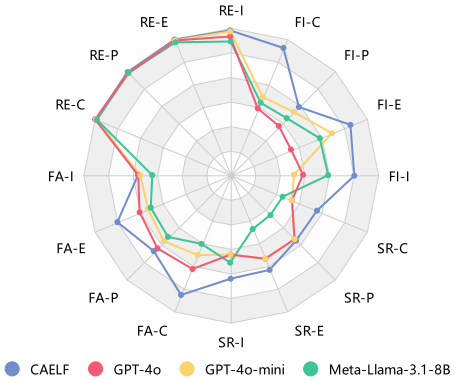


Fig. 2. Human evaluation results, including four human evaluation metrics on each feedback dimensions. For example, Readability-Issue (RE-I) represents the readability of feedback in issue dimension.

Human Evaluation Result. To evaluate the feedback quality, we conducted a manual analysis of the textual content generated by the experiments. We adopted feedback evaluation criteria proposed in [15] and included the following four dimensions in our evaluation: **Readability (RE)**, **Factuality (FA)**, **Self-Regulation (SR)** and **Future Improvement (FI)**. This process resulted in a total of 3,200 ratings across all criteria (800 for each models). Figure 2 shows the results of the human evaluation. CAELF achieved average scores of 4.943, 4.331, 3.344, and 4.363 for the four metrics of Readability (RE), Factuality

(FA), Self-Regulation (SR), and Future Improvement (FI), respectively. With CAELF, we observed all methods scored highly for Readability, and significant improvements in another three feedback dimensions compared to the baseline models. These results highlight CAELF's ability to provide more accurate and actionable feedback, especially in helping students self-reflect and improve their future performance.

5 Conclusion

In this paper, we propose a Contestable AI-Empowered LLM Framework for Interactive Feedback Generation (CAELF), aimed at automating the interactive feedback process and systematically addressing the weaknesses of LLMs in current interactive educational environments. We conducted a case study of critical thinking essay assessment using a dataset of 500 essays and a four-dimensional assessment rubric, including automated experiments and additional human evaluation. The results show that CAELF performs much better than the baseline model in interactive feedback task. This work demonstrates the significant potential of CAELF for applications in interactive learning environments, providing hope for overcoming the time and resource constraints that have historically hindered the widespread adoption of interactive feedback.

Acknowledgements. This research is supported by the Ministry of Education, Singapore (Grant IDs: RG22/23, RG17/22, LKCMedicine Start up Grant).

References

1. Alfrink, K., Keller, I., Kortuem, G., Doorn, N.: Contestable AI by design: towards a framework. Mind. Mach. **33**(4), 613–639 (2023)
2. Association of American Colleges and Universities: Value rubric development project (2019). https://www.aacu.org/value/rubrics
3. Baroni, P., Gabbay, D., Giacomin, M., van der Torre, L. (eds.): Handbook of Formal Argumentation. College Publications, London, England (2018)
4. Boud, D., Molloy, E.: Rethinking models of feedback for learning: the challenge of design. Assess. Eval. High. Educ. **38**(6), 698–712 (2013)
5. Dai, W., et al.: Can large language models provide feedback to students? A case study on ChatGPT. In: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 323–325. IEEE (2023)
6. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate (2023). https://arxiv.org/abs/2305.14325
7. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif. Intell. **77**(2), 321–357 (1995) https://doi.org/10.1016/0004-3702(94)00041-X, https://www.sciencedirect.com/science/article/pii/000437029400041X
8. Hagging Face: Essay grade v1 project (2024). https://huggingface.co/datasets/deepachalapathi/essay_grade_v1

9. Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**(1), 81–112 (2007). https://doi.org/10.3102/003465430298487
10. Hirsch, T., Merced, K., Narayanan, S., Imel, Z.E., Atkins, D.C.: Designing contestability: interaction design, machine learning, and mental health. In: Proceedings of the 2017 Conference on Designing Interactive Systems, pp. 95–99. DIS 2017, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3064663.3064703
11. Hopfenbeck, T.N., Zhang, Z., Sun, S.Z., Robertson, P., McGrane, J.A.: Challenges and opportunities for classroom-based formative assessment and AI: a perspective article. In: Frontiers in Education, vol. 8, p. 1270700. Frontiers Media SA (2023)
12. Kostic, M., Witschel, H.F., Hinkelmann, K., Spahic-Bogdanovic, M.: LLMs in automated essay evaluation: a case study. In: Proceedings of the AAAI Symposium Series, vol. 3, pp. 143–147 (2024)
13. Leofante, F., et al.: Contestable AI needs computational argumentation (2024). https://arxiv.org/abs/2405.10729
14. Liang, T., et al.: Encouraging divergent thinking in large language models through multi-agent debate. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 17889–17904. Association for Computational Linguistics, Miami, Florida, USA (2024)
15. Mitra, C., Miroyan, M., Jain, R., Kumud, V., Ranade, G., Norouzi, N.: RetLLM-E: retrieval-prompt strategy for question-answering on student discussion forums. Proc. AAAI Conf. Artif. Intell. **38**(21), 23215–23223 (2024). https://doi.org/10.1609/aaai.v38i21.30368, https://ojs.aaai.org/index.php/AAAI/article/view/30368
16. Moore, S., Nguyen, H.A., Chen, T., Stamper, J.: Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods. In: Viberg, O., Jivet, I., Muñoz-Merino, P., Perifanou, M., Papathoma, T. (eds.) Responsive and Sustainable Educational Futures. EC-TEL 2023. LNCS, vol. 14200, pp. 229–245. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-42682-7_16
17. Nicol, D.: From monologue to dialogue: improving written feedback processes in mass higher education. In: Approaches to Assessment that Enhance Learning in Higher Education, pp. 11–27. Routledge (2014)
18. Sadler, D.R.: Formative assessment and the design of instructional systems. Instr. Sci. **18**(2), 119–144 (1989)
19. Schmidt, S.J.: Using writing to develop critical thinking skills. NACTA J. 31–38 (1999)
20. Sharadgah, T.: Developing critical thinking skills through writing in an internet-based environment. In: Society for Information Technology & Teacher Education International Conference, pp. 2178–2185. Association for the Advancement of Computing in Education (AACE) (2014)
21. Shute, V.J.: Focus on formative feedback. Rev. Educ. Res. **78**(1), 153–189 (2008)
22. Tang, X., et al.: Medagents: large language models as collaborators for zero-shot medical reasoning (2024). https://arxiv.org/abs/2311.10537
23. Wang, B., Yue, X., Sun, H.: Can ChatGPT defend its belief in truth? Evaluating LLM reasoning via debate (2023). https://arxiv.org/abs/2305.13160
24. Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., Trausan-Matu, S.: Automated essay scoring in applied games: reducing the teacher bandwidth problem in online training. Comput. Educ. **123**, 212–224 (2018)

25. Xiu, Y., Xiao, Z., Liu, Y.: LogicNMR: probing the non-monotonic reasoning ability of pre-trained language models. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 3616–3626. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, December 2022. https://doi.org/10.18653/v1/2022.findings-emnlp.265
26. Yancey, K.P., Laflair, G., Verardi, A., Burstein, J.: Rating short l2 essays on the CEFR scale with GPT-4. In: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pp. 576–584 (2023)



WILLM: A System for Academic Writing Improvement Based on Large Language Models

Yongli Mou(✉), Fatemeh Fathi, Ben Thillen, and Stefan Decker

Chair of Computer Science 5, RWTH Aachen University, Aachen, Germany
{mou,fathi,decker}@dbis.rwth-aachen.de, ben.thillen@rwth-aachen.de

Abstract. Academic writing poses significant challenges for students, particularly non-native English speakers. While existing tools, such as Grammarly, provide surface-level corrections, they often lack detailed explanations, long-term skill development and personalized support. In this paper, we introduce WILLM, a system for Academic Writing Improvement based on Large Language Models (LLMs). WILLM provides context-aware feedback on grammar, vocabulary, coherence, and organization while integrating active recall quizzes and personalized reviews to enhance long-term writing proficiency. A three-week user study with 19 non-native English-speaking participants demonstrated improvements in grammar and vocabulary scores and high usability ratings (SUS score: 84.21). Results suggest that further refinement of long-term improvement features is essential for enhancing academic writing development.

Keywords: Academic English writing · Large language models · Prompt engineering

1 Introduction

Academic writing is essential for scholarly communication [1], but it presents challenges, particularly for non-native English speakers [2], including maintaining clarity, adhering to stylistic conventions, ensuring proper grammar, and avoiding plagiarism [3]. Recently, Large Language Models (LLMs) such as GPT-4 demonstrate exceptional capabilities in understanding, generating, and evaluating text [4], which makes them promising tools for academic writing enhancement [5]. However, existing tools leveraging LLMs often focus on surface-level corrections like grammar, spelling, and punctuation, or generic feedback, leaving gaps in addressing deeper aspects of academic writing, such as coherence, argumentation, and critical analysis [6].

In this paper, we propose WILLM, a system for academic **W**riting **I**mprovement based on **L**arge **L**anguage **M**odels, as a novel solution designed to address these gaps by combining LLM-powered feedback with personalized learning, gamified elements, and user-centric design. WILLM aims to improve

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 36–43, 2025.
https://doi.org/10.1007/978-3-031-98462-4_5

not only the immediate quality of writing but also the long-term development of academic writing skills. Our main contributions in this study are the following: **(1)** Design and implementation of WILLM, a novel academic writing improvement system that leverages LLMs to provide real-time, context-aware feedback on grammar, vocabulary, coherence, and organization, alongside automated scoring based on a custom rubric; **(2)** Integration of long-term skill development features, including active recall quizzes and personalized reviews, to foster sustained writing proficiency; **(3)** Empirical evaluation of WILLM through a three-week user study with non-native English speakers, demonstrating practical improvements in writing quality and high usability (SUS score: 84.21).

2 Related Work

The AI-driven academic writing tools has been greatly influenced by advances in natural language processing (NLP). Recent studies highlight the complementary roles of formative and summative assessments, with formative methods particularly effective due to their timely and targeted feedback mechanisms [7,8]. Writing assessments, which were previously relying on human evaluators, are increasingly adopting automated scoring methods, becoming increasingly aligned with human judgments [9–11]. Modern LLMs, particularly GPT variations, enhance it by accurately predicting writing quality across diverse contexts, facilitating real-time grammatical, stylistic, and coherence adjustments [12–14]. Beyond assessment, writing improvement tools, especially AI-powered platforms such as Grammarly, QuillBot, and DeepL, enhance writing quality by providing real-time and advanced functionalities like paraphrasing, stylistic revisions, and content generation [15–17]. Empirical studies consistently demonstrate these tools' effectiveness in reducing errors and improving overall writing proficiency [18,19]. Beyond immediate corrections, research emphasizes the potential of AI-supported strategies for long-term skill reinforcement, including spaced repetition, adaptive quizzes, and reflective practices [20–22]. Additionally, gamification techniques, evidenced by tools like Duolingo, have shown promise in motivating sustained engagement through badges, points, and competitive elements, although their effectiveness varies by learner and context [23–25].

3 Methodology

3.1 System Design

WILLM addresses the multifaceted challenges of academic writing by integrating LLMs with advanced educational strategies. The system employs a client-server architecture designed for scalability and efficiency. As illustrated in Fig. 1, WILLM comprises a user interface frontend, API implementation backend, user data storage, and LLM components.

Developed as a web-based application, WILLM provides real-time feedback, facilitates long-term learning, and incorporates gamification elements to sustain

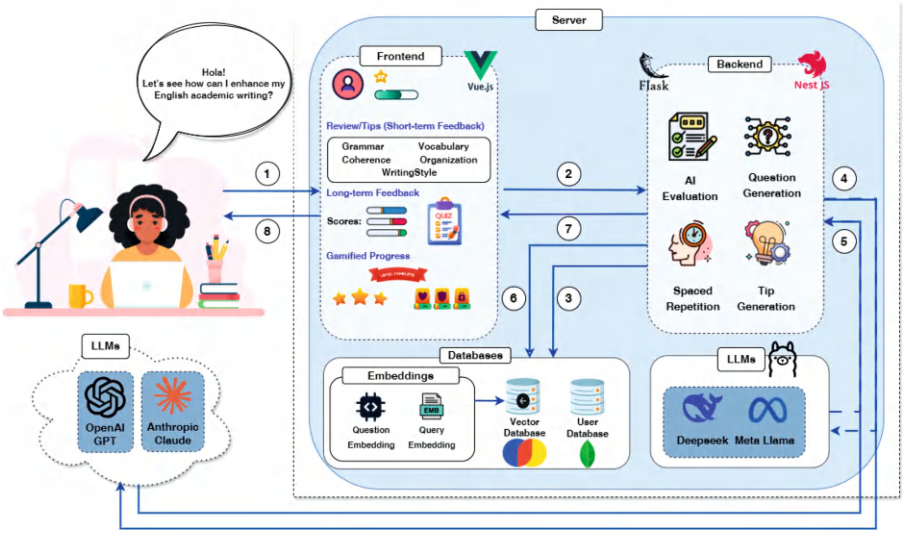


Fig. 1. The architectural overview of the WILLM

user engagement. After submitting, WILLM provides actionable feedback on five key elements of academic writing: grammar, vocabulary, organization, coherence, and writing style. The system evaluates submission quality using custom scoring rubrics, helping users understand their performance and identify areas for improvement. This combination of detailed corrections and structured scoring provides immediate assistance while establishing a foundation for long-term skill development. To cultivate sustained academic writing proficiency, WILLM features specialized enhancement mechanisms including quizzes and personalized reviews based on advanced educational principles such as active recall, spaced repetition, and reflective learning [26]. Recognizing motivation's role in learning, WILLM incorporates gamification elements including Experience Points (XP), progress tracking, and achievement badges. Users earn XP by completing tasks such as submitting drafts or taking quizzes, enabling level progression. Badges and achievements are unlocked based on specific milestones like consistent usage or significant writing score improvements. These features appear prominently in user profiles, creating a sense of accomplishment and encouraging continued platform engagement.

3.2 Implementation Details

WILLM employs LLMs for automated feedback, quiz generation, and spaced repetition to enhance immediate and long-term writing skills.

Feedback Generation. WILLM offers real-time feedback on grammar, vocabulary, coherence, organization, and style. Users submit texts, analyzed by an

LLM via API, identifying issues and assigning scores using a structured rubric aligned with TOEFL, IELTS, and PTE standards (see Fig. 2). The rubric evaluates: **Grammar**: Sentence structure and correctness, **Vocabulary**: Word choice and lexical diversity, **Organization**: Logical flow and paragraph structure, **Coherence**: Transitions and contextual consistency and **Writing Style**: Formal tone and clarity. The system and user prompts for generating scores based on the customized rubrics are depicted in Fig. 2.

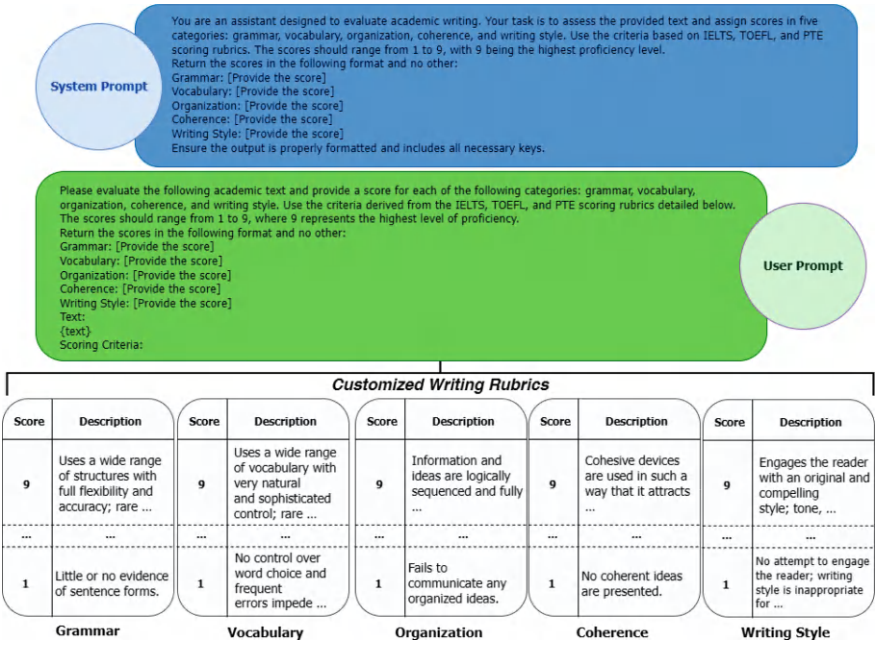


Fig. 2. Customized Writing Rubric-based System and User Prompt for Scoring

Quiz Generation. WILLM generates quizzes based on user-specific mistakes from previous texts, emphasizing grammar, vocabulary, organization, coherence, and style. Questions, stored as vector representations, employ spaced repetition (initial quiz within 24 h, followed by intervals of two, four, and five days) to reinforce retention. Quizzes contain multiple-choice and text-input questions, providing immediate feedback with explanations.

Review Generation. WILLM compiles and analyzes users' recent submissions, identifying progress and areas for improvement, especially in organization and coherence, supplemented by personalized writing tips to enhance long-term proficiency.

3.3 User Study

To evaluate the effectiveness and usability of WILLM, we conducted a user study involving non-native English-speaking participants.

Participant Demographics. The study involved 19 non-native English speakers: 5 undergraduates, 9 bachelor's degree holders, and 5 master's students. Participants were selected based on academic background (minimum bachelor's level or current enrollment), non-native English proficiency, and familiarity with writing tools or LLMs. All participants provided informed consent and were briefed on data privacy and ethical guidelines.

Experiment Setup. A within-subjects design allowed participants to serve as their own controls, directly comparing pre- and post-test performance, minimizing the influence of external variables. Additionally, the study also assessed the system usability through the System Usability Scale (SUS) [27] to measure user satisfaction.

4 Results

In this section, we present the results of the user study regarding the effectiveness of WILLM in improving academic writing skills, including user engagement, pre- and post-test performance comparisons, and usability evaluations.

4.1 User Engagement

User engagement metrics were assessed through analysis of submissions, quiz participation, and review activities over a three-week period. Participants averaged 14.69 weekly submissions, with considerable variability observed (range: 3.50–39.12). Overall, each user submitted an average of 23.32 pieces (range: 7–95, SD = 20.26). Quiz engagement showed a completion rate of 42.55% (20 out of 47 quizzes, totaling 56 questions). The overall accuracy was 62.50%, varying notably by question type; the highest accuracy was observed in academic sentence (81.82%) and synonym (80.00%) questions, whereas revision questions yielded the lowest accuracy rate (20.00%). A total of 78 reviews were generated, averaging 4.88 reviews per user and 1.77 per session. Weekly engagement increased significantly during the second week (Week 1: 7 reviews; Week 2: 44 reviews; Week 3: 25 reviews).

4.2 Pre and Post-test Performance

User performance was evaluated using automated essay scoring (AES) via GPT-3.5, aligned with previous research demonstrating LLMs' reliability in AES applications [17,28,29]. Each writing sample was scored across five dimensions: grammar, vocabulary, organization, coherence, and writing style, with the average of five repeated evaluations per submission recorded.

Comparative analysis revealed performance improvements from pre- to post-tests in all dimensions. Grammar scores improved from 6.22 (SD = 0.94) to 6.76 (SD = 0.65), coherence from 6.68 (SD = 0.84) to 7.10 (SD = 0.70), organization from 7.11 (SD = 0.73) to 7.47 (SD = 0.66), vocabulary from 6.35 (SD = 0.71) to 6.72 (SD = 0.70), and writing style from 6.07 (SD = 0.76) to 6.45 (SD = 0.60). Additionally, a reduction in standard deviations indicates increased consistency in user performance.

4.3 Usability and User Feedback

System usability and user satisfaction were assessed using the SUS, a validated instrument comprising ten items evaluated on a 5-point Likert scale. Responses were scored and adjusted according to SUS standard protocols, yielding a final usability score scaled to 100. Scores above 68 indicate acceptable usability, aligning with industry benchmarks.

4.4 Discussion and Limitations

Although WILLM improved writing proficiency, several factors impact its effectiveness. Model selection balances GPT-3.5's affordability and GPT-4o's superior quality; however, GPT-4o's cost limits usage frequency. Grammar and vocabulary corrections utilize either model, whereas quizzes and reviews exclusively employ GPT-4o. Zero-shot prompting effectively minimized token usage but occasionally produced inconsistencies in coherence feedback. LLM hallucinations necessitated post-processing filters. The three-week, 19-participant study showed modest, non-significant improvements, indicating that larger samples and longer durations are required for conclusive results. Low engagement limited the impact of spaced repetition and personalized reviews. Improving interface usability and review material presentation could enhance effectiveness. Additionally, WILLM's single-document limitation and reliance on API-based LLMs constrain scalability and adaptability.

5 Conclusion

This paper introduces WILLM, an academic writing improvement system based on LLMs, providing comprehensive feedback and integrated quizzes and gamification for sustained skill enhancement. The initial evaluation showed promising grammar and vocabulary improvements and high user acceptance. Future work should involve larger, diverse participant samples over extended periods, enhanced user interface designs, and reduced reliance on costly API-based LLMs by exploring robust open-source alternatives such as DeepSeek-V3. Addressing these areas will enhance WILLM's practicality and effectiveness in academic writing support.

Acknowledgments. This work was partially supported by the German Federal Ministry of Education and Research (BMBF) under projects WestAI (Grant no. 01IS22094D) and MyEduLife (Grant no. 21INVI02).

References

1. Tang, X., Duan, X., Cai, Z.G.: Are LLMs good literature review writers? evaluating the literature review writing ability of large language models. *arXiv preprint*arXiv:2412.13612 (2024)
2. Yuvayapan, F., Bilginer, H.: Identifying the needs of postgraduate students: the first step of academic writing courses. J. Lang. Linguist. Stud. **16**(2), 595–611 (2020)
3. Wallwork, A.: English for Academic Research. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-53174-3
4. Achiam, J., et al. Gpt-4 Technical report. arXiv preprint arXiv:2303.08774 (2023)
5. Li, C., et al.: Teach LLMs to personalize–an approach inspired by writing education. arXiv preprint arXiv:2308.07968 (2023)
6. Ishak, F.: Investigating the relationship between Algerian EFL learners' use of Grammarly and academic grammar writing: The case of master students at Biskra University (2023)
7. Chen, D., Zhang, L.: Formative assessment of academic English writing for Chinese EFL learners. TESOL Int. J. **12**(2), 47–64 (2017)
8. Horstmanshof, L., Brownie, S.: A scaffolded approach to discussion board use for formative assessment of academic writing skills. Assess. Eval. High. Educ. **38**(1), 61–73 (2013)
9. Hunter, D.M., Jones, R.M., Randhawa, B.S.: The use of holistic versus analytic scoring for large-scale assessment of writing. Can. J. Program Eval. **11**(2), 61–86 (1996)
10. Ghalib, T.K., Al-Hattami, A.A.: Holistic versus analytic evaluation of EFL writing: a case study. Engl. Lang. Teach. **8**(7), 225–236 (2015)
11. Xie, J., Cai, K., Kong, L., Zhou, J., Qu, W.: Automated essay scoring via pairwise contrastive regression. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 2724–2733 (2022)
12. Dong, F., Zhang, Y., Yang, J.: Attention-based recurrent convolutional neural network for automatic essay scoring. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 153–162 (2017)
13. Yongchao, W., Henriksson, A., Nouri, J., Duneld, M., Li, X.: Beyond benchmarks: spotting key topical sentences while improving automated essay scoring performance with topic-aware BERT. Electronics **12**(1), 150 (2022)
14. Mansour, W., Albatarni, S., Eltanbouly, S., Elsayed, T.: Can large language models automatically score proficiency of written essays? *arXiv preprint*arXiv:2403.06149 (2024)
15. Gain, A., Rao, M., Bhat, S.K.: Usage of Grammarly–online grammar and spelling checker tool at the health sciences library, Manipal academy of higher education, Manipal: a study. Library Philosophy and Practice, 1–13 (2019)
16. Kurniati, E.Y., Fithriani, R.: Post-graduate students' perceptions of Quillbot utilization in English academic writing class. J. English Lang. Teach. Linguist. **7**(3), 437–451 (2022)

17. Mizumoto, A., Eguchi, M.: Exploring the potential of using an AI language model for automated essay scoring. Res. Methods Appl. Linguist. **2**(2), 100050 (2023)
18. Tran, T.: Ai tools in teaching and learning English academic writing skills. Proc. AsiaCALL Int. Conf. **4**, 170–187 (2023)
19. Dja'far, V.H., Hamidah, F.N.: The effectiveness of AI technology in improving academic English writing skills in higher education. J. Lang. Lit. Stud. **4**(3), 579–593 (2024)
20. Munday, P.: *Duolingo*. gamified learning through translation. J. Span. Lang. Teach. **4**(2), 194–198 (2017)
21. Ruan, S., et al.: Quizbot: a dialogue-based adaptive learning system for factual knowledge. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2019)
22. Nanayakkara, H.: Reflection to the success of language learning. Master's thesis, H. Nanayakkara (2023)
23. Dehganzadeh, H., Dehganzadeh, H.: Investigating effects of digital gamification-based language learning: a systematic review. J. English Lang. Teach. Learn. **12**(25), 53–93 (2020)
24. Loewen, S., et al.: Mobile-assisted language learning: a Duolingo case study. ReCALL **31**(3), 293–311 (2019)
25. Faiella, F., Ricciardi, M.: Gamification and learning: a review of issues and research. J. e-learn. Knowl. Soc. **11**(3), 1–9 (2015)
26. Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., Willingham, D.T.: Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychol. Sci. Publ. Interest. **14**(1), 4–58 (2013)
27. Brooke, J.: SUS: a retrospective. J. Usabil. Stud. **8**(2), 1–12 (2013)
28. Swiecki, Z., et al.: Assessment in the age of artificial intelligence. Comput. Educ. Artif. Intell. **3**, 100075 (2022)
29. Tate, T.P., et al.: Can AI provide useful holistic essay scoring? Comput. Educ. Artif. Intell. **7**, 100255 (2024)



What Makes Teamwork Work? A Multimodal Case Study on Emotions and Diagnostic Expertise in an Intelligent Tutoring System

Xiaoshan Huang(✉), Haolun Wu, Xue Liu, and Susanne P. Lajoie

McGill University, Montreal, QC, Canada
{xiaoshan.huang,haolun.wu}@mail.mcgill.ca,
{xue.liu,susanne.lajoie}@mcgill.ca

Abstract. Teamwork is pivotal in medical settings when professionals with diverse skills and emotional states collaborate to make critical decisions. This case study examines the interplay between emotions and professional skills in group decision-making during collaborative medical diagnosis within an Intelligent Tutoring System (ITS). By comparing verbal and physiological data between high-performing and low-performing teams of medical professionals working on a patient case within the ITS, alongside individuals' retrospective collaboration experiences, we employ multimodal data analysis to identify patterns in team emotional climate and their impact on diagnostic efficiency. Specifically, we investigate how emotion-driven dialogue and professional expertise influence both the information-seeking process and the final diagnostic decisions. Grounded in the socially shared regulation of learning framework and utilizing sentiment analysis, we found that social-motivational interactions are key drivers of a positive team emotional climate. Furthermore, through content analysis of dialogue and physiological signals to pinpoint emotional fluctuations, we identify episodes where knowledge exchange and skill acquisition are most likely to occur. Our findings offer valuable insights into optimizing group collaboration in medical contexts by harmonizing emotional dynamics with adaptive strategies for effective decision-making, ultimately enhancing diagnostic accuracy and teamwork effectiveness.

1 Introduction

Effective teamwork allows groups to achieve more than what individuals working in isolation can accomplish. However, working together does not guarantee optimal outcomes [29]. Collaborative dynamics are essential in team success, especially when emotions are contagious and shared through social interactions [1].

Educational technologies play a key role in supporting learners knowledge acquisition [19], where socio-emotional interactions are crucial to their knowledge co-construction [10,13]. Challenges remain in understanding teamwork within technology-rich environments, where interactions involve not only human-to-human dynamics but also human-computer interactions. These interactions

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 44–52, 2025.
https://doi.org/10.1007/978-3-031-98462-4_6

introduce complexity, and multimodal data analysis [11], which captures effects through various channels such as speech (e.g., emotional tones) [5] and physiological responses (e.g., heart rate variabilities) [2,6], offers avenues to understand these nuanced dynamics [33].

2 Background

2.1 The Role of Affective States in Teamwork

Teamwork operates on both cognitive and affective dimensions, with emotions playing a pivotal, yet less-understood role in shaping team dynamics and overall performance. Positive affective states, such as optimism or curiosity, have been shown to enhance team engagement, foster collaboration, and improve problem-solving efficacy [9]. On the other hand, negative emotions can disrupt communication, hinder decision-making, and ultimately reduce team effectiveness [28]. Understanding the intricate relationship between emotions and teamwork is critical, particularly in complex tasks within technology-rich environments.

Team emotional climate, the shared affective states within a team have impacts on collaborative experiences [20,30] and team effectiveness [7]. Emotional tone, expressed through vocal intonation or written communication [31], reflects individuals' emotional states, which in turn influences cognitive processes [3]. It can significantly affect team performance, highlighting the importance of considering affective factors alongside cognitive ones [15].

2.2 Interactions and Team Emotional Climate in Teamwork

Grounded from the Socially Shared Regulation of Learning framework (SSRL [25]) we argue that team effectiveness hinges on shared regulation of emotions, motivation, cognition and metacognition [17]. SSRL becomes crucial when meaningful interactions are elicited among team members [18,24].

These interactions are closely associated with active participation and overall success in teamwork [16]. Metacognitive interaction facilitates justification of behavior to one and others [8], enhancing team cohesion and performance. Socio-cognitive interactions, such as monitoring and reflection, are also strongly linked to team success. Emotional and motivational interactions, while closely related to socio-cognitive interactions, are more challenging to distinguish and measure [23]. The complexity of human-computer interactions adds another layer to these interactions, affecting learner engagement dynamically [4,12]. Emotional contagion, which shapes team emotional climate, plays a vital role in influencing team efficiency. Analyzing emotional sharing and regulation thus offers a key lens for optimizing teamwork [26].

In this case study, we investigated two research questions (RQs) are raised within an Intelligent Tutoring System (ITS) context: ***1.*** *How do interaction types relate to a team's emotional climate?* ***2.*** *What factors trigger emotional changes during team decision-making?*

3 Methodology

3.1 Participants and Learning Environment

This case study consisted of 4 dyads of female medical professionals ($\text{Mean}_{age} = 28.13$) enrolled in a residency program at a large North American university. The study was conducted after receiving approval from the ethics board of the institute where participants were involved. BioWorld [12,21], an intelligent tutoring system designed to assist medical students in learning through problem-solving tasks, was used as the media and learning platform. Each dyad was co-located to share the same screen. The BioWorld interface includes four primary functions: *Patient Case*, *Tests*, *Library*, and *Hypotheses*, accessible via a left-side navigation bar where learners can freely navigate between these functions.

3.2 Measurements

Verbal data were collected through transcribed dialogues from each team. Team emotional climate was quantified using compound sentiment scores derived from sentiment analysis, while various interaction types were identified within the dialogue transcriptions.

Physiological data were obtained during the diagnostic task. Participants were fitted with the Empatica E4 bracelet, a non-intrusive device, on their non-dominant hands. Average heart rate was derived from the raw blood volume pulse (BVP) signal, sampled at a default rate of 1 Hz. Heart rate change points were analyzed to indicate moments of heightened emotional intensity or stress, reflecting shifts in the team's emotional climate [27]. We utilized the Bayesian Estimator of Abrupt change, Seasonal change, and Trend (BEAST) algorithm to probabilistically identify periods of substantial heart rate variation [32]. This approach helps to pinpoint moments of physiological arousal changes linked to teamwork experience. Analysis was conducted using the RBeast[1].

Team effectiveness was measured using objective performance indices including diagnostic efficiency and subjective reflections gathered through semi-structured interviews.

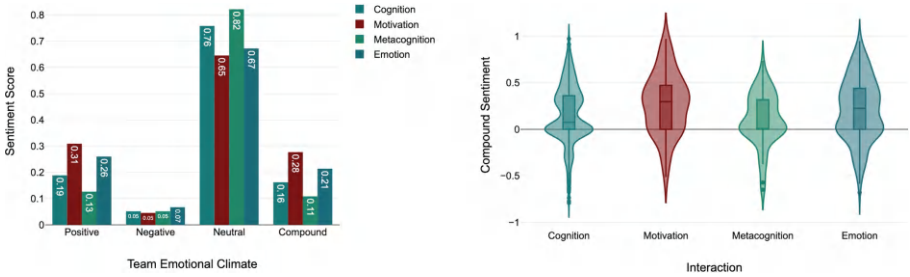
3.3 Analytic Method

We used sentiment analysis to identify team emotional climate [14]. The compound score serves as an aggregated measure, encapsulating the sentiment polarity of the text, ranging from −1 (extreme negativity) to +1 (extreme positivity).

We adapted a SSRL scheme[2] to code the dialogue. Specifically, we categorized interactions into four types: *Metacognitive*, *Cognitive*, *Emotional*, and *Motivational*. Two researchers coded the dialogues and solved discrepancies

[1] https://pypi.org/project/Rbeast/.

[2] https://github.com/katherine-huang/AIED-2025/blob/main/coding_scheme_SSRL.png.



(a) Interaction by Team Climate (b) Compound Sentiment by Interaction

Fig. 1. Interaction by team climate and compound sentiment by interaction.

among codes. The interrater reliability for each interaction category exceeded 0.7, met the accepted threshold for Cohens kappa [22]. Content analysis using multimodal data was conducted by linking heart rate change point timestamps to team dialogues, connecting them to specific interactions.

4 Results

4.1 RQ1. Interaction and Team Emotional Climate

The dialogue in teams shows an overall neutral to slightly positive emotional climate during the task (Fig. 1), with a statistically significant effect of interaction type on emotional tone, $F(3, 1116) = 4.96$, $p = .002$. The compound sentiment of socio-motivational interactions (mean = 0.28), is significantly higher than both cognitive (mean = 0.11, $p = .02$) and metacognitive interactions (mean = 0.16, $p = .01$) based on Tukey's Honestly Significant Difference (HSD) tests.

4.2 RQ2. Triggers to Team Emotional Climate Changes

We examined differences in emotional climate shifts during decision-making between high-performing (Group A) and low-performing (Group B) teams. In Group A, heart-rate changes were observed as team members navigated complex medical terminology and decision-making related to diagnostic assumptions (*P1* and *P2* are short for *participant 1* and *participant 2* respectively):

> (P1): "Okay. So then, what do you think is going on? [**searching in the BioWorld library**] ...something in the adrenal glands? Okay, click." **[heart rate change detected]**
> (P2): "Catecholamine."
> (P1): "And then... let's see. So headache, sweating, and tachycardia."
> (P2): "Yeah, pretty exciting. [**reading on the searching result from the library**]" **[heart rate change detected]**

This sequence shows that when the team reached a pivotal decision point that they summarized in their report, such as identifying the test results with related

symptoms ("elevated catecholamines" and "headache, sweating, and tachycardia"), heart rate changes occurred. This suggests engagement during these critical exchanges are necessary for advancing their problem-solving process. Additionally, the ITS assists in knowledge acquisition, helping the team develop a deeper understanding of the case.

Furthermore, heart-rate changes were detected during the following conversation. In this conversation, P1 was seeking clarification about some lab results, specifically values related to measurements which are not fully understood. P2 provided clarification, stating that the values are not elevated, and emphasizes the importance of normal thyroid-stimulating hormone (TSH) levels. The conversation reveals P1's relief upon understanding that their results are within acceptable ranges and reassures them that there's nothing concerning beyond normal checks. They also mentioned a key test to order for this given patient profile – TSH, which is an important cue for them to correctly diagnose.

(P1) "And then it's... 1560 for the free (urine test). Ah, it said the total was elevated, I don't know. and then 1600 is good?"
(P2) "These are micrograms. We're not in micrograms. They just aren't elevated." **[heart rate change detected]** ...
(P1) "And then, what helped us?"
(P2) "The catecholamine [laughs]"
(P1) "I know."
(P2) "I mean, for sure, the fact that everything else is good. Like, for sure, the fact that your TSH is normal is helpful. And then, like the cortisol. This one is also helpful. I think the rest is kinda just, like, make sure they're fine." **[heart rate change detected]**
(P1) "Well, it's good."

In contrast, Group B experienced emotional fluctuations during critical knowledge exchange moments but failed to turn them into productive learning events:

(P2) "Now that everything is thrown back to normal, I think your anxiety one (i.e., the highlighted patient symptom for diagnosis) is higher on this."
(P1) "It's good. Okay. Panic attacks (an incorrect diagnosis was raised)."

Further discussion about whether other symptoms were relevant to their final hypothesis during which another emotional fluctional was detected:

(P2) "Like dizzy and faint, not necessarily."
(P1) "I agree. Palpitations, yes."
(P2) "Yeah. You're trying to bring it back up?"
(P1) "Sweating? I am not sure."

The dialogue reveals lingering uncertainties and unresolved issues, which led to a misdiagnosis. Their failure to align their emotional shifts with focused problem-solving discussions. Intervention of hints or regulation are therefore needed to prevent negative emotions and foster clarity in decision-making.

Most of heart rate changes in both groups occurred during critical knowledge exchanges (sharing, refining, or questioning). However, Group A's emotional fluctuations aligned with confident knowledge acquisition, while Group B's emotional fluctuations coincided with uncertain and unsolved inquiries.

5 Discussion and Conclusion

5.1 Motivational Expressions Drive Positive Team Climate

Our findings reinforce the critical role of expressions of motivation in teamwork. During team decision-making, the "mission possible" mentality—where team members believe in the attainability of their goals—helps sustain motivation and positively influences the team's emotional well-being. By fostering these socio-motivational interactions, ITSs could be designed to better support positive team experiences, thus leading to enhanced teamwork.

5.2 Emotional Fluctuations as Catalysts in Team Decision-Making

The results indicate that shifts in emotional climate occur during critical moments of decision-making, where the succeed team took benefits from knowledge sharing and professional skill transfer. These moments are essential for learning, as they highlight the interaction-driven nature of teamwork, where team members actively contribute to each other's learning process. Emotional fluctuations also indicate cognitive engagement and deeper processing of new information. However, only by appropriate regulation and critical reflection within the team, the team potential signaled by the team's emotional shift could be maximized. Without deliberate regulation, emotional shifts risk tipping from productive curiosity into emotions of disengagement (e.g., boredom) or overload (e.g., anxiety).

For teams that struggled, additional prompts could provide supports to teamwork effectiveness. Providing real-time, adaptive feedback that prompts reflection during moments of emotional highs or lows can guide teams toward more structured, collaborative decision-making. For instance, customized scaffolds that encourage re-evaluation of evidence when emotional tension or uncertainty arises could help mitigate the negative impact of unresolved discrepancies.

These interventions can help struggling teams to manage emotional variability constructively, turning moments of doubt into opportunities for deeper cognitive engagement and collective problem-solving.

5.3 Implications for ITS Design and Teamwork Research

These findings underscore the importance of integrating artificial emotional intelligence in ITS. Specifically, future ITS could integrate features that detect and

respond to social interaction cues, encouraging interactions that boost team emotional climate. Additionally, ITSs designed to recognize and support periods of emotional fluctuation as part of the learning process, providing scaffolds that guide teams through challenging moments while capitalizing on opportunities for deeper engagement during knowledge exchange. For example, future design can combine emotion detection with behavioral logs to enable affect-aware tutoring, and to provide hints or meta-cognitive prompts such as:

- *"Take a moment to reflect on and discuss about the differing opinions."*
- *"Do you need to revisit the key information to resolve this uncertainty?"*

Furthermore, analyzing how team effectiveness evolves with fluctuations in cognitive and affective states offers valuable insights into team dynamics. Applying this analytical lens to high-stakes teamwork environments offers unique opportunities to: (1) uncover challenge-dependent teamwork mechanisms, (2) identify inflection points where performance pathways diverge, and (3) develop contingency-based strategies for sustaining effectiveness under stress. This approach moves beyond static team assessments toward a process-oriented understanding of adaptive collaboration within technology-rich environments.

References

1. Andersen, P.A., Guerrero, L.K.: Principles of communication and emotion in social interaction. In: Handbook of Communication and Emotion, pp. 49–96. Elsevier (1996)
2. Appelhans, B.M., Luecken, L.J.: Heart rate variability as an index of regulated emotional responding. Rev. Gen. Psychol. **10**(3), 229–240 (2006)
3. Chang, H.S., Lee, C.Y., Wang, X., Young, S.T., Li, C.H., Chu, W.C.: Emotional tones of voice affect the acoustics and perception of mandarin tones. PLoS ONE **18**(4), e0283635 (2023)
4. Chen, F., Li, S., Lin, L., Huang, X.: Identifying temporal changes in student engagement in social annotation during online collaborative reading. Educ. Inf. Technol. 1–24 (2024)
5. Dehbozorgi, N., Mohandoss, D.P.: Aspect-based emotion analysis on speech for predicting performance in collaborative learning. In: 2021 IEEE Frontiers in Education Conference (FIE), pp. 1–7. IEEE (2021)
6. Donker, M., van Aken, S., Mainhard, T.: Using heart rate to tap into motivational and emotional processes during teaching and learning. In: Motivation and Emotion in Learning and Teaching across Educational Contexts, pp. 258–273. Routledge (2023)
7. Farh, C.I., Seo, M.G., Tesluk, P.E.: Emotional intelligence, teamwork effectiveness, and job performance: the moderating role of job context. J. Appl. Psychol. **97**(4), 890 (2012)
8. Frith, C.D.: The role of metacognition in human social interactions. Philos. Trans. Roy. Soc. B Biol. Sci. **367**(1599), 2213–2223 (2012)
9. Haas, M., Mortensen, M.: The secrets of great teamwork. Harv. Bus. Rev. **94**(6), 70–76 (2016)
10. Huang, X., Lajoie, S.P.: Social emotional interaction in collaborative learning: why it matters and how can we measure it? Soc. Sci. Humanit. Open **7**(1), 100447 (2023)

11. Huang, X., Lajoie, S.P.: A scoping review on effective measurements of emotional responses in teamwork contexts. Curr. Psychol. 1–22 (2024)
12. Huang, X., Li, S., Lajoie, S.P.: The relative importance of cognitive and behavioral engagement to task performance in self-regulated learning with an intelligent tutoring system. In: International Conference on Intelligent Tutoring Systems, pp. 430–441. Springer (2023)
13. Huang, X., Wu, H., Liu, X., Lajoie, S.: Examining the role of peer acknowledgements on social annotations: unraveling the psychological underpinnings. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–9 (2024)
14. Hutto, C., Gilbert, E.: Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, pp. 216–225 (2014)
15. Imai, Y.: Emotions in SLA: new insights from collaborative learning for an EFL classroom. Mod. Lang. J. **94**(2), 278–292 (2010)
16. Isohätälä, J., Järvenoja, H., Järvelä, S.: Socially shared regulation of learning and participation in social interaction in collaborative learning. Int. J. Educ. Res. **81**, 11–24 (2017)
17. Järvelä, S., et al.: Socially shared regulation of learning in CSCL: understanding and prompting individual-and group-level shared regulatory activities. Int. J. Comput.-Support. Collab. Learn. **11**, 263–280 (2016)
18. Järvelä, S., Nguyen, A., Vuorenmaa, E., Malmberg, J., Järvenoja, H.: Predicting regulatory activities for socially shared regulation to optimize collaborative learning. Comput. Hum. Behav. **144**, 107737 (2023)
19. Jerinic, L., Devedzic, V.: The friendly intelligent tutoring environment. ACM SIGCHI Bull. **32**(1), 83–94 (2000)
20. Kelly, J.R., Barsade, S.G.: Mood and emotions in small groups and work teams. Organ. Behav. Hum. Decis. Process. **86**(1), 99–130 (2001)
21. Lajoie, S.P.: Student modeling for individuals and groups: the bioworld and howard platforms. Int. J. Artif. Intell. Educ. **31**(3), 460–475 (2021)
22. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia Medica **22**(3), 276–282 (2012)
23. Näykki, P., Isohätälä, J., Järvelä, S., Pöysä-Tarhonen, J., Häkkinen, P.: Facilitating socio-cognitive and socio-emotional monitoring in collaborative learning with a regulation macro script-an exploratory study. Int. J. Comput.-Support. Collab. Learn. **12**, 251–279 (2017)
24. Nguyen, A., Järvelä, S., Rosé, C., Järvenoja, H., Malmberg, J.: Examining socially shared regulation and shared physiological arousal events with multimodal learning analytics. Br. J. Edu. Technol. **54**(1), 293–312 (2023)
25. Panadero, E., Järvelä, S.: Socially shared regulation of learning: a review. Eur. Psychol. (2015)
26. Perry, N.E., Winne, P.H.: Tracing students' regulation of learning in complex collaborative tasks. In: Interpersonal Regulation of Learning and Motivation, pp. 45–66. Routledge (2013)
27. Pijeira-Díaz, H.J., Drachsler, H., Järvelä, S., Kirschner, P.A.: Sympathetic arousal commonalities and arousal contagion during collaborative learning: how attuned are triad members? Comput. Hum. Behav. **92**, 188–197 (2019)
28. Polo, C., Simonian, S., Chaker, R.: Affording emotional regulation of distant collaborative argumentation-based learning at university. Educ. Technol. Int. **23**(1), 1–39 (2022)

29. Salas, E., Sims, D.E., Burke, C.S.: Is there a “big five” in teamwork? Small Group Res. **36**(5), 555–599 (2005)
30. Somech, A., Drach-Zahavy, A.: Translating team creativity to innovation implementation: the role of team composition and climate for innovation. J. Manag. **39**(3), 684–708 (2013)
31. Syrjämäki, A.H., et al.: Emotionally toned online discussions evoke subjectively experienced emotional responses. J. Media Psychol. Theor. Methods Appl. **35**(1), 55 (2023)
32. Zhao, K., et al.: Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A bayesian ensemble algorithm. Remote Sens. Environ. **232**, 111181 (2019)
33. Zhao, L., et al.: Analysing verbal communication in embodied team learning using multimodal data and ordered network analysis. In: International Conference on Artificial Intelligence in Education, pp. 242–254. Springer (2023)

Beyond Code: Leveraging ChatGPT to Enhance (Not Replace) Traditional Programming Education

Adetunji Adeniran[1(✉)] and Manal Alhathli[2]

[1] University of Dundee, Scotland, UK
aadeniran001@dundee.ac.uk
[2] College of Computer and Information Sciences, Princess Nourah Bint Abdul Rahman University, Riyadh, Saudi Arabia
maalhathli@pnu.edu.sa

Abstract. The rise of OpenAI tools like ChatGPT has sparked debate in programming education, balancing benefits for learning against concerns over reduced critical thinking and reliance on traditional methods. This study explores ChatGPT's impact on student learning in a university programming course through assessment analysis and student feedback. Findings show that, when used as a scaffold rather than a shortcut, ChatGPT can enhance understanding and support effective educational integration.

Keywords: OpenAI · Programming Education · ChatGPT · Student Learning

1 Background

ChatGPT's advanced code-generation capabilities enhance computational thinking, programming self-efficacy, and student motivation [2, 5, 17, 19]. It aids coding assignments by offering suggestions, hints, and code snippets, promoting effective coding practices [14].

However, concerns arise about students' overreliance on AI, potentially undermining critical thinking and problem-solving skill development [6, 7, 14, 15]. While traditional resources encouraged deeper learning, ChatGPT's instant solutions may reduce cognitive effort and bypass essential problem-solving steps [17]. The key question is: *How can programming courses be designed to ensure students develop foundational skills while positioning ChatGPT as a support tool rather than a placeholder for their knowledge*?

This study addresses these concerns and proposes a solution framework by exploring four key objectives: *examining how students use ChatGPT in programming courses*, *identifying learner variables that influence ChatGPT use*, *analysing its impact on learning outcomes*, and *exploring strategies to optimize ChatGPT's role in programming education.*

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 53–59, 2025.
https://doi.org/10.1007/978-3-031-98462-4_7

2 Methodology

2.1 Literature Review

We selected studies based on relevance to generative AI in student learning, narrowing to five key 2024 papers on ChatGPT use: Brender et al., Keuning et al., Kiesler et al., Prather et al., and Xu et al.

2.2 Literature Review

Student vs. ChatGPT Performance. We compared student submissions with ChatGPT-generated solutions for two assignments in an *introductory programming course* (2023/2024). ChatGPT responses were graded using the same rubrics as student work, addressing the research question: *Does the type of assessment influence how ChatGPT usage could obscure inferences about student learning*? This sub-research question contributes to addressing the overarching research question of this study.

Student Perception of ChatGPT. A survey (N = 24) was conducted to explore student perceptions of ChatGPT in programming education, with ethical approval (UOD-ASC-TPG-2023–172) and informed consent obtained. The survey did not collect demographic data and focused on: *self-reported programming knowledge before and after the semester*, *perceived effectiveness of ChatGPT compared to other learning resources*, and *the correlation between prior knowledge and perceptions of ChatGPT's educational value.*

3 Result

3.1 ChatGPT and Learning: Key Issues from Literature

Student Usage Pattern of ChatGPT in Programing Learning Activities. Brender et al. (2024) classify student use of ChatGPT into three categories: *Conceptual Explorers*, who use it to understand concepts and tasks; *Debuggers*, who frequently exchange code and errors between ChatGPT and their editor; and *Practical Developers*, who rely on ChatGPT to solve tasks step by step. This typology is supported by other studies. Keuning et al. (2024) highlight debugging as a common use, consistent with the Debugger category. Kiesler et al. (2024) describe behaviors such as generating unverified solutions, seeking explanations, and requesting feedback. Prather et al. (2024) distinguish between *Uncritical Acceptance* (copying code without evaluation) and *Strategic Engagement* (critically integrating ChatGPT suggestions). Similarly, Xu, Sun, and Li (2024) identify patterns like asking conceptual questions, submitting error messages, reviewing feedback, and copying code. These patterns broadly align with Brender et al.'s framework, which we adopt to conceptualize student use and inform strategies for effective ChatGPT integration in programming education.

Influence of Prior Knowledge on Students' ChatGPT Usage Pattern in Programming Learning Activities. **Influence of Prior Knowledge on Students' ChatGPT Usage Pattern in Programming Learning Activities.** Studies have shown that prior programming knowledge influences students' ChatGPT usage patterns, which in turn affect the benefits or drawbacks they experience from the tool. Brender et al. (2024) found that while both high- and low-knowledge students could use ChatGPT to complete lab activities, students with higher prior-knowledge demonstrated improved conceptual understanding in post-tests. This suggests that ChatGPT can help low-knowledge students' complete tasks but is less effective in supporting conceptual learning.

Brender et al. (2024) distinguish between *performance* (students' ability to complete tasks) and *learning outcomes* (conceptual understanding measured through post-tests). They posit that, while ChatGPT use may boost performance, students with lower prior knowledge show weaker learning outcomes.

Prather et al. (2024) further observed that novice student' using ChatGPT encountered challenges such as misleading guidance, metacognitive difficulties, and false sense of progress, hindering deep conceptual understanding. In contrast, students categorized as "Not Struggling" used ChatGPT to accelerate performance.

Xu, Sun, and Li (2024) found high-performing students used ChatGPT more often and effectively than low-performing (LP) peers, echoing Brender et al. (2024), who showed that stronger prior knowledge improves ChatGPT use. Programming knowledge shapes usage patterns, influencing whether ChatGPT aids or hinders learning. Three main themes emerge in the literature: T1 (analyzing students' varied ChatGPT usage), T2 (influence of prior knowledge on usage), and T3 (how T1 and T2, or T1 alone, affect learning outcomes). Studies addressing these themes include [1, 8, 9, 12], and [16], all employing mixed methods. Most studies explored T1 and T3; some ([1, 12, 16]) also examined T2, while [8] and [9] focused primarily on usage and outcomes without fully considering prior knowledge.

3.2 Explorative Studies and Analysis

Building on the core themes (T1, T2, T3), we argue that performance reflects learning outcomes, with assessment scores serving as observable indicators of learning [10]. Based on Brender et al. (2024), performance here relates to scores on formative assessments (e.g., lab activities), while learning outcomes align with summative assessments. Formative assessments demand that performance closely tracks learning gains, whereas summative assessments capture overall achievement.

Drawing from Koedinger et al.'s (2012) KLI model and Brender et al.'s (2024) description, formative assessments must promote active learning that transfers conceptual knowledge for later testing. This raises concerns about how ChatGPT's direct-answer capabilities affect assessment design, prompting us to explore differences across assessment types and content.

Student vs. ChatGPT Performance on Assignments 1 and 2 is illustrated in Figs. 1 and 2 below.

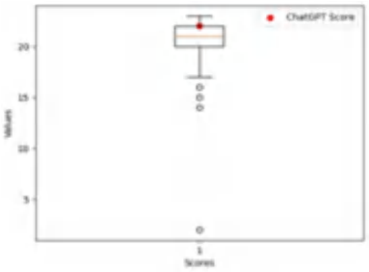


Fig. 1. Assignment 1[1]: ChatGPT scored 22/23, ranking above the 75th percentile among 53 students.

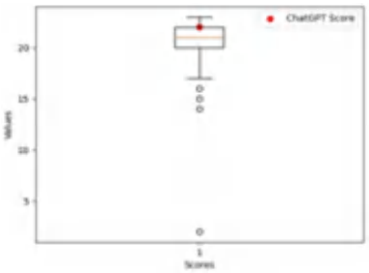


Fig. 2. Assignment 2[2]: ChatGPT's code failed to compile, scoring 11/23, falling into the 25th percentile.

Assignment 1 covered basic programming concepts (control structures, functions), while *Assignment 2* involved advanced topics (OOP, file handling, error tracing). ChatGPT's lower performance on Assignment 2 supports prior findings [4] that it struggles with complex tasks.

These results suggest ChatGPT can assist with basic programming tasks without ensuring deep conceptual understanding, potentially contributing to challenges like the "Junior-Year Wall" [3]. This raises a key question: can we design basic programming assessments that ChatGPT cannot easily solve? As AI improves, this may become a losing battle. Instead, assessments should encourage students to use ChatGPT as Conceptual Explorers rather than Practical Developers [1].

Student Perception of ChatGPT's Impact on Learning Programming. We conducted a survey (ethics ID: UOD-ASC-TPG-2023–172) in an introductory programming course for non-computing students, some with prior experience. The survey assessed students' self-perceived learning progress and their views on various learning methods, including in-class instruction, online resources, and ChatGPT.

Participants self-rated their programming knowledge before (Pre-SA) and after (Post-SA) the course on a 5-point scale (1 = Novice, 5 = Expert). Results show improvement:

[1] View Assignment 1 here: https://tinyurl.com/5cu75wm4.
[2] View Assignment 2 here: https://tinyurl.com/4st6zdf3.

the media rose from 2 to 3, and the mean from 2.2 to 3.04, indicating increased confidence. Skewness was low, suggesting symmetrical responses without major outliers.

The results suggest a self-reported improvement in programming knowledge after completing the course. The median score increased from 2 (Novice) in the Pre-SA to 3 (Intermediate) in the Post-SA, indicating a shift towards greater programming expertise. The mean score also increased from 2.2 to 3.04, further supporting the conclusion that students gained confidence in their programming skills. The skewness values (Pre-SA: 0.05, Post-SA: 0.03) indicate a relatively symmetric distribution of responses, with no significant outliers in self-assessment data.

Perceived Usefulness of Different Learning Mediums. Students rated the usefulness of learning mediums (1 = Least useful, 5 = Most useful). Online text-based resources and videos (mean = 3.8) were rated most useful, followed by in-class instruction (mean = 3.6). ChatGPT received the lowest rating (mean = 2.7, median = 3) with slight positive skew, indicating mixed perceptions.

Perception of Usefulness and Knowledge-Level. We analyzed correlations between Post-SA scores and perceived usefulness of learning resources. Students with higher self-reported skills rated text-based resources as more useful ($r = 0.42$), with weak correlations for in-class instruction ($r = 0.03$) and online videos ($r = 0.09$). ChatGPT showed a negative correlation ($r = -0.39$), indicating that students with less programming knowledge found it more helpful. While this may seem to contradict prior findings [1, 12, 16], the discrepancy likely stems from the question's framing: "Please rate the following medium of learning opportunities on a scale of 1 to 5, with 1 being the lowest and 5 being the highest, based on how much they help you achieve authentic and deep learning of programming." Advanced students often use ChatGPT to validate, debug, and explore, but may not see it as a primary source of deep learning [1, 12, 16].

4 Discussion of Result, Implication for Practice and Conclusion

Advanced programming students use ChatGPT to offload cognitive tasks, speed up coding, generate snippets, and clarify concepts, leveraging their clear goals to assess errors or limitations in its output. A key challenge is guiding novices to use ChatGPT for conceptual exploration rather than mere code copying [1]. Research highlights live coding as vital for building debugging skills, procedural knowledge, and syntax mastery in beginners [11, 13]. To foster deeper engagement, we propose integrating ChatGPT with active learning strategies, such as:

Flipped Classroom Approach. The flipped classroom model requires students to learn course content independently before class, reserving class time for active, deeper engagement. In our proposed implementation, students complete a pre-class review test, which closes once the class session begins, followed by a graded post-class test available afterward. The pre-class test encourages early engagement with the material, while the graded post-class test provides motivation for students to prepare in advance.

Rethinking Assessment Design. An example of this integrated approach is a Java programming assignment from an introductory data structures and algorithms course,

assessing students' procedural knowledge of set data structures as part of a broader summative framework (documented here[3]). We recommend using Koedinger et al.'s (2012) framework for assessment design, which breaks tasks into knowledge components (KCs) reflecting key cognitive processes. These KCs, such as Association, Fact, Category, Production, Rules & Plans, and Principles, help create structured, reproducible assessments and support data-driven insights into effective learning strategies.

Programming Instruction in the Era of ChatGPT. In the ChatGPT era, in-class instruction should prioritize live coding and complex problem-solving to promote deeper learning, targeting Production, Schema, and Skill-level knowledge components (KCs) as outlined by Koedinger et al. (2012). For example, when teaching HTML and CSS basics, a KC include correctly linking an external CSS file and understanding relative vs. absolute file paths. Assessments can focus on:

- Fact: Knowing the syntax to link a CSS file: < link rel = "stylesheet" href = "styles.css" >
- Category: Understanding how file paths and folder structures impact linking.
- Production, Schema, Skill: Applying this knowledge to practical development decisions.

Hands-on exercises in-class could have students create and link HTML and CSS files, move folders to test file path behaviour, and adjust links, accordingly, clarifying the practical differences between relative and absolute paths.

In conclusion, ChatGPT benefits advanced learners but can mislead beginners. We propose a structured framework combining a flipped classroom model with pre- and post-class assessments to promote active learning and engagement. Adapting the KLI model helps define key knowledge components (KCs) for assessing both procedural and conceptual understanding. In-class live coding fosters real-world problem-solving, enhancing learning, particularly in the AI-assisted era. These strategies can improve programming instruction, promote deeper learning, and, like calculators in math, OpenAI could become a valuable tool for scaffolding programming education as its role is better understood and integrated.

References

1. Brender, J., El-Hamamsy, L., Mondada, F., Bumbacher, E. Who's helping who? when students use ChatGPT to Engage in Practice Lab Sessions. In: International Conference on Artificial Intelligence in Education, pp. 235–249. Springer Nature Switzerland, Cham (2024)
2. Denny, P., et al.: Prompt problems: A new programming exercise for the generative aiera. In: Proceedings of the 55th ACM Technical Symposium on Computer Science Education, vol. 1. pp. 296–302 (2024)
3. Dickey, E., Bejarano, A., Garg, C.: AI-Lab: A framework for introducing generative artificial intelligence tools in computer programming courses. SN Comput. Sci. **5**(6), 720 (2024)
4. Dunder, N., Lundborg, S., Wong, J., Viberg, O.: Kattis vs ChatGPT: Assessment and evaluation of programming tasks in the age of artificial intelligence. Proceedings of the 14th Learning Analytics and Knowledge Conference, pp. 821–827 (2024)

[3] https://tinyurl.com/2p83496x.

5. Fuchs, K.: Exploring the opportunities and challenges of nlp models in higher education: is chat gpt a blessing or a curse? Front. Educ. **8**, 1166682 (2023)
6. Hoq, M., Shi, Y., Leinonen, J., Babalola, D., Lynch, C., Akram, B.: Detecting chatgpt-generated code in a cs1 course. LLM@ AIED, pp. 53–63 (2024)
7. Jošt, G., Taneski, V., Karakatič, S.: The impact of large language models on programming education and student learning outcomes. Appl. Sci. **14**(10), 4115 (2024)
8. Keuning, H., Alpizar-Chacon, I., Lykourentzou, I., Beehler, L., Köppe, C. d., Sosnovsky, S.: et al. Students' perceptions and use of generative AI Tools for programming across different computing courses. In: Proceedings of the 24th Koli Calling International Conference on Computing Education Research p. 1012) (2024b)
9. Kiesler, N., Scholz, I., Albrecht, J., Stappert, F., Wienkop, U.: Novice learners of programming and generative ai-prior knowledge matters. In: Proceedings of the 24th Koli Calling International Conference on Computing Education Research, pp. 1–2 (2024)
10. Koedinger, K., Corbett, A., Perfetti, C.: The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cogn. Sci. **36**(5), 757–798 (2012)
11. Lin, Y., Yeh, M., Tan, S.: Teaching programming by revealing thinking process: Watching experts' live coding videos with reflection annotations. IEEE Trans. Educ. **65**(4), 617–627 (2022)
12. Prather, J., et al.: The widening gap: The benefits and harms of generative ai for novice programmers. In: Proceedings of the 2024 ACM Conference on International Computing Education Research, vol. 1 , pp. 469–486 (2024)
13. Raj, A., Patel, J., Halverson, R., Halverson, E.: Role of live-coding in learning introductory programming. In: Proceedings of the 18th kOli Calling International Conference on Computing Education Research, pp. 1–8 (2018)
14. Silva, C., Ramos, F., de Moraes, R., Santos, E.: Chatgpt: challenges and benefits in software programming for higher education. Sustainability **16**(3), 1245 (2024)
15. Singh, H., Tayarani-Najaran, M., Yaqoob, M.: Exploring computer science stu dents' perception of chatgpt in higher education: a descriptive and correlation study. Educ. Sci. **13**(9), 924 (2023)
16. Xu, J., Sun, D., Li, Y.: Using learning analytics in understanding college students' behavior in ChatGPT-facilitated programming learning. In: Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024, pp. 75–82. International Society of the Learning Sciences (2024)
17. Xue, Y., Chen, H., Bai, G., Tairas, R., Huang, Y.: Does chatgpt help with introductory programming? an experiment of students using chatgpt in cs1. In: Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training, pp. 331–341 (2024)
18. Yilmaz, R., Yilmaz, F.: The effect of generative artificial intelligence (ai)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. Comput. Educ. Artifi. Intell. **4**, 100147 (2023)

One Code to Predict Them All: Universal Encoding for Inquiry Modeling

Jade Mai Cock[1(✉)], Valentine Delevaux[1], Ido Roll[2], Richard Davis[3], and Tanja Käser[1]

[1] EPFL, Lausanne, Switzerland
{jade.cock,valentine.delevaux,tanja.kaeser}@epfl.ch
[2] Technion, Haifa, Israel
roll@technion.ac.il
[3] KTH, Stockholm, Sweden
rldavis@kth.se

Abstract. Interactive simulations enhance science education and foster inquiry skills, but their open-ended nature can be cognitively overloading. While adaptive systems offer timely support, research on predicting conceptual understanding in these environments is limited. Most models are simulation-specific, leading to time-consuming and non-generalizable solutions. In this paper, we introduce a universal encoding that converts lower-level interaction data into higher-level features applicable across various open-ended learning environments (OELEs). This encoding aims to offer a general framework to model inquiry across environments and to alleviate challenges such as the "cold start" problem. Our findings demonstrate that models trained on the universal encoding perform comparably to or better than study-specific encodings across multiple contexts. Code is provided in https://github.com/epfl-ml4ed/universal-oele.

Keywords: Interactive Simulations · Inquiry-based Learning · Generalized Encoding · Behavioral Analysis · Conceptual Understanding

1 Introduction

Interactive simulations (ISs) are recognized as effective tools for enhancing science education, fostering inquiry skills [25,36]. They enable a hands-on approach to exploring complex scientific phenomena, deepening conceptual understanding [37]. Yet, the open-ended nature of inquiry can overwhelm students [2,15].

This is where adaptive support can come into play through inferring student knowledge and identify struggling behavior patterns in ISs [35]. To find these patterns, prior works have used clustering techniques to build learner profile with different intervention needs [9,11,13,35], pattern mining to extract recurring behavioral patterns [35], and student models weights' analyses to understand predictive features [6,33]. Unfortunately, while substantial work has been done to develop these models [35], most research on ISs has focused on modeling one specific simulation [5,13,22,26,35,39], making it hard to generalize findings.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 60–67, 2025.
https://doi.org/10.1007/978-3-031-98462-4_8

Only few studies have modeled student interactions across multiple ISs. Those that have [6,9,27] were limited by allowing variation in only one aspect (either the learning task, user interface, or simulation topic) restricting their applicability to a narrow range of scenarios.

There are compelling reasons to pursue more generalized approaches to modeling student behavior in ISs. First, the datasets available for specific simulations or tasks are typically small and cannot be shared for privacy reasons [32]. A generalizable model could provide a robust starting point for new ISs when not enough data is available for training [33], reducing the time and data required to achieve effective personalization. Second, there are significant implications for the learning sciences. By employing the same encoding and model across different ISs, we might surface generalizable insights about student learning across various simulations and contexts, allowing for a more comprehensive understanding of student behavior and learning patterns in ISs.

Building on these motivations, in this paper we introduce a generalized encoding, grounded in inquiry-based learning theory [7,23,34], that adapts to various ISs, regardless of the subject or interface. To explore the usefulness and scalability of our generalized encoding, we analyze five datasets from different environments, tasks, and student populations, showing that our approach structures unstructured log data into meaningful features across environments. With our experiments, we address the following research question: *Can we model student interactions on interactive simulations and extract valuable insights through the proposed generalized encoding?* Our results demonstrate that the proposed generalized encoding performs on par with dataset-specific encodings.

2 Method

To design and evaluate a generalized data representation for ISs, we first collected clickstream data and conceptual knowledge gains of students working with ISs in five different prior studies. We then extracted learning indicators for each in two different ways: dataset-specific indicators from prior work (**specific** encoding) and our generalized data representation (generalized encoding). Finally, we trained deep-learning predictors of conceptual knowledge gains separately for each dataset and encoding to assess the quality of our proposed representation.

2.1 Learning Context

PhET [37] is a collection of ISs that emulate scientific phenomena, allowing students to explore how dependent variables (DVs) change based on (ir)relevant independent variables (IVs). We worked with datasets from 5 different PhET user studies in which participants used an IS to solve an inquiry task. Their conceptual knowledge was assessed through post-tests. All participants gave informed consent, and studies were approved by relevant ethics boards (Nr. 050-2020, 010-2023). Table 1 shows dataset characteristics and participant demographics.

Table 1. Characteristics and demographics of participants per dataset. Gender is self-reported; using *f* for female, *m* for male, *o* for other/non-binary, and *na* for no answer.

Dataset	IS	Inquiry Task	Population	Nr. of Participants	Gender
`rank-cap-uni` [5]	Capacitor	Ranking	University	193	NA
`inq-cap-uni` [7]	Capacitor	Open	University	138	f: 89, m: 47, o: 1, na: 1
`rank-beer-voc` [6]	Beer's Law	Ranking	Apprentices	254	f: 108, m: 135, o: 4, na: 7
`inq-beer-uni` [7]	Beer's Law	Open	University	147	f: 95, m: 50, o: 1, na: 1
`inq-beer-voc` [31]	Beer's Law	Open	Apprentices	159	f: 90, m: 57, o: 2, na:10

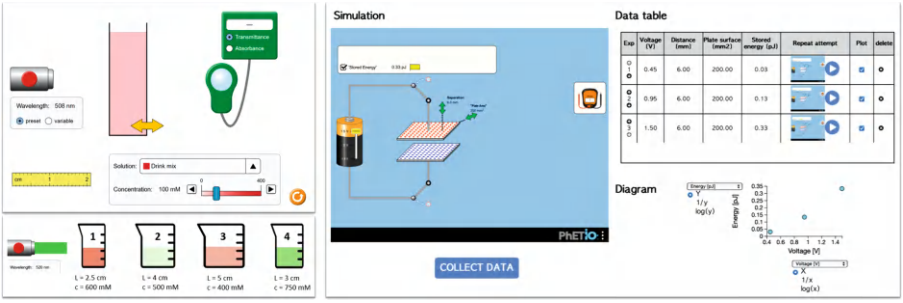


Fig. 1. Standard *Beer's Law* simulation with corresponding ranking task (left) and augmented *Capacitor* simulation used for the open inquiry task (right).

Interactive Simulation. Students participating in the user study used one of two simulations (see Fig. 1). The *Beer's Law* simulation models how light absorbance in a fluid depends on three IVs: flask width (*width*), solution concentration (*concentration*), and the laser/solution color (*color*). The *Capacitor* simulation models stored energy or plate charge based on battery voltage (*voltage*), plate area (*area*), and plate separation (*separation*).

Inquiry Task. In the studies, participants engaged in either goal-oriented inquiry via a ranking task (*Ranking*) or open inquiry (*Open*). In *Ranking*, they ranked four non-reproducible simulation configurations by DV magnitude. In *Open*, they explored DV–IV relationships using an augmented IS with a measurement table and plotting tool. Figure 1 shows the *Ranking* version of the *Beer's Law* simulation (left) and the augmented *Open* version of the *Capacitor* simulation (right).

Population. All studies were conducted with university students (Bachelor's level) or lab technician apprentices in professional education (age: 15–19 years).

2.2 Model Building

From the clickstream data, we extracted each student's sequence of clicks, including (i) mouse press timestamp, (ii) simulation element clicked, (iii) mouse release timestamp, and (iv) any change in DV or IVs. We then derived dataset-specific learning indicators from prior work and applied our generalized encoding to train a model predicting students' conceptual knowledge gains post-interaction.

Specific Encoding. *Specific encodings* are state-of-the-art indicators from prior work, specific to a single study and dataset. For both *ranking* datasets, we used the encoding suggested by [6], while we used the encoding suggested by [7] where each action is encoded into a sequence of *phase-action*.

Generalized Encoding. Our generalized encoding is informed by theoretical and empirical evidence from learning sciences. We created the generalizable learning indicators by describing each "click" using four dimensions: 1) the current state of the simulation (`state`), 2) the inquiry phase the student was in (`phase`), 3) the type of variable the student was interacting with (`variable`), and 4) the duration of the click interaction (`time`). For implementation, we concatenated the one-shot encoded vectors for the `phase`, `state`, and `variable`, scaling the `variable` by the duration `time`.

`State.` The `state` is determined by the configuration of the simulation at the start of the mouse click and can take four different values: *not-observable*, *difficult*, *not-optimal*, and *optimal*. It describes to what degree the present set-up supports inferences about the DV. It is driven by observations that students who manage to set up the simulation correctly from the start perform better later on [6,31].

Not-Observable. A state in which a change in the IV does not result in an observable change in the DV is qualified as *not-observable*. According to prior work, a considerable number of students enter this state when they fail to measure the correct DV [5,6,21] or to set up measuring instruments correctly [6].

Difficult states are configurations of the simulations which make understanding the relations between the IVs and the DV challenging. As shown in [6,7], students exploring in this state after having explored in an optimal state usually grasp more complex concepts. A difficult state in the *Capacitor* IS, would be opening the circuit as it reverses the relations between the independent (area, separation) and dependent (plate charge) variables. In the *Beer's Law* IS, exploring settings in which the absorbance is weak yet necessary to complete the task is also considered difficult (green laser with a green solution).

Not-Optimal. We call a state *not-optimal* when the student can observe the DVs and its changes, but it does not further their understanding of the task at hand. This `state` is only relevant for the *ranking* datasets.

Optimal. We call a configuration *optimal*, if it is not classified as *not-observable*, *difficult*, or *not optimal* through the rules described above. For Beer's Law, this is typically a green laser with a red solution.

`Phase.` Inquiry learning is a cycle of three phases: 1) posing questions and making hypotheses, 2) gathering and analyzing data, and 3) empirically verifying or rejecting the hypotheses emitted in phase 1 [12,18].

Explore. Students often struggle to plan experiments [10] and have difficulties identifying which variables can influence the system [1,4,14,16,19,38]. This phase models whether the student manipulates the simulation in a way that would allow them to learn about how different variables influence the DV. In

the *Ranking* datasets, every interaction is categorized as*explore* as no data collection and visualization tools are available. In the *Open* datasets, every active change made to the parameters in the simulation is considered as *explore*.

Record. Some students struggle to record their progress effectively. Signs of struggle can for example be expressed when students record one or few samples only [20], record the same sample repeatedly [3], or record a vast amount of data points making it difficult to keep track of their data collection. In contrast, methodically recording data points varying one variable at a time is an effective strategy [34]. *Recording* is thus the second phase of the inquiry cycle: gathering and analyzing data. We qualify a click as record whenever a student presses the "record" button in the simulations with a data collection tool.

Analysis. Correctly analyzing the data collected is essential for students to confirm or reject their hypothesis. Insufficient data collection prevents students from verifying hypotheses [28,30] and can lead to confirmation bias [8,16,17]. Understanding what type of data students are analyzing and how much time they have spent doing so is thus important to understand whether students were able to confirm potential (implicit) hypotheses. We qualify a click as analysis whenever the students interact with the graph tool.

`Variable.` The `Variable(s)` encoded at each mouse click depend on the inquiry phase. In the *explore* `phase`, it is the IV the student is manipulating. In the *record* `phase`, it is the `variable(s)` whose values have changed since the last *record* action. Finally, in the *analysis* `phase`, it consists in the variables denoted by the axes of the graph tool. We encode the IVs by their complexity as it has been shown to determine the perceived difficulty of inquiry [7]. *easier* (area, width), *easy* (separation and concentration), *complex* (voltage, color). We also introduce two further categories: breaks (periods of inactivity of 20 s [7]), and other (anything else than breaks and IVs). The former is essential as it can model periods of reflections, plan experiments and conduct inquiry [10,24,29].

Finally, the `time` denotes the difference between the start and end time stamps of the interaction (end time stamp−begin time stamp). Breaks are chunks of 20s [7]. Breaks are an important learning marker which can model periods of reflections, essential to plan experiments and conduct inquiry [10,24,29].

Success Prediction. Our goal was to build a model predicting whether a student will achieve a `high` ($label = 1$) or `low` ($label = 0$) conceptual understanding after solving an inquiry task using the IS (determined in pior work [5–7]). Specifically, we trained a classifier to predict the success *label* of a student based on the generalized (or specific) encoding of their clickstream data from the IS. Following prior work [6], we used a classifier (SA-GRU) consisting of a self-attention layer, followed by a GRU layer, and a classification layer.

3 Experimental Evaluation

We assessed the quality of our generalized encoding by comparing performance of the SA-GRU model between the generalized and specific encodings. We trained

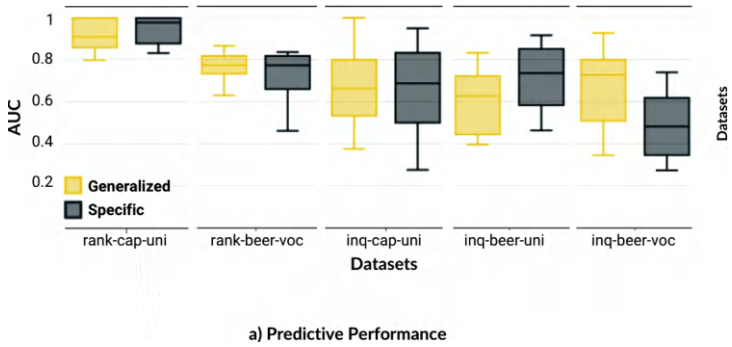


Fig. 2. Predictive performance in terms of AUC of the SA-GRU classifier across five datasets using the **specific** and the generalized encodings.

separate SA-GRUs for each dataset using a nested stratified 10-fold cross validation and optimized over the following hyperparameters: *attention hidden size*: $\{8, 16, 32, 64\}$, *gru hidden size*: $\{8, 16, 32, 64\}$, *epochs*: $\{15, 30, 50\}$. Figure 2 shows model performance in terms of AUC across the five datasets. Pairwise t-tests with a Benjamini-Hochberg correction showed that the generalized encoding significantly outperformed the **specific** encoding on `inq-beer-voc` (ΔAUC: 0.25, $p < .05$). In contrast, the specific encoding performedsignificantly better on `inq-beer-uni` (ΔAUC: 0.11, $p < .05$) only. There were no significant differences across encodings for the other three datasets.

In summary, our universal encoding is broadly effective, while leaving room for specialized approaches in some contexts.

4 Discussion

In this paper, we proposed a generalized encoding grounded in learning science theory to address the scalability and generalizability challenges in modeling student behavior across different ISs and assessed this encoding's quality in terms of predictive performance. We demonstrated the scaling of the generalized encoding across various contexts, with models trained on it performing comparably to or better than study-specific encodings in nearly all cases. Based on our findings, we suggest that the generalized encoding can offer multiple benefits. First, a framework to model students interactions in ISs no matter the topic, interface, student population, and learning task. This is especially useful if no prior work has been conducted on the environment at hand. Second, even if a baseline for the environment already exists, ***generalized encoders*** could still demonstrate value by enabling researchers to use them as pretrained models. Although the generalized encoding performed well across environments, study-specific features occasionally outperformed it, indicating a need for customization in some cases.

Future work could explore hybrid models combining generalized and task-specific features to boost performance and extend the encoding to broader learning environments beyond science simulations.

Acknowledgments. This project was substantially financed by the Swiss State Secretariat for Education, Research and Innovation SERI. Thank you to the Alliance scholarship who funded the trip which initiated this collaboration.

References

1. Al Mamun, M.A., Lawrie, G., Wright, T.: Exploration of learner-content interactions and learning approaches: the role of guided inquiry in the self-directed online environments. Comput. Educ. **178**, 104398 (2022)
2. Alfieri, L., Brooks, P.J., Aldrich, N.J., Tenenbaum, H.R.: Does discovery-based instruction enhance learning? Educ. Psychol. **103**(1) (2011)
3. Buckley, B.C., Gobert, J.D., Horwitz, P.: Using log files to track students' model-based inquiry. In: Proceedings of ISLS (2006)
4. Chinn, C.A., Brewer, W.F.: The role of anomalous data in knowledge acquisition: a theoretical framework and implications for science instruction. Rev. Educ. Res. **63**(1) (1993)
5. Cock, J.M., Marras, M., Giang, C., Kaser, T.: Early prediction of conceptual understanding in interactive simulations. In: Proceedings of EDM (2021)
6. Cock, J.M., Marras, M., Giang, C., Käser, T.: Generalisable methods for early prediction in interactive simulations for education. In: Proceedings of EDM (2022)
7. Cock, J.M., Roll, I., Käser, T.: Consistency of inquiry strategies across subsequent activities in different domains. In: Proceedings of AIED (2023)
8. Dunbar, K.: Concept discovery in a scientific domain. Cogn. Sci. **17**(3) (1993)
9. Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., Goldhammer, F.: Exploring behavioural patterns during complex problem-solving. JCAL **36**(6) (2020)
10. Glaser, R., Schauble, L., Raghavan, K., Zeitz, C.: Scientific Reasoning Across Different Domains. Springer, Heidelberg (1992)
11. Gobert, J.D., Kim, Y.J., Sao Pedro, M.A., Kennedy, M., Betts, C.G.: Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. Think. Skills Creativity **18** (2015)
12. Gobert, J.D., Sao Pedro, M.A.: Digital assessment environments for scientific inquiry practices. In: The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications, pp. 345–371 (2016)
13. Käser, T., Schwartz, D.L.: Exploring neural network models for the classification of students in highly interactive environments. In: Proceedings of EDM. ERIC (2019)
14. Keselman, A.: Supporting inquiry learning by promoting normative understanding of multivariable causality. J. Res. Sci. Teach. **40**(9) (2003)
15. Kirschner, P., Sweller, J., Clark, R.E.: Why unguided learning does not work: an analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. Educ. Psychol. **41**(2) (2006)
16. Klahr, D., Dunbar, K.: Dual space search during scientific reasoning. Cogn. Sci. **12**(1) (1988)
17. Klayman, J., Ha, Y.W.: Confirmation, disconfirmation, and information in hypothesis testing. Psychol. Rev. **94**(2) (1987)
18. Kuhn, D., Black, J., Keselman, A., Kaplan, D.: The development of cognitive skills to support inquiry learning. Cogn. Instr. **18**(4) (2000)
19. Kuhn, D., et al.: Strategies of knowledge acquisition. Monogr. Soc. Res. Child Dev. **157** (1995)

20. Kuhn, D., Schauble, L., Garcia-Mila, M.: Cross-domain development of scientific reasoning. Cogn. Instr. **9**(4) (1992)
21. López-Tavares, D., Perkins, K., Reid, S., Kauzmann, M., Aguirre-Vélez, C.: Dashboard to evaluate student engagement with interactive simulations. In: Proceedings of PERC (2018)
22. McBride, E., Vitale, J.M., Gogel, H., Martinez, M.M., Pardos, Z., Linn, M.C.: Predicting student learning using log data from interactive simulations on climate change. In: Proceedings of L@S (2016)
23. Pedro, M., Jiang, Y., Paquette, L., Baker, R., Gobert, J.: Identifying transfer of inquiry skills across physical science simulations using educational data mining. In: Proceedings of ICLS (2014)
24. Perez, S., et al.: Identifying productive inquiry in virtual labs using sequence mining. In: Proceedings of AIED (2017)
25. Perkins, K., Moore, E., Podolefsky, N., Lancaster, K., Denison, C.: Towards research-based strategies for using PhET simulations in middle school physical science classes. In: Proceedings of AIP (2012)
26. Poitras, E., Butcher, K.R., Orr, M., Hudson, M.A., Larson, M.: Predicting student understanding by modeling interactive exploration of evidence during an online science investigation. Inter. Learn. Environ. **30**(5) (2022)
27. Roll, I., et al.: Understanding the impact of guiding inquiry: the relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning. Instr. Sci. **46** (2018)
28. Schauble, L., Glaser, R., Raghavan, K., Reiner, M.: Causal models and experimentation strategies in scientific reasoning. J. Learn. Sci. **1**(2) (1991)
29. Shih, B., Koedinger, K.R., Scheines, R.: A response time model for bottom-out hints as worked examples. In: Proceedings of EDM (2008)
30. Shute, V.J., Glaser, R.: A large-scale evaluation of an intelligent discovery world: Smithtown. Inter. Learn. Environ. **1**(1) (1990)
31. Shved, E., Bumbacher, E., Mejia-Domenzain, P., Kapur, M., Käser, T.: Teaching and measuring multidimensional inquiry skills using interactive simulations. In: Proceedings of AIED (2024)
32. Stinar, F., Xiong, Z., Bosch, N.: An approach to improve "k"-anonymization practices in educational data mining. JEDM **16**(1) (2024)
33. Swamy, V., Marras, M., Käser, T.: Meta transfer learning for early success prediction in MOOCs. In: Proceedings of L@S (2022)
34. Tschirgi, J.E.: Sensible reasoning: a hypothesis about hypotheses. Child Dev. (1980)
35. Wang, K.D., Cock, J.M., Käser, T., Bumbacher, E.: A systematic review of empirical studies using log data from open-ended learning environments to measure science and engineering practices. BJET **54**(1) (2023)
36. Wieman, C.E., Adams, W.K., Loeblein, P., Perkins, K.K.: Teaching physics using PhET simulations. Phys. Teach. **48**(4) (2010)
37. Wieman, C.E., Adams, W.K., Perkins, K.K.: PhET: simulations that enhance learning. Science **322**(5902) (2008)
38. Wu, H.K., Hsieh, C.E.: Developing sixth graders' inquiry skills to construct explanations in inquiry-based learning environments. Int. J. Sci. Educ. **28**(11) (2006)
39. Xing, W., Li, C., Chen, G., Huang, X., Chao, J., Massicotte, J., Xie, C.: Automatic assessment of students' engineering design performance using a Bayesian network model. J. Educ. Comput. Res. **59**(2) (2021)



"I Would Share It, But..." Exploring Ways to Optimize the Privacy-Personalization Trade-Off in Intelligent Tutoring Systems

Adrienn Toth(✉), Neele Roch, Noé Zufferey, and Verena Zimmermann

ETH Zurich, Zurich, Switzerland
{adrienn.toth,neele.roch,noe.zufferey,verena.zimmermann}@gess.ethz.ch

Abstract. Personalized learning is increasingly improving through AI-enhanced intelligent tutoring systems (ITS). However, ethical and privacy aspects of ITS, such as the privacy-personalization trade-off, are under-researched. We conducted an interview study with $N = 32$ university students and found that students were not largely concerned about privacy as such but implications of data collection for social aspects, their education, and society at large. Students preferred ITS to supplement rather than replace human teachers in certain tasks, leveraging the benefits of both. We provide recommendations on addressing the identified concerns for ITS design and successful integration into a curriculum, for which AI literacy and student autonomy emerge as crucial factors.

Keywords: Intelligent Tutoring System · Human-centered Artificial Intelligence · Data Privacy · Higher Education

1 Introduction

Personalized learning tools can substantially help students in their learning process while enabling professors to more efficiently use the scarce in-person time in classrooms. While intelligent tutoring systems (ITS) [22] are increasingly considering human-centered aspects such as usability [2,20], students are generally absent from the ITS development process [3] and are often only recruited for post-deployment evaluation [3]. In addition, there is little attention to aspects outside of learning outcomes, such as the question of data privacy and ethics [23,30,35]. As personalization relies on personal data processing (e.g., students' performance, behavior, and personal data), it is crucial to take the privacy-personalization trade-off (PPTo) into account when developing ITS [4], involving actual students in the process. In this article, we aim to answer the following resulting research question: *In the context of higher education, what factors play a role in students' assessments of the PPTo associated with ITS, and how do these factors shape their decision-making?*

We conducted an interview study with $N = 32$ university students to understand their requirements and concerns regarding ITS and the PPTo. Students'

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 68–76, 2025.
https://doi.org/10.1007/978-3-031-98462-4_9

concerns range from social, educational, to high-level societal implications of ITS, which can be addressed in different ways. Based on our findings, we derive concrete design recommendations for future ITS, focusing on technical functionalities and successful integration into course curricula.

2 Related Work

There is a growing interest in using AI-related education techniques [17,35]. ITS provide students with tailored feedback, and instruction such as hints [22]. Various interventions, like high and low scaffolding, [7], how ITS could support self-regulated learning [24], how gamification can be integrated into an ITS [21,24], and comparing ITS to traditional methods [5] have shaped ITS development. ITS can use students' data to support collaborative learning [6,34], and for time management [1]. However, privacy concerns, as well as a lack of understanding of how ITS work, generally result in not trusting ITS [11,27,29], and therefore hinder educational institutions from adopting these tools [8]. Research in this area is limited [23,35], even though students could provide crucial insights to support the design process [3,28], and privacy research calls for improved human-centered approaches [18,25]. Only a few recent works focus specifically on designing systems considering privacy [33]. Kwapisz et al. [19] and Greenhalgh et al. [14] explored privacy concerns of students regarding data sharing practices of learning management systems, while others focused on analytics dashboards [31] or collaboration analytics [36]. Our study takes a first step toward addressing the lack of student involvement in ITS design, aiming to derive concrete recommendations for the data handling of future ITS.

3 Method

We conducted individual interviews (45–60 min. each) with $N = 32$ students from ETH Zurich and the University of Zurich, from various fields of study. They were all fluent in English and had already completed at least three semesters. They were 24 years old in average ($SD = 2$, $min = 20$, $max = 28$), 15 were men, 17 were women, 13 were Bachelor students, 18 were Master students, and one was a PhD student.

Each interview consisted of four main sections aiming to (1) capture students' definition of personalization, their main usage context, and their beliefs, (2) explore how personalization could improve educational practices, (3) critically reflect on the benefits and risks associated with personalized ITS, and (4) explore the willingness of data-sharing with respect to the types of data, the students' need for transparency and control over data collection, storage, and processing. In this short paper, only the results of the two last points are presented in detail with other crucial insights mentioned where necessary.

Data Analysis. The interviews took place online with audio-recording and transcription for anonymized analysis using *trint*, yet manually validated to ensure correctness. The transcripts were analyzed using MAXQDA24. One

researcher first generated an initial code structure that was revised based on other researchers' feedback. This revision process was repeated multiple times, involving a total of four researchers.

Ethical and Open Science Considerations. Participants signed a consent form and received compensation in line with our institution's suggestions. Our method followed ethical guidelines for psychological research and has been IRB-approved. The interview guideline and supporting materials are available online[1]. Interview transcripts are made available upon request.

4 Results

We developed a framework summarizing the key concerns toward ITS and potential solutions based on the interview outcomes (see Fig. 1).

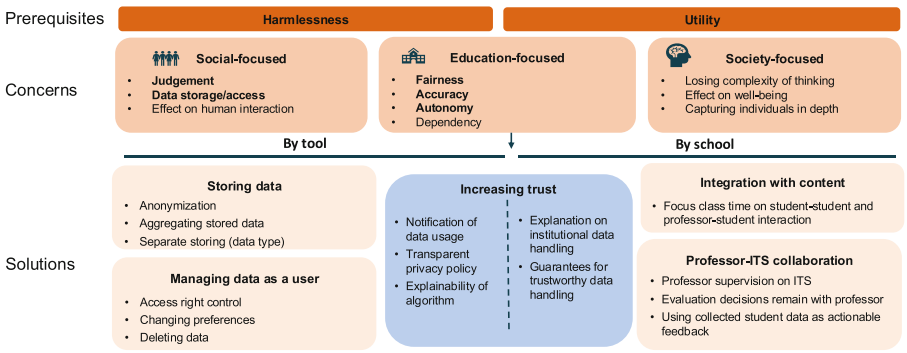


Fig. 1. Concerns toward ITS and potential solutions. Concerns directly addressable by ITS features (in bold) and those requiring measures from the institution are shown.

Unmet Prerequisites Can Prevent Students from Using an ITS. Participants generally showed a positive attitude toward ITS, but described aspects, such as certain ITS features being *harmful* for their learning process ("*I think it's [...] maybe harmful [...] to not be involved in the process at all.*" - P12) to prevent them from using these systems. *Utility* was also frequently mentioned, when discussing which data they "*wouldn't see [...] necessary for the service.*" (P24) and thus wouldn't want to be collected, despite not having particular ethical concerns.

Social Concerns were Common. Many participants' concerns were related to data access and use by others. *Judgement* refers to concerns about lecturers' perception of the data and potential negative effects, e.g., "*I still think that [the lecturer] could then be biased by what I've done in a certain test.*" (P5) or what other students think, e.g., "*not everyone is confident being talked about their*

[1] Data available on OSF.

results in front of everyone" (P7). *Data storage/access* refers to who can access educational data; while students generally accepted staff accessing their data, they opposed sharing it with peers or external parties. Finally, *effect on human interaction* related to how ITS use can affect the overall learning experience, e.g., how it would reduce interaction with peers.

Concerns about the ITS' Effect on Educational Outcomes. Students were worried about *fairness*, noting that extensive access to study data could lead lecturers to be biased (e.g., *"I would provide [time and frequency of study sessions], but only if then they're not graded."* - P30). Additionally, concerns about whether the ITS has enough data to make *accurate* predictions were raised (e.g., *"How do we know that there are no blind spots on the algorithm side?"* - P8). Students believe that *autonomously* regulating their study process should be something they do independently and were *"afraid to lose the ability to do it on my own."* (P19), which is why they prefer to be in control of their study progress and only use the support of ITS on the task-level. Finally, *dependency* on the tool, which refers to students not being able to efficiently study without an ITS, was found to prevent them from learning the soft skills necessary in a work environment later, as also described in previous research [13].

Concerns on the Effects of an ITS on Society. The final category is related to the widespread usage of personalized tools on society, such as people *"los[ing] the ability to create a thought"* (P13) or seeing each other as complex beings as an ITS would fail to *capture individuals in depth*. Finally, some were concerned that an ITS could have a negative effect on their mental health due to comparison with others, constantly being reminded of their weaknesses, or reduced social interaction. These concerns, although rare, still show the importance of addressing general worries toward AI systems, not only ITS.

Potential Solutions Exist both on the ITS and Institution Level. The acceptance of ITS depended heavily on its integration into the students curriculum, therefore not all concerns can be tackled by ITS features alone. Within ITS, the best way of *storing data* depended on perceived data sensitivity. A lot of students were only slightly concerned about sharing data as long as it is anonymized, or, in the case of their personal data, stored separately from their detailed results and behavioral data. More concerned participants preferred the data to only be available in aggregated format from a class *"for the professor to understand which topic is less clear or if he should focus more on that [topic]"* (P5), without providing their individual records. Some suggested restricting access to certain recipients or data types individually, whereas others preferred to choose the functionalities and the system to only collect the data necessary for those, e.g., *"I imagine it would be tied to some services, what data you need to use or you need to provide"* (P24). Almost all participants considered being able to *change their data-sharing preferences* easily and *deleting* parts or all of their data from the system very important. They illustrated how continuous transparency could *increase their trust* in the tool. They also mentioned that understanding better *"how [the tool is] coming to the suggestions"* (P13) would

convince them of the accuracy of the ITS. However, certain recommendations require the institution's involvement. In case an ITS would be a core element of their course, some participants expected the information to come directly from the university instead of the tool, saying "*it would make sense to maybe have a [...] class that just focuses on what this tool [does].*" (P10). Also, the ITS and classroom teaching have to complement each other instead of just coexisting or replacing the other ("*I feel like they could work well together instead of like against each other*" - P6).

5 Discussion

Personalization should Happen at the Task Level, while Students Remain in Control Over what and when to Study. Based on students' concerns related to losing autonomy over their learning and becoming overly dependent on the ITS, it is not recommended to design the ITS in a way that takes high-level control of the learning process and behaviors. Instead, the personalization should remain on the task level, providing step-by-step guidance and feedback. This helps students maximize their learning outcomes on the micro-level, while also helping them with self-regulated learning, e.g., supported by the integration of pedagogical agents shown to be successful in improving learning outcomes [12]. This can prevent harmful cognitive offloading that can cause the decline of critical thinking skills if students are too dependent on the tool [13] and reduce the enjoyment of learning [15].

The Tool should be Functional Regardless of a Student's Data Sharing Preferences. Tools should ensure that users' privacy preferences are adhered to without limiting their ability to use the system [16], as privacy preferences vary greatly depending on context [18]. This is in line with some students' thought that if an ITS was used as part of a course, they should still be free to grant and revoke consent to data collection without suffering major drawbacks. This could be achieved by integrating different'learning modes' associated with different data collection practices. For example, students can target their weak areas with relevant exercises, but for that rich interaction data has to be collected from them. However, if they wish to only review material without specific focus on their weaknesses, then no or little data is required.

Information on Data Collection should be Provided from a Learning, not a Legal Perspective. Students showed trust in their institution to store and handle their data securely, reflecting previous studies [23]. Therefore, a traditional privacy notice might not be the most efficient way of informing them. Instead, as students named multiple concerns related to how data collected by an ITS could negatively affect them within the institution (e.g. unfair evaluations, judgment from classmates if they can access the data), designing a student-centered notice and consent system could be beneficial. For example, instead of asking "Do you consent to provide your data about the time and duration of your study sessions?", the questions should be "Would you like to get feedback

related to your study schedule and productivity?" along with an explanation of the data collection implications of the decision. This would also better equip students to make similar future decisions and improve AI literacy, which many in higher education are currently lacking [14].

Consider the ITS as a Complementary Asset to the Course, not as a Standalone Tool. Many participants agreed that there are some advantages to learning from a human which digital tools simply cannot substitute: showing empathy, explaining new material in the most suitable way, and engaging in reflective discussions. However, an ITS is advantaged in its availability and capacity to provide an infinite amount of exercises. Therefore, instead of looking at ITS as a substitute for human teaching, institutions should take advantage of both when designing courses. Teaching new, complex materials are best to stay in the hands of human lecturers who can engage students the best. Meanwhile, ITS could serve as a tool for individual practice and revision, the required amount of which might differ greatly between students, so they can benefit greatly from continuous personalized feedback [31,32]. Previous research has also demonstrated the success of combining human mentoring with AI tutoring [9].

Be Clear about the Capabilities of the Tool and Its Role in the Curriculum. Understandably, most students optimize their learning strategy such that they get the best grades, therefore they might not use an ITS seriously if it is not part of their evaluation [36]. At the same time, our results show that students are uncomfortable with being evaluated based on data from an ITS, therefore it is not advised that institutions do so. They can however make sure to clarify the role of the ITS in the curriculum (e.g., independent practice of complex exercises) and how that is useful to achieve the desired learning outcomes to boost engagement and eliminate potential concerns.

Make the most of ITS by Providing Feedback to Teachers on Class Performance. Despite a focus on the student perspective, 16 participants mentioned that the ITS data could provide valuable insights to lecturers about how students interact with the course. Most of them preferred it to be in an aggregated format based on the whole class' data instead of making rich interaction data from individuals available. [10] came to similar conclusions regarding visualizing cognitive-affective states in online learning. The exact ways of how this information could be most effectively utilized on teacher dashboards is definitely worth further investigation [26].

5.1 Conclusion and Future Work

The study revealed that the question of privacy cannot be fully separated from other societal and ethical factors once we discuss using an ITS integrated into a course. Therefore, the final recommendations cover a wider scope than just data handling and suggest focusing on (1) ways to ensure flexibility and autonomy of students in using and providing data to the ITS, (2) efficiently informing students about implications of data handling not just through the ITS, but the

institution itself, and (3) designing curricula in a way that benefits of an ITS as well as human teaching are maximized.

However, there are some limitations to this study. The interviews were conducted with only local students in Switzerland. Their answers might vary from individuals studying in other countries with different education systems and data protection regulations. Future work could conduct similar studies in other cultural and educational environments. Also, the interview questions referred to the hypothetical scenario of an ITS used in university courses, and their declared opinion may be different from their actual behavior. Future studies could investigate students' perceptions in the context of actual interactions with ITS.

Acknowledgements. This research has been supported by the EPFL-ETHZ Joint Doctoral Program in the Learning Sciences and the Innovedum Fund at ETH Zurich.

References

1. Akyuz, Y.: Effects of intelligent tutoring systems (its) on personalized learning (pl). Creat. Ed. **11**(0606) (2020)
2. Alfredo, R., Echeverria, V., Jin, Y., Swiecki, Z., Gašević, D., Martinez-Maldonado, R.: Slade: a method for designing human-centred learning analytics systems. In: Proceedings of the 14th LA and Knowledge Conference (2024)
3. Alfredo, R., et al.: Human-centred learning analytics and AI in education: a systematic literature review. artificial intelligence. Comp. Edu. (2024)
4. Awad, N., Krishnan, M.: The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. MIS Quart. **30** (2006)
5. Borchers, C., et al.: What makes problem-solving practice effective? comparing paper and AI tutoring. In: Responsive and Sustainable Educational Futures, Springer (2023)
6. Borchers, C., Yang, K., Lin, J., Rummel, N., Koedinger, K.R., Aleven, V.: Combining dialog acts and skill modeling: what chat interactions enhance learning rates during AI-supported peer tutoring? (2024)
7. Borek, A., Mclaren, B., Karabinos, M., Yaron, D.: How much assistance is helpful to students in discovery learning? EC-TEL 2009, Springer (2009)
8. Cermakova, A.L., Clary, A., Havinga, B.: Evaluation mechanisms: a way to increase trust in educational technologies? (2024), https://www.edtechstrategylab.org/s/2024_Deep-Dive-I_Evaluation-Mechanisms_EdTech-Strategy-Lab.pdf
9. Chine, D.R., et al.: Educational equity through combined human-AI personalization: a propensity matching evaluation. In: AI in Education (2022)
10. Corza-Vargas, V.M., Martinez-Maldonado, R., Escalante-Ramirez, B., Olveres-Montiel, J.: Students' ethical, privacy, design, and cultural perspectives on visualizing cognitive-affective states in online learning. J. L.A. (2024)
11. Cukurova, M., Miao, X., Brooker, R.: Adoption of artificial intelligence in schools: unveiling factors influencing teachers' engagement. In: AIED (2023)
12. Dever, D.A., Sonnenfeld, N.A., Wiedbusch, M.D., Azevedo, R.: Pedagogical agent support and its relationship to learners' self-regulated learning strategy use with an intelligent tutoring system. In: AIED (2022)

13. Gerlich, M.: Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies **15**(11) (2025)
14. Greenhalgh, S.P., DiGiacomo, D.K., Barriage, S.: Platforms, perceptions, and privacy: ethical implications of student conflation of educational technologies. Info. Learn. Sci. **124**(9/10) (2023)
15. Han, B., Nawaz, S., Buchanan, G., McKay, D.: Ethical and pedagogical impacts of AI in education. In: AIED (2023)
16. Hann, I.H., Hui, K.L., Lee, S.Y.T., Png, I.P.: Overcoming online information privacy concerns: an information-processing theory approach. J. Man. Inf. Sys. **24**(2) (2007)
17. HolonIQ: artificial intelligence in education. Surv. Insights (2023), https://www.holoniq.com/notes/artificial-intelligence-in-education-2023-survey-insights
18. Knijnenburg, B.P., et al.: User-tailored privacy. In: Modern Socio-Technical Perspectives on Privacy (2022)
19. Kwapisz, M.B., Kohli, A., Rajivan, P.: Privacy concerns of student data shared with instructors in an online learning management system. In: CHI 2024 (2024)
20. Lang, C., Davis, L.: Learning analytics and stakeholder inclusion: what do we mean when we say "human-centered"? In: LAK23 (2023)
21. Long, Y., Aleven, V.: Gamification of joint student/system control over problem selection in a linear equation tutor. In: ITS, p. 378–387 (2014)
22. Mousavinasab, E., Zarifsanaiey, N.R., Niakan Kalhori, S., Rakhshan, M., Keikha, L., Ghazi Saeedi, M.: Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. Int. Lear. Env. **29**(1) (2021)
23. Mutimukwe, C., Viberg, O., Oberg, L.M., Cerratto-Pargman, T.: Students' privacy concerns in learning analytics: model development. Brit. J. Ed. Tech. **53**(4) (2022)
24. Nagashima, T., et al.: Designing playful intelligent tutoring software to support engaging and effective algebra learning. In: EC-TEL 2022 (2022)
25. Nissenbaum, H.: A contextual approach to privacy online. Daedalus **140**(4) (2011)
26. de Quincey, E., Parker, M.: User-centred guidelines for the design of curriculum analytics dashboards (2021)
27. Renz, A., Vladova, G.: Reinvigorating the discourse on human-centered artificial intelligence in educational technologies. Tech. Inn. Mgmt. Rev. **11** (2021)
28. Sun, K., Mhaidli, A.H., Watel, S., Brooks, C.A., Schaub, F.: It's my data! tensions among stakeholders of a learning analytics dashboard. In: CHI 2019 (2019)
29. Tan, M., Subramonyam, H.: More than model documentation: uncovering teachers' bespoke information needs for informed classroom integration of chatgpt. In: CHI 2024 (2024)
30. Viberg, O., Mutimukwe, C., Grönlund, A.: Privacy in la research: understanding the field to improve the practice. J. L.A. **9** (2022)
31. Vreugd, L., van Leeuwen, A., Jansen, R., Schaaf, M.: Learning analytics dashboard design and evaluation to support student self-regulation of study behaviour. J. L.A. (2024)
32. Weydner-Volkmann, S., Bär, D.: Student autonomy and learning analytics: philosophical considerations for designing feedback tools. J. L.A. **11**(33) (2024)
33. Williamson, B., Bayne, S., Shay, S.: The datafication of teaching in higher education: critical issues and perspectives. Teach. in HiEd **25**(4) (2020)
34. Yang, K.B., et al.: Pair-up: prototyping human-AI co-orchestration of dynamic transitions between individual and collaborative learning in the classroom. In: CHI 2023 (2023)

35. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int. J. Ed. Tech. in HiEd **16**(1) (2019)
36. Zhou, Q., et al.: Investigating students' experiences with collaboration analytics for remote group meetings. In: AIED 2021 (2021)

A Participatory Strategy for AI Ethics in Education and Rehabilitation Grounded in the Capability Approach

Valeria Cesaroni[1], Eleonora Pasqua[4,5], Piercosma Bisconti[3], and Martina Galletti[2,3(✉)]

[1] University of Perugia, Perugia, Italy
valeria.cesaroni@dottorandi.unipg.it
[2] Sony Computer Sciences Laboratories, Paris, France
martina.galletti@sony.com
[3] Dipartimento di Ingegneria Informatica, Automatica e Gestionale "A. Ruberti", Sapienza University of Rome, Rome, Italy
[4] Centro Ricerca e Cura di Roma, Rome, Italy
[5] Dipartimento di Organi di Senso, Sapienza University of Rome, Rome, Italy

Abstract. AI-based technologies have significant potential to enhance inclusive education and clinical-rehabilitative contexts for children with Special Educational Needs and Disabilities. AI can enhance learning experiences, empower students, and support both teachers and rehabilitators. However, their usage presents challenges that require a systemic-ecological vision, ethical considerations, and participatory research. Therefore, research and technological development must be rooted in a strong ethical-theoretical framework. The Capability Approach - a theoretical model of disability, human vulnerability, and inclusion - offers a more relevant perspective on functionality, effectiveness, and technological adequacy in inclusive learning environments. In this paper, we propose a participatory research strategy with different stakeholders through a case study on the ARTIS Project, which develops an AI-enriched interface to support children with text comprehension difficulties. Our research strategy integrates ethical, educational, clinical, and technological expertise in designing and implementing AI-based technologies for children's learning environments through focus groups and collaborative design sessions. We believe that this holistic approach to AI adoption in education can help bridge the gap between technological innovation and ethical responsibility.

Keywords: Participatory Design · Special Education Needs and Disabilities · Capability Approach · AI ethics · NLP

1 Introduction

The integration of artificial intelligence (AI) in educational and clinical contexts necessitates a multidimensional approach that considers cognitive, relational, and ethical dimensions. First, the literature underscores that artificial intelligence, by directly influencing cognitive and relational structures [3] can

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 77–84, 2025.
https://doi.org/10.1007/978-3-031-98462-4_10

play a crucial role in the scaffolding function. This, in turn, plays a pivotal role in proximal development, significantly shaping subjectification, qualification, and socialization processes [4]. Second, an inclusive perspective requires an ecological-systemic approach to the technology design that can take into account the interaction between individuals, technologies, and their broader social environments [5]. For these reasons, the development and the implementation of such technologies must be anchored within an ethical framework which is appropriate and grounded in a theoretical model suitable for childhood, disability, and human vulnerability [6]. Moreover, it is essential to design the development of the technology by engaging diverse stakeholders, each bringing distinct epistemic and methodological perspectives.

This study makes several key contributions. Drawing inspiration from the ARTIS Project [7,8], which develops an AI-based interface to rehabilitate children with text comprehension difficulties. The study contributes to: (I) Propose a theoretical framework based on the capability approach; (II) Introduce a co-creation strategy applicable to various contexts; (III) Present implementation results using ARTIS as a case study. The study illustrates a research and development pathway linking inclusion, technological innovation and critical assessment of personal capabilities, proposing a holistic and participatory model for ethical and inclusive AI development aimed at supporting vulnerable learners.

2 Related Work

The field of AI in Education (AIEd) has grown alongside developments in big data and adaptive technologies, with AI typically supporting students, teachers, or institutions [9]. While widely adopted in mainstream education for administration and personalized learning [10], its role in inclusive and special education remains limited and often instrumental [11]. Scholars have criticized the lack of ethical and pedagogical consideration in AI integration [12] noting that many tools are not designed with the needs of learners with disabilities in mind [9].

Recent research has moved beyond AI's technical functions to highlight its ethical and social implications in education [13]. AI ethics research often remains narrowly technical—focused on fairness, bias, and transparency—while neglecting deeper pedagogical and inclusive concerns [5,14]. This functionalist approach sidelines critical, theory-driven perspectives and rarely involves educators or specialists in early design stages [15,16]. The lack of an ecological and systemic outlook results in AI tools misaligned with pedagogical values and classroom realities, especially in inclusive education contexts.

Finally, AI has demonstrated promise in diagnosing neurodevelopmental disorders through machine learning analysis of MRI [17] and EEG [18] data, providing effective screening support [2]. Despite benefits for clinical decision-making [19,20] and personalized learning [21], AI's role in structured therapeutic interventions and long-term rehabilitation remains underexplored. Research on AI in structured therapeutic interventions is still limited [22], and few programs provide adequate training for educators and clinicians.

This study addresses the lack of AI-driven rehabilitation tailored to diverse cognitive needs and the shortage of training for practitioners by positioning AI as a core component of clinical and educational interventions that support cognitive and linguistic development. The ARTIS project [7,8] exemplifies a research approach based on an ethical framework of human development and provides an example of the context-sensitive use of technology. Based on van Dijk's neuropsycholinguistic model [23,24], and developed with speech therapists and tested in public and clinical settings [8], ARTIS aims to support the rehabilitation process.

3 The Theoretical Framework

Existing ethical frameworks for AI, often focused on risk mitigation (e.g. fairness, bias, privacy), prove inadequate in educational and rehabilitative contexts involving children. These models tend to be object-oriented and contractualist, assuming autonomous, equal agents [6]. Such assumptions neglect the fundamental conditions of vulnerability, dependence, and asymmetry that characterize childhood and disability [6]. We argue for a shift toward a holistic perspective that sees technology as a sociotechnical mediator rather than a neutral tool [1]. This requires an ethical model grounded in vulnerability, interdependence, and the complex interplay between individuals, technology, and their environments. AI, especially in childhood contexts, challenges instrumentalist perspectives and demands a rethinking of inclusion through a more relational and systemic lens. To address this gap, we adopt the Capability Approach (CA) as a foundational framework. Originating with Amartya Sen [26] and expanded by Martha Nussbaum [25], the CA redefines justice not in terms of equal resources, but of real freedoms - what individuals are effectively able to do and be. It emphasizes conversion factors - individual (e.g. impairments), social (e.g. norms, expectations), and environmental (e.g. infrastructures) - that shape how resources like technology translate into meaningful opportunities (capabilities and functionings) [28]. In this light, AI must be evaluated not just on its technical performance but on how it shapes learning, inclusion, and well-being. It calls for research questions attentive to human diversity, contextual factors, and the mediating role of technology in subject formation. With this approach, we can interpret the interplay between technology, context, and users' situated conditions in a more nuanced way, enabling a deeper, more critical understanding of the role technology plays in inclusive processes. Especially relevant to AI - defined as a third-level technology [27] for its pervasive cognitive and hermeneutic role - the CA supports a systemic, hybridized view of educational processes.

In the ARTIS project, the Capability Approach has shaped the formulation of key research questions that guide both design and evaluation. These include: whether the technology is developmentally appropriate for its intended users; what user model underpins the system's design choices; and which capabilities are promoted or, conversely, constrained by the intervention. Additional considerations include how the processes of subjectification, qualification, and socialization are influenced by the technology, and what forms of human–AI interaction

best support learner agency. In sum, the CA offers a rich ethical and epistemological lens to reframe inclusion and justice in the age of AI, moving beyond risk management toward a vision of relational, situated, and developmental justice.

4 The Co-creation Framework: Principles and Practical Applications

4.1 General Principles

The ARTIS project has developed a participatory research strategy for interface design that combines the Capability Approach with Value Sensitive Design [8, 29]. This ensures AI-driven educational tools align with real-world needs and pedagogical goals by involving educators, clinicians, and developers from the start. Unlike traditional methods that seek input after development, ARTIS emphasizes early, interdisciplinary co-creation. This approach, adaptable beyond ARTIS, serves as a model for ethical, effective AI in education and rehabilitation. In this section, we introduce CORE – the Co-Creation framework for Optimized, Responsible, and Ethical AI Development – our second technical contribution, detailing its structure, methodology, and principles.

CORE unfolds in three phases to align diverse perspectives across tech, clinical, and educational fields. Phase one builds a proof of concept. Phase two refines the interface through ethical, pedagogical, and functional insights grounded in focus groups and interviews with developers and clinicians. Phase three evaluates the tool's impact and ethics in real-world settings, using the Capability Approach to ensure meaningful, responsible learning outcomes. To provide a clear overview, Table 1 presents the CORE model as a structured framework for guiding AI development in educational and therapeutic contexts. It highlights six key dimensions that support ethical, pedagogical, and human-centered design, implementation, and evaluation. Each dimension includes key points and guiding questions to prompt critical reflection and decision-making. Results from the ARTIS project are discussed in detail in Sect. 4.2.

The CORE framework is grounded in the following guiding principles:

1. **Child-Centered Design** – Defines objectives, age-appropriateness, and developmental alignment while mapping pedagogical models and impacted capabilities.
2. **Human Interaction and Context** – Specifies use cases and aligns expectations across educational and clinical environments.
3. **Participatory Engagement** – Involves stakeholders (e.g., teachers, families, clinicians, children), addresses value tensions, and supports iterative co-design.
4. **Agency** – Examines the balance between automation and human oversight, autonomy, privacy and its impact on users' self-efficacy and working conditions.
5. **AI Quality and Robustness** – Focuses on data quality, fairness, transparency, and privacy, especially in sensitive child-focused applications.

Table 1. This table outlines a structured framework for developing AI technologies. The table identifies six fundamental aspects essential for evaluating AI applications in child-focused contexts, highlights critical factors within each dimension, and poses reflective questions to ensure ethical and developmental alignment.

Conceptual Dimension	Key Considerations	Guiding Questions
Children-Centered Framework	– Clear definition of technology's objectives – Alignment with developmental stages – Identification of impacted capabilities – Explicit pedagogical models	– What are the intended learning outcomes? – How does it support developmental stages? – What cognitive and social skills are targeted? – Which pedagogical models underpin the design?
Human Interaction and Setting	– User experience and accessibility – Clear interaction models	– What are the primary use cases? – How can stakeholder alignment be achieved?
Democratic Engagement	– Inclusive stakeholder involvement – Addressing value conflicts – Systematic feedback integration	–How can educators, parents, and therapists contribute? – What mechanisms ensure inclusive evaluation?
Agency	–Balance between automation and human oversight – Ensuring autonomy and privacy	– Is the technology overly intrusive? – How does it influence learning and interaction? – What are its implications for educators and clinicians?
AI Quality and Robustness	– Data quality and bias mitigation – Algorithmic fairness and transparency – Protection of user privacy	–How is data preprocessed and validated? – What measures ensure algorithmic fairness? – How is children's data privacy safeguarded?
Experimental Validation	– Technology acceptance assessment – Rigorous testing and risk analysis	– What social acceptance frameworks are considered? – What barriers may hinder stakeholder adoption? – Has a comprehensive risk mitigation plan been implemented?

6. **Experimental Validation** – Assesses functionality, acceptance, and risks, ensuring alignment with user values and identifying barriers and enablers.

We introduce Table 1, as a conceptual and practical resource, supporting AI development that emphasizes ethical integrity, user needs, and contextual relevance in educational and clinical environments.

4.2 Practical Applications

The ARTIS project is currently in Phase Two of the CORE Framework, focusing on qualitative analysis through two focus groups and rounds of semi-structured interviews with two key practitioner groups: AI engineers/developers and clinical professionals. While the former manages software design and implementation, the latter evaluates its rehabilitative functionality. Insights from these sessions highlighted the need for a more holistic, ethically grounded development strategy, emphasizing alignment among the project's technical, clinical, and educational pillars. Initial studies involved the "AI Group" ($n = 6$), comprising full-stack developers and engineers who assessed data quality, pictogram and vocabulary appropriateness, human oversight, and human-computer interaction. Subsequently, the "Clinicians Group" ($n = 6$) from the Research and Care Center in Rome (CRC) evaluated the technology's rehabilitative relevance, inclusivity, user expectations, and ethical implications in clinical contexts.

One of the most extensively analyzed aspects in the discussion with clinicians was the modalities of human oversight and interaction, both in the direct engagement between clinicians and the software and within the broader clinical-rehabilitative setting. The primary objective was to understand the needs and challenges associated with integrating technology into rehabilitation contexts and to evaluate the theoretical alignment between the AI Group and the Clinician Group. Within the framework of the Capability Approach, achieving an optimal balance of human oversight is critical. Specifically, it is essential to mediate between (a) the degree of automation, which ensures that the technology remains unobtrusive and facilitates seamless interaction, and (b) the necessity for clinicians to retain control and the ability to intervene when using the interface. While excessive automation would entail risks associated with the loss of adult agency in the use of technology, an excessive need for intervention could potentially disrupt the fluidity of the setting, leading to a reversal of the technology's intended function. The goal is for the technology to serve as a scaffolding tool, subtly supporting the interaction between the child and the educator or rehabilitator rather than becoming the dominant element in the learning process. If human oversight is not appropriately balanced, there is a risk that the interaction will shift from a child-centered learning dynamic to one where managing the technology takes precedence. Another key finding was the need to align theoretical and practical models between groups, particularly in user experience, interface design, and interaction mechanisms. In response to the focus group, a second software prototype was developed to better integrate technical and therapeutic requirements.

5 Conclusion and Future Work

This study underscores the importance of a multidimensional, ethical, and inclusive approach to integrating AI in education, emphasizing its potential to foster cognitive and relational development in alignment with educational goals such as subjectification, qualification, and socialization. Adopting a systemic-ecological perspective, it considers the dynamic interplay between individuals, technology, and environments—particularly for vulnerable populations. Using an AI-powered reading comprehension interface as a case study, the research contributes by: (I) establishing a capability-based theoretical framework for AI development, (II) implementing a stakeholder-driven co-creation strategy, and (III) demonstrating it through a case study. Future work will develop the framework into a systematic methodology for evaluating ethical issues in child-centered AI. This methodology will offer a systematic approach to assessing ethical considerations across various AI applications.

References

1. Latour, B.: Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press (2005)
2. Khan, R.U., Cheng, J.L.A., Bee, O.Y.: machine learning and dyslexia: diagnostic and classification system (DCS) for kids with learning disabilities. Int. J. Eng. Technol. **7**(3.18), 97–100 (2018)
3. Chiriatti, M., Ganapini, M., Panai, E., Ubiali, M., Riva, G.: The case for human-AI interaction as system 0 thinking. Nat. Hum. Behav. **8**(10), 1829–1830 (2024)
4. Tuomi, I.: Artificial intelligence, 21st century competences, and socio-emotional learning in education: more than high-risk? Eur. J. Educ. **57**(4), 601–619 (2022)
5. Cesaroni, V., Galletti, M., Pasqua, E., Nardi, D., et al.: Towards trustworthy AI in inclusive education: a co-creation approach rooted in ecological frameworks. In: Ital-IA 2024: 4th National Conference on Artificial Intelligence, Organized by CINI (2024)
6. Tuomi, I.: A framework for socio-developmental ethics in educational AI. In: Hawaii International Conference on System Sciences 2023 (HICSS-56) (2023)
7. Galletti, M., et al.: a reading comprehension interface for students with learning disorders. In: Companion Publication of the 25th International Conference on Multimodal Interaction, pp. 282–287 (2023)
8. Galletti, M., Pasqua, E., Calanca, M., Marchesi, C., Tomaiuoli, D., Nardi, D., et al.: ARTIS: a digital interface to promote the rehabiliatation of text comprehension difficulties through artificial intelligence. In: Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, 29–30 May 2024, Naples, Italy (2024)
9. Holmes, W., Tuomi, I.: State of the art and practice in AI in education. Eur. J. Educ. **57**(4), 542–570 (2022)
10. Pagliara, S.M., et al.: The integration of artificial intelligence in inclusive education: a scoping review. Information **15**(12), 774 (2024)
11. Barua, P.D., et al.: Artificial intelligence enabled personalised assistive tools to enhance education of children with neurodevelopmental disorders–a review. Int. J. Environ. Res. Public Health **19**(3), 1192 (2022)
12. Mouta, A., Pinto-Llorente, A.M., Torrecilla-Sánchez, E.M.: Uncovering blind spots in education ethics: insights from a systematic literature review on artificial intelligence in education. Int. J. Artif. Intell. Educ. **34**(3), 1166–1205 (2024)
13. Holmes, W., Porayska-Pomsta, K.: The Ethics of AI in Education. Practices, Challenges, and Debates. Routledge New York (2023)
14. Cesaroni, V.: Inclusive education in the age of artificial intelligence: a critical perspective on policy guidelines through the lens of ecological-systemic theory of technological mediation. Educ. Sci. Soc-Open Access **15**(2) (2024)
15. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education-where are the educators? Int. J. Educ. Technol. High. Educ. **16**(1), 1–27 (2019)
16. Seon, J., Holmes, W.: When AI in education takes off without educators. In: How to Use Digital Learning with Confidence and Creativity, pp. 269–277. Edward Elgar Publishing (2024)
17. Cao, M., Martin, E., Li, X.: Machine learning in attention-deficit/hyperactivity disorder: new approaches toward understanding the neural mechanisms. Transl. Psychiatry **13**(1), 236 (2023)
18. Ahire, N., Awale, R.N., Patnaik, S., Wagh, A.: A comprehensive review of machine learning approaches for dyslexia diagnosis. Multimedia Tools Appl. **82**(9), 13557–13577 (2023)

19. Merzon, L., et al.: Eye movement behavior in a real-world virtual reality task reveals ADHD in children. Sci. Rep. **12**(1), 20308 (2022)
20. Vadalà, C.: L'uso dell'intelligenza artificiale in ambito sanitario: riflessioni a partire da una sperimentazione per lo sviluppo di un SAMD per la diagnosi di autismo infantile. BioLaw Journal-Rivista di BioDiritto **1S**, 375–388 (2024)
21. Reiss, M.J.: The use of AI in education: practicalities and ethical considerations. London Rev. Educ. **19**(1) (2021)
22. Bhardwaj, A., Sharma, M., Kumar, S., Sharma, S., Sharma, P.C.: Transforming pediatric speech and language disorder diagnosis and therapy: the evolving role of artificial intelligence. Health Sci. Rev. 100188 (2024)
23. Kintsch, W.: Toward a model of text comprehension and production. Psychol. Rev. (1978)
24. Van Dijk, T.A., Kintsch, W.: Strategies of discourse comprehension. Language **62** (1983)
25. Nussbaum, M.C.: Frontiers of Justice: Disability, Nationality. Harvard University Press, Species Membership (2007)
26. Sen, A.K.: What is development about. Front. Dev. Econ. Future Perspect. **1**, 506–513 (2000)
27. Floridi, L.: The Fourth Revolution: How the Infosphere is Reshaping Human Reality. OUP Oxford (2014)
28. Coeckelbergh, M.: "How I learned to love the robot": Capabilities, information technologies, and elderly care. In: The Capability Approach, Technology and Design, pp. 77–86. Springer (2012)
29. Friedman, B., Hendry, D.G.: Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press (2019)

Teacher Perceptions on AI Support for Bi/multilingual Learners: Superpowers and Constraints

Alina Guha(✉), Shamya Karumbaiah, Cynthia Baeza, Mariana Castro, Kiyo White, and Diego Roman

University of Wisconsin, Madison, WI 53706, USA
{ajguha,shamya.karumbaiah,ctbaeza,mcastro,diego.roman}@wisc.edu, kmwhit26@colby.edu

Abstract. Bi/multilingual learners in U.S. classrooms navigate the dual challenge of learning academic content while acquiring a new language. Artificial intelligence-based tools (e.g., those powered by Multilingual Large Language Models (MLLMs)) can facilitate new forms of languaging in these classrooms. We ran a survey on teachers of bi/multilingual learners (n = 232) across the U.S. and analyzed their perceptions of AI using a mixed-methods approach. We found that teachers desire multilingual abilities and feel positively towards AI to promote engagement, personalized support, and equity. They worry about the dangers of overreliance, mistakes/misinformation, and the loss of language nuance when using AI for bi/multilingual learners. Educators are divided on the potential social isolation caused by AI and confidence in supporting their students in a foreign language. Contextual factors (e.g., language background, grade level) also affect teachers' views on AI. These findings can inform the future development of AI used to support bi/multilingual students. The fears, desires and contextual factors can indicate what changes and designs may be most impactful in different classrooms.

Keywords: Multilingualism · Teachers · AI · Perceptions · Survey · Needs Finding

1 Introduction

Bi/multilingual students are underserved in many U.S. classrooms: pervasive.

monolingual ideology frames their language capabilities as a deficit [3], they might not have teachers that speak their home languages, and the assessment of their learning may be inhibited by their language skills [7]. Allowing students access to their home language in classrooms can improve their self-confidence and academic outcomes [3, 12] moving towards a more equitable education for bi/multilingual learners.

Advances in AI, most specifically Multilingual Large Language Models (MLLMs; machine learning models trained on large amounts of data that have the ability to interpret, translate, and generate text), provide new opportunities to transform the way we approach

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 85–91, 2025.
https://doi.org/10.1007/978-3-031-98462-4_11

teaching and learning for bi/multilingual students. They can help with translation so students have access to their home language, thus promoting equity and engagement [1, 5]. These tools can help monolingual teachers or teachers that do not speak all their students' languages. They can also provide immediate, personalized feedback and assist with teachers' workloads [1, 13].

The efficacy and frequency of technology implementation depends on teacher perceptions [2]. There are many potential reasons for teachers to be apprehensive about introducing AI in their classrooms including an interference with students' content or language learning, data security and privacy, and social isolation or a reduction of human interaction [1, 6, 10, 13]. Teachers' classroom experience can also inform the future development of AI and in what areas it might be most impactful. Thus, it is vital to understand teachers' perspectives on using AI to support bi/multilingual learners. Through our paper we aim to answer the following research questions:

1. What are some of the areas of support for teaching bi/multilingual students identified by educators that AI can help with?
2. Why might teachers hesitate to adopt AI as a tool for supporting bi/multilingual learners?
3. How do contextual factors affect educators' views on AI for bi/multilingual students.

2 Methods

2.1 Survey Development and Distribution

Survey Development. We designed a survey for teachers of bi/multilingual learners across the U.S. on AI and multilingualism. This paper focuses on the sections directly related to AI. We will be analyzing responses from two open-ended questions:

1. *If you had any superpower to help you support your bi/multilingual students, what would it be? List as many as you like.* (adapted from Nazaretsky et al. [8] to identify potential areas of support without restricting teachers' creativity based on their perceptions of AI)
2. *When you hear "AI support for bi/multilingual students," what concerns, fears, or constraints come to mind?*

We will also be analyzing Likert scale items in the "Early Input on AI" section. The Likert scale questions are inspired by literature on teacher perceptions of AI and language technologies [1, 4, 6, 13] (for example items, refer to footnote on Table 1). The scale has 4 options from 'Strongly Agree' to 'Strongly Disagree'.

Survey Distribution. We intentionally sampled using Prolific to include bilingual Latine teachers and both bilingual and monolingual teachers. Given that Latine and Spanish-speaking students make up the majority of students classified as English Learners in the U.S. [9, 11], we searched for evidence for whether shared ethnic backgrounds might influence perception. The project was IRB approved and we received 232 responses in total from Prolific and through UW Madison's mailing list of educators.

We received responses from teachers with varying years of experience (<1:28;1–2:50,3–5:48;6–10:42;11 +:61), teaching different grade levels (Preschool:21; Primary:47; Intermediate:56; Middle school:58; High school:65; Post-secondary:57), and across different racial and ethnic categories (Asian/ Pacific Islander:38, Hispanic/Latine:74, White/Caucasian:102, Black or African American:26, American Indian or Alaskan Native:4). Respondents could choose more than one option for grade levels and race/ethnicity.

2.2 Data Analysis

Qualitative Coding. The responses from the two open-ended questions were coded and grouped into appropriate themes. The short answers were labeled to identify the respondents' fears or superpowers and these labels were grouped based on overarching themes or patterns. The labels and themes were reviewed and any redundancies or missed patterns were re-coded.

Linear Regression for Contextual Factors. We conducted Ordinary Least Squares (OLS) regression analysis for each of the Likert scale items on AI, converting responses into binary variables. Responses of 'Strongly Agree' and 'Agree' were coded as 1, while other responses were coded as 0. Each item was considered against the demographic variables (e.g., years of experience, grade level) to identify any significant correlations. Our goal was to explore if these factors might provide further insight into any differences in opinion between teachers on these items.

3 Results

3.1 RQ1: What Are Some of the Areas of Support for Teaching Bi/Multilingual Students Identified by Educators that AI Can Help with?

Superpowers. Five of the most frequently coded superpowers are (1) a desire to have multilingual abilities, (2) mind reading, (3) emotional support for their students, (4) multicultural abilities, and (5) extra time to provide personalized support.

Teachers reported wanting to "become fluent in any language". Some reasons reported are to make "connections between all students' L1s and English" or to "explain certain issues like grammar in a way that makes more sense". Many teachers mentioned mind reading and telepathy "to pin point the exact struggle/ hurdle that they are facing" or "to know how to best explain things to them". The next most common category was providing emotional/psychological support such as "removing the level of stress they have about acquisition," or "giv[ing] them all the confidence they need to feel safe and take risks as they learn," Some teachers mentioned wanting to "understand culturally where students are coming from" and how this might help them "Explain things in culturally relevant ways to them." They also mentioned wanting to "clone [themselves]" so that they could have extra time to provide more personalized support.

Besides this some of the desired superpowers include changing others' negative perceptions of bi/multilingual learners, creating visualizations and resources, effecting

administrative change, empathy, patience, enhanced memory and supporting students' family and home life.

AI Support. 75%-84% of teachers responded either 'Strongly Agree' or 'Agree' to many of the positive possibilities of AI use for bi/multilingual students. They report an interest in AI supporting bi/multilingual learners using in-the-moment language support, promoting equity by allowing learners access to their home language, increasing student engagement, saving time by providing personalized feedback, and acting as a conversational partner.

3.2 RQ2: Why Might Teachers Hesitate to Adopt AI as a Tool for Supporting Bi/Multilingual Learners?

Fears. The five most frequently coded fears are (1) overreliance, (2) mistakes, (3) plagiarism/cheating, (4) social isolation, and (5) language nuance.

For overreliance, teachers' most reported fear is that students might be taking "shortcuts that will undermine the foundational building blocks of learning and education" and they "don't do as much cognitive work when they rely on translations of long passages of texts all the time." The next highly reported fear is that of "AI inaccuracies and lack of oversight," or "misinformation that the computer picked up from sources I don't know or trust." This becomes particularly challenging in the case of a teacher not knowing the language it is being used in - "that the information provided to the student in their home language is erroneous." Another frequently reported worry is that of plagiarism or cheating. Teachers ask "Can [students] use it during assessments." They are also worried that "I will not easily be able to tell what their level of English competency is." Social isolation - "risk of diminishing human interaction and personalized support" - is another reported worry. The fifth is that AI cannot understand nuances in language. "AI struggles to detect context-reliant language conventions" leading to potential "misunderstandings or misinterpretations, especially with complex academic content."

Some of the other reported fears include a lack of oversight or ability to monitor students when they use AI, inequity in AI for under-resourced languages, teachers overusing AI, a desire for training in AI use, knowing how to guide students in how to use AI, and issues of safety, privacy and bias.

Likert Scale Questions. 56% teachers responded 'Slightly' or 'Not at all' and the remaining 44% responded that they were either 'Extremely' or 'Very' comfortable using AI in their classroom. 84% of participants responded 'Strongly Agree' or 'Agree' when asked if they would like to receive more training in AI before evaluating its safety. This item had the highest number of respondents choosing 'Strongly Agree' (99) compared to other questions.

According to teachers' reported responses on their concerns about AI, they seem to be in agreement about several of the issues; with a range of 74–84% agreement on privacy and data security, AI bias, overreliance, and mistakes as concerns for AI incorporation. Many teachers also agree that AI cannot understand language nuances to support bi/multilingual students (65% agreement). Fewer teachers report disallowing the use of AI on the basis of cheating/plagiarizing (35% agreement).

Teachers are divided on knowing how to support their bi/multilingual students using AI in a language they are not familiar with themselves. 48% responded 'Disagree' or 'Strongly Disagree' indicating that they would be comfortable. 41% of teachers reported agreeing that AI for bi/multilingual students could lead to isolation between peers and teachers, indicating another potentially divided area.

3.3 RQ3: How Do Contextual Factors Affect Educators' Views on AI for Bi/Multilingual Students?

Monolingualism as a Factor. 61% of monolingual teachers (19 out of 31) agreed (chose 'Agree' or 'Strongly Agree') that they wouldn't know how to support their students using AI in an unknown language, as compared to 50% of non-monolingual teachers (101 out of 201) who agreed. This group also cited cheating or plagiarizing as a cause for concern at a slightly lower rate as compared to other teachers ($p < 0.05$).

Grade Level as a Factor. 85% of post-secondary teachers were worried about bias, with middle school teachers as the next highest group at 80%. The rest of the grade levels were between 70–78% and had no significant relationships with this question. 71% of preschool teachers agreed on AI being risky due to mistakes and 74% of intermediate teachers did the same. Whereas the rest of the groups ranged from 81–86% agreement. 91% of high school teachers agreed with using AI as a conversational partner, while the rest ranged from 70–88%.

Table 1. Table of coefficients from OLS regression for selected questions and contextual factors. Bold coefficients are significant, * indicates $p < 0.05$, ** indicates $p < 0.01$.

	unknown language[1]	cheating[2]	bias[3]	mistakes[4]	conversational partner [5]
const	0.485	0.469	0.698	0.836	0.833
preschool	−0.007	0.081	0.068	**−0.182***	0.105
intermediate	−0.143	−0.047	-0.032	**−0.138***	−0.085
middle	0.059	0.093	**0.188****	0.005	0.080
high	−0.135	0.044	0.113	−0.062	**0.126***
post-secondary	−0.074	−0.037	**0.233****	−0.022	0.096
monolingual	**0.267***	**0.250***	0.090	−0.040	−0.145

[1] I won't know how to support my multilingual students if they use AI in a language I don't speak
[2] I do not allow my multilingual students to use AI because this would be cheating/plagiarizing.
[3] I am worried about AI biases (e.g., stereotypes about a given language) that may disproportionately impact my multilingual students.
[4] I think AI can make mistakes sometimes, making it risky for students to use it in a language I don't speak.
[5] AI can be a good conversational partner for my multilingual students who don't have peers speaking their home language

4 Discussion

Overview. We conducted a teacher survey to learn the ways that AI could help and hinder their bi/multilingual learners' language, academic and social development. This information can help in understanding the areas for support as reported by teachers with varying backgrounds and classrooms.

Teachers' fears and desires can inform the development of AI tools to support bi/multilingual learners. More than half of the survey respondents mentioned multilingual abilities or translation as a desired superpower. They mentioned several possible ways to incorporate this in the classroom (e.g., subtitles, connections between languages). Some of these tools could simultaneously address some of the other superpowers mentioned (i.e., instilling confidence, personalization). Teachers also reported significant concerns that need to be addressed for AI use to be effective in the classroom. Mistakes and language nuance need to be improved to increase educators' trust in these technologies. They also mentioned a lack of oversight when such issues do occur, thus encouraging new mechanisms for teachers to engage with AI in a way that gives them more information and control in the classroom. The most reported fear relating to AI is overreliance. It would be useful to explore different ways to address this concern, both at the level of designing and developing the technology as well as finding thoughtful ways to use it effectively in the classroom.

Individual Differences must be Taken into Consideration. Other insights from the survey (i.e., contextual factors) can have a significant effect on a teachers' perception of particular use cases or dangers. For instance, the correlations between grade level and some of the survey items can provide an indication of what features may be appropriate for certain grade levels and what teachers of these student groups would be most looking forward to. An AI conversational partner may be more accepted for high school students as compared to those in primary school. We can also explore monolingual teachers' use of language technologies. While this group is the one that would benefit the most from a tool like AI, they seem to be more hesitant about its use. This could be due to a lack of trust in both students and the AI to report what the students are truly discussing. Other contextual factors should also be considered in future studies such as school policies and teachers' roles and subjects.

AI training can help support teachers of bi/multilingual learners. According to some of the responses to open-ended questions and Likert scale questions, teachers would like to have a deeper understanding of what AI is and how it works, how to utilize AI in the classroom, and how to guide their students in using AI more effectively.

Limitations and areas for further exploration. Given that data was collected through a survey, there are limitations as to how much detail can be extrapolated from teacher responses. The ideas shared in the survey can be developed further through storyboarding and individual interviews would also be necessary to understand the impact of language knowledge or grade level on using AI. The data from the survey can provide several potential directions for research into AI for this group of students.

Conclusion. Teachers' classroom experience and direct interactions with bi/multilingual learners can be invaluable in the goal of creating more equitable classrooms using AI. Their insights can not only inform what issues and tools may be most

necessary to resolve and design but also will be a deciding factor in the effective and impactful implementation of these technologies. This survey can act as a first step in a collaboration between researchers and educators of bi/multilingual students to develop equitable and ethical AI solutions for this student group.

Acknowledgments. Support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation.

References

1. Alharbi, W.: AI in the foreign language classroom: a pedagogical overview of automated writing assistance tools. Educ. Res. Inter. 4253331. (2023). https://doi.org/10.1155/2023/4253331
2. Farjon, D., Smits, A., Voogt, J.: Technology integration of pre-service teachers explained by attitudes and beliefs, competency, access, and experience. Comput. Educ. **130**, 81–93 (2019). https://doi.org/10.1016/j.compedu.2018.11.010
3. García, O., Wei, L.: Translanguaging. Palgrave Macmillan UK (2014). https://doi.org/10.1057/9781137385765
4. Kim, N. J., Kim, M.K.: Teacher's perceptions of using an artificial intelligence-based educational tool for scientific writing. Frontiers **7** (2022). https://doi.org/10.3389/feduc.2022.755914
5. Kelly, R., Hou, H.: Empowering learners of English as an additional language: translanguaging with machine translation. Lang. Educ. **36**(6), 1–16 (2021). https://doi.org/10.1080/09500782.2021.1958834
6. Kizilcec, R.F.: To advance AI use in education, focus on understanding educators. Int. J. Artif. Intell. Educ. **34**(1), 12–19 (2023). https://doi.org/10.1007/s40593-023-00351-4
7. Marsh, V.L.: Best practices for educating English language learners: History, controversy, and a path forward. Research Brief: Center for Urban Education Success (2018). https://www.rochester.edu/warner/cues/wp-content/uploads/2023/12/ELLS-brief_FINAL-with-header_proofed-TD.pdf
8. Nazaretsky, T., Bar, C., Walter, M., Alexandron, G.: Empowering teachers with AI: co-designing a learning analytics tool for personalized instruction in the science classroom. 1–12 (2022). https://doi.org/10.1145/3506860.3506861
9. Statista. Hispanic population groups in the United States in 2022, by origin of country, 5 July (2024). https://www.statista.com/statistics/234852/us-hispanic-population/
10. Taghizadeh, M., Hasani Yourdshahi, Z.: Integrating technology into young learners' classes: Language teachers' perceptions. Comput. Assist. Lang. Learn. **33**(8), 982–1006 (2020). https://doi.org/10.1080/09588221.2019.1618876
11. United States Census. Hispanic or Latino population quick facts (2022). https://www.census.gov/quickfacts/fact/table/US/RHI725222
12. Vogel, S., García, O. Translanguaging. In Noblit, G., Moll, L. (eds.), Oxford Research Encyclopedia of education. Oxford University Press (2017). https://doi.org/10.1093/acrefore/9780190264093.013.181
13. Zhang, K., Ayse, B.A.: AI Technologies for education: recent research & future directions. Comput. Educ. Artifi. Intell. (2021). https://doi.org/10.1016/j.caeai.2021.100025



Towards a Multimodal Document-Grounded Conversational AI System for Education

Karan Taneja[1(✉)], Anjali Singh[2], and Ashok K. Goel[1]

[1] Georgia Institute of Technology, Atlanta, GA 30332, USA
{karan.taneja,ashok.goel}@cc.gatech.edu
[2] University of Texas at Austin, Austin, TX 78712, USA
singhanj@umich.edu

Abstract. Conversational AI systems in education have predominantly focused on studying text-based interactions but education, especially in STEM fields, necessitates the use of visual depictions such as optical ray diagrams, molecular geometry, and physiological structures. Further, multimedia learning using text and visuals has been shown to improve learning outcomes compared to text-only instruction. In this paper, we examine MuDoC, a Multimodal Document-grounded Conversational AI, as a tool for multimedia learning. MuDoC leverages text and visuals from documents to generate multimodal blog-like responses. We compare it to a text-only system to explore the differences in learner engagement, trust, and learning outcomes. Our findings show that both visuals and verifiability of content enhance engagement and foster trust; however, no significant impact in performance was observed. We interpret the findings in light of theories from cognitive and learning sciences to derive implications for multimodal conversational AI systems in education.

Keywords: Multimedia Learning · Multimodal AI · Dialog Systems

1 Introduction

Existing conversational AI systems such as Jill Watson [7,15] and PET [18] ground their responses in instructor-approved course materials and predefined template-based slides respectively to answer student questions. These course documents typically contain visuals to support understanding and engagement, especially in STEM fields, but they have not been utilized to generate multimodal responses. Further, multimedia learning involving the processing of verbal and visual elements is considered more effective for transfer and retention compared to learning from one modality [9]. With recent AI advancements, multimodal dialog systems [3,6,8,12,19] have received increasing attention, but their application to multimedia learning remains under-explored. In this paper, we evaluate MuDoC, a Multimodal Document-grounded Convers-ational AI system [14], for multimedia learning. MuDoC generates responses interleaved with text and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 92–99, 2025.
https://doi.org/10.1007/978-3-031-98462-4_12

images in a format similar to blog articles[1]. Based on GPT-4o, it uses past chat as context to understand user queries and responds based on visuals and text retrieved from documents. MuDoC's user interface allows seamless navigation to image and text sources to enable verification of AI-generated content to foster trust in its responses. We compare MuDoC to a text-only system in a user study with students to understand the differences in learner engagement, trust on AI, and their subsequent performance on problem-solving tasks. Prior studies have isolated the learning and the evaluation process in two steps [9] but we consider a more authentic scenario where students solve problems while learning by asking questions to an AI system. This setting requires students to self-regulate their learning [17] as they form questions, read AI responses, and solve problems in a reasonable time frame. Student feedback shows that both visuals and verifiability of AI content support learner engagement and trustworthiness in multimodal conversations, though results show no significant difference in performance. Our analysis provides insights into the cognitive processing involved in this authentic scenario, based on which we provide directions for future research.

2 Related Work

Multimedia Learning: Multimedia Learning Theory (MLT) [9] suggests that learning and retention improves with verbal and visual information as it is processed through separate, limited-capacity channels and integrated together along with prior knowledge to form meaningful mental representations. Cognitive Load Theory (CLT) [13] posits that while short-term working memory has limited capacity, effective learning depends on leveraging the unlimited long-term memory to support the acquisition and storage of new knowledge. In this work, we investigate whether multimedia effect holds true with MuDoC [14] in a realistic setting where cognitive load is high as students learn while solving problems.

Grounded Conversational AI: LLM-based Jill Watson [7,15] is a conversational AI for classrooms which uses a retrieval-augmented generation (RAG) pipeline to answer student questions by retrieving relevant information and prompting the LLM to generate text-based responses. Similarly, PET [18] is a dialog system to use predefined template-based slides for answering student questions using GPT-4. Intelligent textbook proposed by Olson et al. [10] also display a chat window adjacent to documents similar to PET but does not use visuals or allow navigation to source content. MuDoC [14] uses a RAG pipeline and extract figures from documents during the preprocessing step to utilize them for generating interleaved text and image responses.

3 MuDoCVersus TexDoCStudy

MuDoC, a Multimodal Document-grounded Conversational AI system [14], processes user queries as a part of conversation to retrieve relevant text and

[1] Conversation samples (Google Doc): tinyurl.com/MuDoCSamples
Demo+Documentation (OSF): tinyurl.com/MuDoC-OSF.

images from documents to generate a multimodal response. Its responses contain interleaved images and text in a format similar to blog articles (see supplementary material(See footnote 1)). Further, it allows users to seamlessly navigate to sources of images and texts in its responses by mapping them to the source document. To understand the impact of multimodality and interactive features of MuDoC, we designed a within-subjects research study where we compare MuDoC against a baseline system called **TexDoC**. TexDoC is the Text-only variant of MuDoC with the same backbone, except it does not have an image retrieval pipeline or image-related instructions in prompts. Further, it only displays a ChatGPT-like chat window and the PDF document is provided in a separate browser tab.

We recruited n = 30 graduate (Masters and PhD) students from Georgia Tech (Atlanta, GA, USA) to participate in the study. We screened prospective participants to select those who have finished at least one course in ML or AI domain, but have never enrolled for the Knowledge-based AI (KBAI) course on which the problems for our study were based. We invited participants for a 90-minutes in-person session where they solved two subjective analytical problems using pen and paper while asking questions to TexDoC and MuDoC. The KBAI textbook [5] provided to the participants contains numerous visuals explaining concepts and algorithms taught in the course. Each participant solved a *warm-up task* in 7 min to familiarize themselves with the two systems before respective randomly-assigned *graded transfer tasks* in 20 min on *incremental concept learning* or *analogical reasoning*. The participants also answered survey questions comparing TexDoC and MuDoC and provided feedback in a brief semi-structured post-study interview (more details in supplementary material(Seefootnote 1)).

4 Results

Survey and Interview Feedback: Survey responses to statements comparing TexDoC and MuDoC on a 7-point Likert scale are summarized in Fig. 1. Overall, 22 participants preferred MuDoC while only 4 preferred TexDoC. Out of 30, 28 preferred MuDoC *for exploring the content for problem-solving*. Many explained that navigation using text and images allowed them to check responses for hallucinations, refer to the textbook frequently, get more explanations from the book, and to create more informed follow-up questions. Participants also indicated that MuDoC features helped them engage with the content more actively. *For interest and engagement*, 26 participants preferred MuDoC while only 3 preferred TexDoC. Many characterized images from MuDoC as helpful in learning and capturing attention. Several participants also mentioned that MuDoC responses were more detailed and contained helpful examples. Some participants preferred reading the book and found navigation to be helpful and engaging. Those who favored TexDoC found themselves more focused with its minimalism and similarity to ChatGPT-like interface. 23 participants perceived MuDoC responses to *be more memorable*, while 4 favored TexDoC. Many believed that visuals in MuDoC responses made the content more memorable compared to text-only responses. Several participants appreciated the interactive nature of the

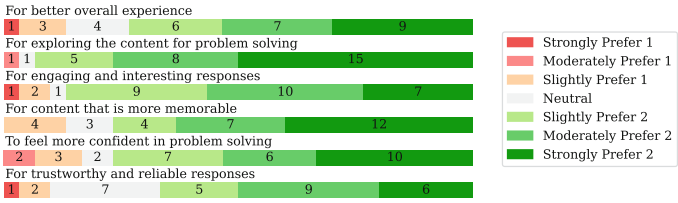


Fig. 1. Participants preferences comparing (1) text-only TexDoC and (2) multimodal MuDoC. Numbers indicate participant count.

interface, noting that it encouraged deeper engagement with the content. Participants in favor of TexDoC preferred conciseness which was important due to time constraints. 23 participants *felt more confident about problem-solving* with MuDoC due to the presence of visuals, ability to navigate to relevant content, preference for reading from the book, response credibility, and ease of understanding due to the ELI10 feature. However, participants who felt more confident with TexDoC expressed concerns about information overload when using MuDoC. Finally, 20 out of 30 participants favored MuDoC *for trustworthiness and reliability of responses* while only 3 preferred TexDoC. The ability to navigate to the source of images and texts using MuDoC was cited as the top reason for its higher credibility. However, some participants raised concerns about MuDoC's errors in navigation using text, and long responses with multiple images.

User Experience Feedback: Based on the survey data, all participants found MuDoC *responses with images* to be useful. Most believed that visuals were more effective in explaining concepts than text alone. Participants placed higher trust on images as they could not be hallucinated and were clipped from the document. 29 participants preferred the convenience of *side-by-side chat & PDF display* because it enabled verification and increased trust. 29 and 26 participants identified *navigation using images and text* respectively as useful due to the ease in referring to the textbook, finding relevant content, and improving credibility. However, some expressed their frustration with navigating using text when the text was mapped incorrectly. 22 and 21 participants found ELI10 and Summarize features to be useful respectively for being engaging and assistive for reading, while others could not find enough time to utilize them.

Student Performance: Two teaching assistants (TA) created an initial rubric and independently evaluated $\approx 25\%$ of all solutions. Following this, the initial rubric was adapted to create a final rubric. After independently evaluating all the solutions again and scoring them on a scale of 0–12, the disagreements were discussed to decide final scores. A linear mixed-effects model (LMM) was used with scores as dependent variable, system (TexDoC or MuDoC) and task as fixed effects, and participants as groups. We observed a small but insignificant negative impact of MuDoC on student scores ($\hat{\beta} = -0.350$, SE = 0.542, p = 0.518). During the tasks, participants asked M = 5.17, SD = 2.40 questions to

MuDoC and M = 5.86, SD = 2.68 to TexDoC (with paired t-test, t = 1.481, p = 0.149 >0.05), and spent an average of 19 min with both. The average response length excluding images for both systems was about 1750 characters (SD = 634). For MuDoC, 58% responses contained 1–5 images (M = 2.02, SD = 1.18).

5 Discussion

Multimedia Effect: Visuals in MuDoC responses were unanimously considered useful by participants for enhancing learning experience and capturing attention. The *multimedia effect* in MLT suggests that integration of verbal and visual channels during learning can lead to higher cognitive retention [9]. This was also concurred by participants as they perceived multimodal responses to be more memorable. Also, many visuals exemplified concepts/algorithms which positively influenced perception of MuDoC responses. This aligns with the *worked example effect* in CLT which suggests that worked examples are more effective for learning compared to unguided problem solving [13]. While higher cognitive retention and worked example effect have been established before, more work is required to realize them with multimodal AI under varying conditions such as learner goals, time constraints, passive or active role, and levels of personalization.

Reliability and Credibility: MuDoC responses were considered as more credible because participants could examine the source of responses on a granular level. The use of document-grounded images also increased trust because MuDoC images cannot be hallucinated. In information seeking, reliability (consistent high quality responses) and credibility (believability of source) are considered the most important predictors of trust [16]. This explains lower trust in TexDoC as it does not provide credible evidence for its responses. However, as indicated by some participants, TexDoC could be perceived as more reliable for its consistent response length, which was not inflated by images. Further, the presence of sources can improve trust, but quality of source attribution is also important factor [2]. Consequently, reliability of MuDoC can be improved by providing concise responses that focus on student queries over content coverage, and through reliable source attribution. Future work should explore methods to obtain granular feedback, for instance, by asking learners to highlight useful content in responses, to communicate their priorities and understanding. Such information can be capitalized to generate succinct personalized responses.

Learning Outcomes: While text in MuDoC responses had the same average length as TexDoC, additional information consisted of ≈ 2 images and source mappings to the document. The *split-attention effect* in CLT [13] suggests that integrating information from multiple sources increases cognitive load which can explain the 'information overload' experienced by some participants. Further, CLT suggests that long-term memory, such as familiarity with an AI system, can reduce load on working memory and improve learning [13]. Some participants indeed acknowledged a potential negative impact of unfamiliarity with MuDoC. Overall effect of these factors can lead to overload on working memory capacity which may explain why MuDoC did not impact scores despite its many positives.

Critical Thinking: The ease of auditing AI responses empowered participants to check for hallucinations and to delve deeper by examining related content. Therefore, the navigation features in MuDoC helped in active reflection and evaluation of content which suggests that MuDoC may promote critical thinking with AI [11]. As a skill, critical thinking requires deliberate practice [4] and future work could examine whether AI systems that enable such forms of critical evaluation can positively impact learner behavior in the long term.

Cognitive Processing: Interestingly, some participants noted that verifying the AI responses can be helpful for longer retention. This could be explained by Craik and Lockhart's *Levels of Processing* framework of memory [1], according to which learning activities that require deeper cognitive processing are more effective for long-term memory compared to mere repetition. Existing AI systems require excessive work to verify correctness, which discourages engagement in this process [2]. We hypothesize that ease of verification in MuDoC positively contributed to cognitive processing, and, therefore, higher perceived retention.

Load Regulation: In contrast to positive impact on retention, CLT suggests that processing multiple novel elements can lead to poor learning if cognitive load exceeds working memory capacity [13]. This could explain why some participants felt overwhelmed by MuDoC's responses. Our future work will focus on striking a balance between cognitive engagement and cognitive load while allowing learners to self-regulate cognitive load through focus mode for chat and document.

Summary: MuDoC provides engaging responses with interleaved text and visuals that have the potential to improve learning outcomes such as cognitive retention, in agreement with MLT. Trust in AI systems for information-seeking tasks depends on its ability to provide consistent high quality responses and to support verification. For MuDoC, document-grounded visuals and navigation using text and images lead to increased credibility. However, more work is needed to generate concise personalized AI responses and to improve granular source attribution. The active process of verification, when convenient to perform, leads to higher engagement and may positively influence critical thinking and cognitive retention. To further leverage such interactive systems for student learning, there is a need to design systems that allow self-regulation of cognitive load.

6 Conclusion

We conducted a user study to compare a multimodal document-grounded system with its text-only counterpart in a realistic setting where participants solved problems while self-learning by asking questions to an AI system. Results show that multimodality lead to higher engagement and the interface promoted trustworthiness as it allowed verification content through seamless navigation. While we did not observe a significant change in performance on problem solving tasks, the qualitative feedback suggests that (1) illustrative visuals and cognitive engagement through verification can improve cognitive retention and critical thinking skills, and (2) document-grounded visuals and verifiability increase

credibility and trust in AI systems. We discussed our findings in light of theories of learning and cognition to understand implications for multimodal AI for education.

Acknowledgments. We are grateful for the support provided by National Science Foundation under Grant No. 2247790 and Grant No. 2112532. We also thank John Kos, Audrey Olson, Sandeep Kakar and Lingqing Wang for their valuable contributions.

References

1. Craik, F.I.M., Lockhart, R.S.: Levels of processing: a framework for memory research. J. Verbal Learn. Verbal Behav. **11**(6), 671–684 (1972). https://www.sciencedirect.com/science/article/pii/S002253717280001X
2. Ding, Y., et al.: Citations and trust in LLM generated responses, January 2025. http://arxiv.org/abs/2501.01303, arXiv:2501.01303 [cs]
3. Feng, J., Sun, Q., Xu, C., et al.: MMDialog: a large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In: Annual Meeting of the Association for Computational Linguistics, pp. 7348–7363, July 2023. https://aclanthology.org/2023.acl-long.405/
4. Gelder, T.V.: Teaching critical thinking: some lessons from cognitive science. College Teach. **53**(1), 41–48 (2005). https://www.tandfonline.com/doi/abs/10.3200/ctch.53.1.41-48
5. Goel, A., Joyner, D.: Knowledge-based Artificial Intelligence: Cognitive Systems (e-book). Georgia Institute of Technology (2018). https://cs7637.netlify.app/Ebook/kbai_ebook-compressed.pdf
6. Huang, M., et al.: DialogGen: multi-modal interactive dialogue system for multi-turn text-to-image generation, July 2024. http://arxiv.org/abs/2403.08857, arXiv:2403.08857 [cs]
7. Kakar, S., et al.: Jill watson: scaling and deploying an AI conversational agent in online classrooms. In: International Conference on Intelligent Tutoring Systems, pp. 78–90 (2024). https://link.springer.com/chapter/10.1007/978-3-031-63028-6_7
8. Kong, F., Wang, P., Feng, S., Wang, D., Zhang, Y.: TIGER: a unified generative model framework for multimodal dialogue response generation. In: Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), pp. 16135–16141, May 2024. https://aclanthology.org/2024.lrec-main.1403/
9. Mayer, R.E.: Multimedia learning. In: Psychology of Learning and Motivation, vol. 41, pp. 85–139. Academic Press (2002). https://www.sciencedirect.com/science/article/pii/S0079742102800056
10. Olson, A., Maiti, P., Goel, A.: The textbook of tomorrow: rethinking course material interfacing in the era of GPT, January 2025. http://arxiv.org/abs/2501.03618, arXiv:2501.03618 [cs]
11. Singh, A., Taneja, K., Guan, Z., Ghosh, A.: Protecting human cognition in the age of AI. In: Tools for Thought Workshop at CHI 2025, April 2025. http://arxiv.org/abs/2502.12447
12. Sun, Q., et al.: Multimodal dialogue response generation. In: 60th Annual Meeting of the Association for Computational Linguistics, pp. 2854–2866, May 2022. https://aclanthology.org/2022.acl-long.204/

13. Sweller, J.: Cognitive load theory. In: Psychology of Learning and Motivation, vol. 55, pp. 37–76. Academic Press (2011). https://www.sciencedirect.com/science/article/pii/B9780123876911000028
14. Taneja, K., Goel, A.K.: MuDoC: an interactive multimodal document-grounded conversational AI system. In: AAAI Symposium Series 2025 Machine Learning and Knowledge Engineering (MAKE) for Trustworthy Multimodal and Generative AI (2025). http://arxiv.org/abs/2502.09843
15. Taneja, K., Maiti, P., Kakar, S., Guruprasad, P., Rao, S., Goel, A.K.: Jill watson: a virtual teaching assistant powered by ChatGPT. In: International Conference on Artificial Intelligence in Education, pp. 324–337 (2024). https://link.springer.com/chapter/10.1007/978-3-031-64302-6_23
16. Thielsch, M.T., Meeßen, S.M., Hertel, G.: Trust and distrust in information systems at the workplace. PeerJ **6**, 5483 (2018). https://peerj.com/articles/5483/
17. Winne, P.H., Hadwin, A.F.: Studying as self-regulated learning. In: Metacognition in Educational Theory and Practice, pp. 277–304 (1998). https://psycnet.apa.org/record/1998-07283-011
18. Wölfel, M., Shirzad, M.B., Reich, A., Anderer, K.: Knowledge-based and generative-AI-driven pedagogical conversational agents: a comparative study of grice's cooperative principles and trust. Big Data Cogn. Comput. **8**(1), 2 (2024). https://www.mdpi.com/2504-2289/8/1/2
19. Zhu, W., et al.: Multimodal C4: an open, billion-scale corpus of images interleaved with text. In: Annual Conference on Neural Information Processing Systems, pp. 8958–8974 (2024). https://arxiv.org/abs/2304.06939



Generative AI in K-12 Education Policy: A Preliminary Global Review

Camila Hidalgo(✉) and Noha Lea Halim

Technical University of Munich, Richard-Wagner-Street 1, 80333 München, Germany
{camila.hidalgo,nohalea.halim}@tum.de

Abstract. As students and teachers adopt Generative Artificial Intelligence (GAI), educational stakeholders are reassessing existing frameworks, revising policies, and designing new guidelines to harness GAI's potential while ensuring equitable and effective implementation. While higher education has received considerable attention, guidelines for K–12 education remain underdeveloped, necessitating further exploration. We conducted a scoping review of global policy documents on GAI use in K–12 to identify key policies and analyze how they address substantive concerns and strategies. Employing thematic analysis, we found new policy priorities for implementing GAI in the classroom, including co-design in research and development, procurement and infrastructure, well-being and mental health, and human alternatives.

Keywords: Generative Artificial Intelligence · K-12 Education · Children

1 Introduction

The free access to Generative AI (GAI) tools by the public has significantly impacted education [1], due to GAI's unique ability to autonomously generate text, images, and data, which brings new layers of complexity and ethical considerations. This scenario raises the question of "how existing [artificial intelligence (AI)] policies must be [re]interpreted in novel contexts, and new guidance and regulations may need to be developed" [2]. While there has been considerable focus on GAI guidance for higher education, primary and secondary education (K-12 education) remains in the early stages [3, 4]. A United Nations Educational Scientific and Cultural Organization (UNESCO) survey found that fewer than 10% of schools have introduced guidance for GAI, with 40% of these schools reporting that these guidelines had only been communicated verbally [2]. To contribute to reevaluating existing policies in response to GAI, we take stock of the current landscape of GAI in education policy for K-12 to map out emerging trends in policy shaping this field for policymakers, teachers, and learners.

2 Methodology

We conducted a scoping review of international policy documents focused on children, published after November 2022, when ChatGPT was released to the public. By focusing on policy reports from important stakeholders in education with a strong influence on

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 100–107, 2025.
https://doi.org/10.1007/978-3-031-98462-4_13

setting educational policy agendas, these preliminary findings are important for ongoing discussions surrounding AI governance in education. This study excluded academic and legal sources, focusing instead on grey literature that offers policy guidance on GAI in education. Grey literature refers to materials produced by governments, academics, business and industry–both in print and digital formats–that fall outside the realm of commercial publishing [5, 6]. Often written by policy-shaping organizations and not peer-reviewed, this literature plays a crucial role in real-time governance debates. It provides a snapshot of current practice and emerging interventions, here guidelines for GAI in K-12 education, unencumbered by the publication delays typical of academic research [5]. In fast-changing fields like GAI, where institutional actions often outpace scholarly consensus [6], grey literature has become especially valuable. Such literature captures policy while it is being framed, communicated, and implemented, offering timely, practice-based insights. These insights are essential for understanding the evolving regulatory and educational terrain in which GAI is enacted in K-12 settings.

We carried out a keyword-based search of the Google.com search engine using the following search terms and key words [7, 8]: ("Generative Artificial Intelligence" OR "Generative AI" OR "GenAI" OR "GAI") AND ("K-12" OR "primary education" OR "secondary education" OR "children") AND ("genAI + report + education" OR "genAI report education" OR "GenAI Report Education"). We reviewed the resulting 276 hits per by title and brief description.

Three documents met the inclusion criteria: (1) global scope, (2) sources that addressed GAI tools (3) relevance to primary and secondary education (K-12) education and/or child-specific policy focus, and (4) documents such as reports by intergovernmental organizations, white papers, guidelines and toolkits intended to inform decision-making. Documents were excluded if they (1) focused on regional or national contexts rather than a global scope; (2) addressed AI without discussing GAI tools; (3) did not focus on K-12 education or on children as the target population; (4) consisted of opinion pieces, personal statements, academic publications, or advocacy briefs not offering substantive policy guidance.

A keyword-based search yielded three reports based on our search criteria:

1. *UNESCO's Guidance for Generative AI in Education and Research* [9].
2. *World Economic Forum's Shaping the Future of Learning: The Role of AI in Education 4.0* [10].
3. *Organisation for Economic Co-operation and Development's Opportunities, Guidelines, and Guardrails for Effective AI in Education* [11].

After a close reading of the reports, we identified the challenges, core dilemmas, opportunities, and recommendations discussed in these reports. After our first round of analyses, we conducted a deductive thematic analysis [12] of these reports based on Chiu's categorization [4] for the impact of GAI in school education based on Chiu's et al. taxonomy [13], where the researcher identified four main themes and twelve subthemes. The four overarching themes are Student Learning, Teacher Learning, Assessment, and Administration, each of which includes several subthemes. Under Student Learning, the subthemes are: (1.1) AI education in schools, (1.2) critical reasoning and thinking, (1.3) digital, media, and information literacies, and (1.4) generic skills development. The theme of Teacher Learning includes: (2.1) teacher curriculum leadership, (2.2) teacher

AI literacy, (2.3) teacher facilitating skills, and (2.4) interdisciplinary teaching. Within Assessment, the focus is on (3.1) assessment approaches and (3.2) dimensions of assessment. Finally, the Administration theme covers (4.1) effective work with evidence and (4.2) the attitude of school administrative teams. This structure offers a comprehensive lens through which to evaluate the multifaceted impact of GenAI across educational settings. [4] Chiu's categorization was instrumental in guiding our thematic exploration and ensuring that our analysis was robust and aligned with current scholarly discussions. Based on our examination, we included the following themes: co-design in AI research and development, procurement and infrastructure, wellbeing and mental health, and human alternatives.

Limitations. This review identified some valuable trends and suggests potential research directions in GAI in K-12 education for researchers, teachers and policymakers. However, it has some limitations. First, we focus on policy documents with an international scope to map out trends of various AI products developed and deployed across various jurisdictions, which complicates the creation and alignment of guidelines in a fragmented policy landscape. Two, we excluded grey literature from regions (e.g. European Union) that may have provided interesting and complementary insights to focus on global trends to identify potential pathways for alignment. Third, alternative search strings may have yielded different results, for instance, including "EdGPT" into our keyword search ("Generative Artificial Intelligence" OR "Generative AI" OR "GenAI" OR "GAI".) Nonetheless, we believe this keyword "EdGPT" is sufficiently addressed through the AND clause ("K-12" OR "primary education" OR "secondary education" OR "children").

3 Findings and Discussion

The main challenges described in the policy reports refer to student learning, teaching, assessment and administration. The findings show that these commonly discussed topics in AIED – such as ethical and safe use of AI-powered tools, privacy and data protection, biases, AI and digital literacy, and skills – are still important for GAI tools in education. Although at the time of writing, our analysis is ongoing, we can provide provisional results regarding new themes highlighted in the three reports prompted by the GAI impact on K-12 education building on Chiu's categorization. These new themes include co-design in AI research and development, procurement and infrastructure, human alternatives, and wellbeing and mental health.

Co-design in AI Research and Development. Governments should promote co-designing in the research and development of AI tools to be used in education, including teachers, students, and other end-users [11]. This process involves working together to create new AI-powered tools, adapting existing ones for educational purposes, and collaborating on practice-engaged research to develop evidence-based use cases with a strong focus on how these technologies will benefit teachers and learners and under what conditions. Multistakeholder consortium programs should be established and then encouraged to explore the social context in which AI-powered education technologies can best be used and facilitate their uptake in the educational community [11]. This approach will be helpful to navigate GAI outputs, which often undermine pluralistic

and minority opinions resulting in a homogenization of knowledge and limiting creative thinking, since the current knowledge databases and reasoning engines employed reproduce dominant worldviews in their outputs [9]. The WEF report provides examples of these kinds of consortiums. To ensure equitable and appropriate integration of AI into education, policymakers and developers should prioritize participatory co-design processes that give teachers and students an active role in shaping GAI tools from the outset.

Infrastructure and Procurement. Investment in digital infrastructure in and outside the school is key for equitable AI access to help bridge the digital gap [10, 11]. Nevertheless, there are existing dilemmas that policymakers, teachers and developers must grapple with. Such dilemmas encompass developing accessible tools [10] that can be accessed through mobile internet, lending devices to the families that need them, designing AI-powered applications that can run on mobile phones and laptops and be accessed online and offline (e.g., low-tech solutions.) Moreover, when designing GAI tools for education, developers rely on computing power and huge amounts of data, which is generally accessible in leading economies and is often sourced from the majority world, respectively. These tools are then exported to countries in the global majority, which often require greater capacity to develop, deploy, as well as monitor them [9]. Proprietary AI models limit transparency and external review of biases, which in turn may widen the divide in training and controlling GAI models [9].

Human Alternatives. The three reports focused on the role of a human centered approach to governing AI in education, which often implies the augmentation or automatization of certain tasks to optimize teacher roles [10] and students' learning to empower them in the educational processes. The OECD report predicts that AI-diagnoses and suggestions will become the default option as AI adoption increasingly takes on decision processes in the educational context [11]. Nevertheless, this report emphasizes that stakeholders must have access to human alternatives to AI tools and timely human support, especially for high-stakes decisions [11], for instance, to explain a decision-making process for accountability purposes. Even though opting out of AI-based tools should be possible, it raises concerns about equitable learning opportunities and when opting out is even possible or desirable [10]. For example, a user may opt out of employing particular tools selected by the educational institution, which may have the potential to undermine children's learning and fair assessment (Table 1).

Well-being and Mental Health. Teachers and school staff play a critical role in students' well-being. GAI systems employed in educational institutions may pose harmful risks to students' and teachers' well-being due to their excessive use, potential of mimicking human interaction with unknown long-term consequences for end users [9, 10], or the amplifying effect of distributing unethical content [11]. The UNESCO report emphasizes the GAI systems may mimic human interaction, which could have psychological effects on learners' and teachers' cognitive and emotional development and well-being, for instance, the risk of users being manipulated [9]. In contrast, GAI has the potential to carry out conversation-based diagnoses to identify psychological and socio-emotional problems as well as learning difficulties addressing any shortages of teachers and mental

health professionals in schools. Therefore, UNESCO recommends institutional policies that ensures AI use is proportional and does not cause predictable harm to users [9].

The challenges facing K-12 education triggered by GAI revolve around concerns related to copyright, age limitations, and the right to be forgotten. While GAI providers are obligated to safeguard the copyright and portrait rights of AI users, they must also

Table 1. Emerging Policy Trends for GAI in K-12 Education

GAI Guidelines	UNESCO [9]	WEF [10]	OECD [11]
Policy Trends			
Co-Design in Research and Development	**Inclusive GAI Design**. Ensuring equitable representation in GAI model development **Users Fine-Tune EdGPT Pedagogy.** Allowing users to adjust the extent to which EdGPT models go beyond subject knowledge to also focus on student-centered pedagogy, positive teacher-student interactions, and review and adjust GAI-outcomes	**Participatory GAI Development.** Promoting multi-stakeholder engagement in AI tool creation, including teachers, students, researchers, and developers	**Collaborative GAI for Education**. Developing policies that advocate for evidence-based and co-created GAI applications in education
Infrastructure and Procurement	**Equitable Digital Learning**. Ensuring fair access to educational resources – online, offline, low-tech solutions – and internet connectivity	**Equitable GAI Access**. Bridging the digital divide through investment in AI-compatible education infrastructure	**Accessible Digital Infrastructure**. Undertaking coordinated action to ensure accessible and equal distribution of digital infrastructure inside and outside school **Guidelines for Procurement.** Providing clear guidelines on AI procurement in educational settings to ensure transparency and compliance

(*continued*)

Table 1. (*continued*)

GAI Guidelines	UNESCO [9]	WEF [10]	OECD [11]
Human Alternatives	**Human Oversight in GAI**. Ensuring critical decision-making remains with teachers	**Human-Guided AI Learning.** Maintaining human oversight in GAI-assisted learning environments	**Ethical GAI Delegation** Defining boundaries between AI assistance and human decision-making in schools **GAI by Necessity**. Evaluating the need for GAI implementation to uphold the right to human alternatives and to opt-out
Wellbeing and Mental Health	**Psychological Impact of GAI**. Investigating long-term effects of GAI on students' and teachers' mental health	**Ethical GAI for Well-being** Researching the safety, efficacy, and equity implications of GAI-enabled tools in education should be prioritized, for example, the impact these tools may have on children's cognition and development	**Safeguarding GAI Use.** Developing policies to prevent AI-driven stressors in learning **Balancing Learning & Wellbeing**. Navigating conflicting priorities in AI-Driven Education. AIED should prioritize learners' and teachers' mental health and wellbeing and uphold a balance between digital and non-digital activities

recognize that the input these digital tools receive may be integrated into GAI training data, as stated by UNESCO [9]. A key dilemma is balancing the benefits of these tools with the risk of losing control over work and personal data, making it difficult to determine ownership and the originality of the work generated. Moreover, the UNESCO report points out the lack of mechanisms to extract individual data from a GPT model after training, which poses a threat to the right to be forgotten [9]. Children, as primary users of education systems, have unique needs often unmet by adult-focused GAI. Given that most GAI applications are tailored for adults, child learners are vulnerable to risks such as exposure to inappropriate content, manipulation, and uncertainties regarding the long-term impact of GAI in their development. UNESCO stresses the need for age restrictions to protect their rights and well-being [9]. Despite this, questions remain about children's assent formats, the appropriate age threshold, the use of self-reported age for verification, and the responsibility of GAI providers and parents or guardians to oversee children's GAI usage [9].

4 Conclusion

This preliminary global review explores the developing landscape of GAI policy in K-12 education. We conclude that while students and teachers increasingly use GAI, policies and guidelines for this emergent technology in primary and secondary schools are less developed compared to higher education. This review takes stock of the current policy landscape in response to GAI's impact by identifying policy trends and exploring their potential relevance to existing AI policy and the development of future guidance. This research points out emerging priorities to address GAI in this context through a review of key international policy documents. Beyond existing AI topics, commonly discussed with regard to education, such as ethical use, privacy, data protection, bias, and literacy, our analysis highlights the importance of addressing co-design in AI development, consideration of infrastructure and procurement, maintaining human alternatives, and the impact of GAI on teachers' and students' wellbeing and mental health.

In our review, we also identify the challenges related to copyright, age restrictions, and the right to be forgotten in the context of GAI use by children. The next step is to develop a matrix to help policymakers and practitioners navigate the evolving landscape of GAI governance in K–12 education, integrating policy insights with educational research.

Acknowledgments. The authors thank the TUM IEAI for supporting this work and Prof. Urs Gasser for providing constructive feedback and guidance on this research project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gottfried, O.S., Park, E., Jeffrey, G.: About a quarter of U.S. teens have used ChatGPT for schoolwork – double the share in (2023). https://www.pewresearch.org/short-reads/2025/01/15/about-a-quarter-of-us-teens-have-used-chatgpt-for-schoolwork-double-the-share-in-2023/, Accessed 5 Feb 2025
2. UNESCO: UNESCO survey: Less than 10% of schools and universities have formal guidance on AI. https://www.unesco.org/en/articles/unesco-survey-less-10-schools-and-universities-have-formal-guidance-ai, Accessed 5 Feb 2025
3. Steele, J.L.: To GPT or not GPT? Empowering our students to learn with AI. Comput. Educ. Artif. Intell. **5**, 100160 (2023). https://doi.org/10.1016/j.caeai.2023.100160
4. Chiu, T.K.F.: The impact of Generative AI (GenAI) on practices, policies and research direction in education: a case of ChatGPT and Midjourney. Interact. Learn. Environ., 1–17 (2023). https://doi.org/10.1080/10494820.2023.2253861
5. Paez, A.: Gray literature: An important resource in systematic reviews. J. Evid. Based Med. **10**, 233–240 (2017). https://doi.org/10.1111/jebm.12266
6. Benzies, K.M., Premji, S., Hayden, K.A., Serrett, K.: State-of-the-evidence reviews: advantages and challenges of including grey literature. Worldviews Evid. Based Nurs. **3**, 55–61 (2006). https://doi.org/10.1111/j.1741-6787.2006.00051.x
7. Piasecki, J., Waligora, M., Dranseika, V.: Google Search as an additional source in systematic reviews. Sci. Eng. Ethics **24**, 809–810 (2018). https://doi.org/10.1007/s11948-017-0010-4

8. Miake-Lye, I.M., Mak, S., Begashaw, M.M., Shekelle, P.G.: Using Google to search for evidence: how much is enough? One center's experience. Syst. Rev. **14**(92) (2025). https://doi.org/10.1186/s13643-025-02836-w
9. UNESCO: Guidance for generative AI in education and research. UNESCO, Paris, France (2023). https://doi.org/10.54675/EWZM9535, Accessed 5 Feb 2025
10. World Economic Forum: Shaping the future of learning: The role of AI in education 4.0. World Economic Forum. https://www3.weforum.org/docs/WEF_Shaping_the_Future_of_Learning_2024.pdf, Accessed 5 Feb 2025
11. OECD-Education International: Opportunities, guidelines and guardrails on effective and equitable use of AI in education. https://www.oecd.org/content/dam/oecd/en/about/projects/edu/smart-data-and-digital-technology-in-education/Opportunities,%20guidelines%20and%20guardrails%20for%20effective%20and%20equitable%20use%20of%20AI%20in%20education.pdf/_jcr_content/renditions/original./Opportunities,%20guidelines%20and%20guardrails%20for%20effective%20and%20equitable%20use%20of%20AI%20in%20education.pdf, Accessed 5 Feb 2025
12. Roberts, K., Dowell, A., Nie, J.-B.: Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. BMC Med. Res. Methodol. **19**(66) (2019). https://doi.org/10.1186/s12874-019-0707-y
13. Chiu, T.K.F., Xia, Q., Zhou, X., Chai, C.S., Cheng, M.: Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. Comput. Educ. Artif. Intell. **4**, 100118 (2023). https://doi.org/10.1016/j.caeai.2022.100118

Where's the F***ing Filter? A Survey on Handling Problematic and Sensitive Questions in Open Educational Chatbots

Thilo I. Dieing(✉) and Lea Cohausz

University of Mannheim, 68159 Mannheim, Germany
thdieing@mail.uni-mannheim.de, lea.cohausz@uni-mannheim.de

Abstract. Due to the emergence of powerful Large Language Models (LLMs), many open-domain generative educational chatbots have evolved in the past few years, with many more to be expected as best practices and reliable methods emerge. In this paper, we consider a particular best practice that has not received much attention yet: the handling of student questions that are sensitive, politically inappropriate, or contain vulgar language. Although handling such questions wrong can have negative consequences and decrease the trust of users, parents, and teachers, our survey of papers introducing open-domain chatbots reveals that only a small percentage of papers consider them explicitly. In our survey (N = 55 papers), we make a case for why this problem should receive more attention, how it is currently being handled, and which steps future research should take. Moreover, our survey is, to the best of our knowledge, the most complete one considering only open-domain chatbots.

Keywords: Chatbot · open-domain · misuse · LLM · ethical usage

1 Introduction

Recent advancements in Large Language Models (LLMs) have led to the development of various educational chatbots that assist students in their academic lives [5,14]. While studies demonstrate their effectiveness for specific tasks [8,20], one challenge in educational settings is the users—typically children, teenagers, or young adults. This group is more likely to:

- Experiment with vulgar language or inappropriate queries to test the chatbot's responses [11].
- Ask socially or politically sensitive questions or make controversial statements.
- Raise sensitive questions that may not be within the chatbot's domain, but are genuine concerns (e.g. "What can I do when my parents get mad when I don't pass the test?").

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 108–115, 2025.
https://doi.org/10.1007/978-3-031-98462-4_14

These *problematic out-of-domain questions* differ from the chatbot's intended domain and pose a challenge. Traditional rule-based, closed-domain chatbots handled such queries by providing no response [16]. However, LLM-based, open-domain chatbots generate responses to all queries, including out-of-domain ones. For problematic questions, we may want the chatbot to either redirect the user to a more suitable topic or respond empathetically to sensitive issues.

If not managed correctly, these situations can erode trust in the system, particularly among users or their parents. Despite this, many studies on educational chatbots overlook these concerns, and no standard procedure for handling them exists. Some developers place undue trust in LLMs, like ChatGPT, assuming they can handle such queries without additional safeguards, stating they work"quickly and safely." [3, p.55]. However, research shows that even ChatGPT can struggle with these types of questions, leading to safety concerns [13].

Given the importance of the domain and user group, we argue that educational chatbots should address out-of-domain queries with predefined responses and possibly alert responsible adults. To better understand how the questions are currently being addressed, we surveyed existing educational chatbot research and asked: (1) How many papers discuss problematic out-of-domain questions? (2) What procedures exist to handle them? Our goal is to highlight the challenges of problematic-out-of-domain questions and suggest future research directions and practical solutions.

2 Background: Conversational Agents in Educations

Conversational agents, or chatbots, have become increasingly common in education, supporting staff, teachers, and especially students [5,14]. They are typically categorized as teaching-oriented (e.g., tutoring or grading) or service-oriented (e.g., curriculum support) [4,27]. Chatbots also differ by knowledge scope and response generation. They can be open-domain or closed-domain, and use either generative models like LLMs for more dynamic interactions or rule-based systems [2,25]. Currently, most educational chatbots (88.88%, N in the study = 36) are rule-based and closed-domain [16], though the rise of LLMs is expected to shift this trend.

2.1 Related Surveys

Several literature reviews have examined research on educational chatbots. Since this paper focuses on developments influenced by recent LLM advances, we highlight newer surveys. Okonkwo and Ade-Ibijola [27] found that most educational chatbot research emerged after 2016, with 66% focused on teaching, 9% on administration, and the rest on development and assessment. Labadze et al. [17] emphasized chatbots' potential for feedback, explanations, and flexible learning. Gökçearslan et al. [14] showed that most chatbots target college students, with under 15% for younger learners. They also noted that 59.5% of

studies used a variety of chatbots, 8.1% used OpenAI LLMs, and the rest did not specify.

Most of these reviews raised ethical concerns, including chatbot personas, privacy, plagiarism, and originality [14,17,27]. However, little attention is given to user-driven risks—such as misuse through offensive language or politically sensitive queries. Inappropriate responses to such inputs can undermine trust and limit adoption. This paper aims to explore how recent literature addresses this often overlooked concern.

2.2 Search Strategy

As a starting point, we began by selecting papers from three existing surveys that met our inclusion criteria (detailed below), yielding only three results. This low number highlights that even recent surveys mostly cover pre-LLM, closed-domain chatbots. Our review thus fills a gap by focusing on LLM-based educational chatbots.

Next, we searched proceedings from four major AI-in-education conferences[1] and four key journals[2], covering publications from January 1, 2022[3] onward—reflecting the rise of LLM-based systems. Then, we conducted a targeted Google Scholar search using one term from each of the two term sets for chatbots (*chatbot, conversational agent, LLM*) and for the educational context (*education, learning*). We restricted the search to articles published since January 1, 2022, within the first 100 pages (1,000 entries). The applied inclusion criteria state that papers must introduce an open- or semi-open domain chatbot in education, with the chatbot as a main focus, and must be written in English. Our search was conducted in December 2024 and January 2025.

3 Results

Our search yielded 55 papers. Figure 1 shows which search yielded how many papers. The list of all identified papers, along with some information concerning the chatbot method, handling of problematic questions, target group, and problem to be solved (among others), can be found online here. Overall, the vast majority of papers were published in 2024 (60.00%) and 2023 (36.36%) showing that our cut-off date was well-set. 68.52% of the papers focused on college students. Most of the reviewed chatbots were tasked to help students with general Q&A (36.36%) while others were developed for a specific subject such as medicine [10], business [15] or MINT [30]. Scholars mostly relied on the LLMs by OpenAI, while some used open-source models such as LLAMA [22], Mistral [12], commercial LLMs [9], or others [6].

[1] AIED, LAK, EDM, and L@S.
[2] IJAIED, JEDM, BJET, and IJETHE.
[3] ChatGPT 3.5 was released in late 2022.

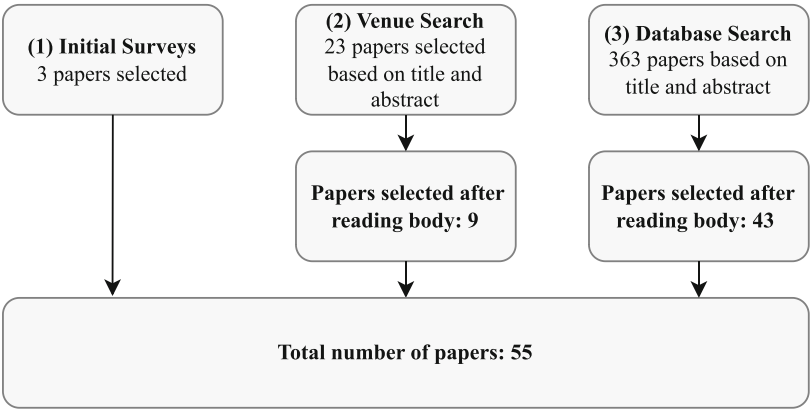


Fig. 1. The process of our search procedure, including the number of resulting papers.

3.1 Findings on Safeguarding

Our analysis shows that only 25.45% of papers (N = 14) acknowledge or address the need to handle problematic user questions–a topic we explore further in the next section. More commonly, concerns focus on issues like data privacy [17,27] and misuse for cheating [24]. In the next section, we examine proposed methods for handling such out-of-domain queries.

4 How to Handle Problematic Out-of-Domain Questions

Our analysis of papers addressing safeguards against inappropriate user queries reveals a range of methods, with no clear gold standard. Approaches generally fall into two categories: those that apply safeguards directly, and those that first detect problematic input before responding. Additional strategies include cleaning training or Retrieval Augmented Generation (RAG) data to reduce bias [28] and stereotyping [19], though these aim to prevent problematic outputs rather than address problematic questions directly.

4.1 Safeguarding Without Detection

One common method for handling problematic queries is prompting the LLM using one-shot or few-shot techniques. Scholars instruct models to be helpful [7,22], to respond appropriately [26], or to avoid out-of-scope questions [23]. While prompting can shape chatbot behavior, it does not prevent the model from generating inappropriate answers. For instance, a chatbot prompted to be helpful might still respond to queries like "Where can I study chicks?" inappropriately. Prompts can be valuable for sensitive questions—e.g., offering empathy when a student feels overwhelmed—but may lack the nuance needed for full safety across

all questions. Overall, prompting alone is a broad, inflexible method that leaves gaps in handling problematic queries, highlighting the need for more targeted safeguards.

4.2 Safeguarding with Detection

To address the limitations of prompting, some scholars use intent classifiers to detect problematic out-of-domain questions before responding. Sajja et al. [29] integrated sentiment detection into a curriculum-assistance chatbot, enabling it to respond more empathetically to negative or sensitive queries. Unlike only using prompting, this method allows the chatbot to dynamically adjust its behavior based on user input. Other researchers also classify user intent but differ in response strategies. Li et al. [18] tag queries as safe or unsafe, passing these to a self-trained model that generates contextually appropriate responses. Their chatbot, trained on Reddit data, produced save replies and revealed that men are more likely to ask unsafe questions.

Another approach to handling out-of-domain questions after they have been detected is the topic avoidance strategy [31], where unsafe queries trigger predefined responses (e.g., "I can't help with that") instead of engaging the LLM. This strategy has been adopted in various educational chatbots [1,9,12,21]. While effective, intent classification requires training on diverse, hard-to-source user data. Still, it remains the most targeted way to limit critical queries, whereas prompting excels at handling emotional ones.

5 Implications and Recommendations

Our survey reveals limited but growing awareness of the need to address problematic out-of-domain questions, though this topic is currently still rarely discussed in application papers. We now turn to the implications of these findings for researchers and practitioners, along with recommendations for future work.

5.1 For Practitioners

We recommend that chatbot developers proactively consider the likelihood of problematic out-of-domain questions in their context and implement countermeasures where needed. A two-step approach—first detecting such queries, then responding with predefined answer patterns—is generally more secure than relying on freely generated responses. However, appropriate replies can be highly context-dependent, and defining what makes a "good" response should be critically discussed. As open-domain educational chatbots are still emerging, more refined and well-evaluated solutions are likely to develop.

We also emphasize the need for explicit testing of chatbot responses to problematic queries. Standard user testing may not surface these issues, as users often ask polite questions in such settings. Instead, purposefully designed test

cases including responses should be reviewed with domain experts (e.g., teachers) to assess their fit. Finally, developers should openly communicate these challenges to teachers and parents. While this may risk initially undermining trust, transparency about the chatbot's safeguards—especially for sensitive topics—can foster greater trust eventually. Letting adults know that certain questions trigger referral messages or predefined responses may increase confidence in the chatbot's responsible use.

5.2 For Researchers

The issue of problematic out-of-domain questions is still emerging, and much research remains to be done, as shown by the varied approaches in the existing literature. Key next steps for research include:

- **Benchmark dataset creation**: We recommend developing a benchmark dataset to help chatbot developers evaluate how well their systems detect and respond to problematic queries. This should include both problematic and regular questions to test differentiation. A benchmark could also vary by target group or topic, enriching the discussion of what defines problematic out-of-domain questions.
- **Testing on benchmark data**: Existing and new chatbots should be tested on this dataset to assess their reliability in detecting and answering problematic questions. Evaluations should focus on both detection accuracy and response quality.
- **Defining good answers**: Future research should explore what constitutes an optimal response to problematic questions, considering factors like question type and target group. Input from experts in related fields (e.g., teachers) would enrich this discussion.
- **Safeguarding methods**: More research should focus on identifying which safeguarding methods work best for handling problematic queries.

6 Conclusion

Motivated by a recent increase in open-domain generative chatbots in education, this literature survey investigated how scholars currently handle users' problematic out-of-domain questions. Building educational chatbots is challenging due to the risk of inappropriate or sensitive questions, often asked by the target group (children or young adults). Allowing such questions to be answered freely can harm trust and usage. This study reviewed 55 papers from 2022 to 2025 to assess how scholars handle such issues. Results show that only 25.45% of papers address problematic questions, typically using two methods: zero-shot or few-shot prompting, or detecting issues via intent classifiers and applying safeguards like topic avoidance or targeted prompting. We recommend that practitioners consider safeguarding against problematic questions and communicate chatbot limitations to teachers and parents. Future research should focus on creating benchmark datasets, developing new strategies for handling such questions, and including experts from various fields.

References

1. Abu-Rasheed, H., Abdulsalam, M.H., Weber, C., Fathi, M.: Supporting student decisions on learning recommendations: an llm-based chatbot with knowledge graph contextualization for conversational explainability and mentoring. arXiv preprint arXiv:2401.08517 (2024)
2. Adamopoulou, E., Moussiades, L.: An overview of chatbot technology. In: IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 373–383. Springer (2020)
3. Alper, A., Tekin, A.: Student experiences on the use of artificial intelligence-based chatbots in programming education (2024)
4. Baral, S., Worden, E., Lim, W.C., Luo, Z., Santorelli, C., Gurung, A.: Automated assessment in math education: a comparative analysis of llms for open-ended responses. In: Proceedings of the 17th International Conference on Educational Data Mining, pp. 732–737 (2024)
5. Berida, O.T., et al.: Smat-bot: an interactive chatbot for facilitating students' comprehension of mathematics. In: 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG), pp. 1–8. IEEE (2024)
6. Bui, T., et al.: Cross-data knowledge graph construction for llm-enabled educational question-answering system: a case study at hcmut. In: Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia, pp. 36–43 (2024)
7. Buraphadeja, V., Chantokul, D.: Supporting student engagement and academic guidance through a chatbot in blended learning. In: International Journal on E-Learning, pp. 241–253. Association for the Advancement of Computing in Education (AACE) (2024)
8. Chakraborty, C., Pal, S., Bhattacharya, M., Dash, S., Lee, S.S.: Overview of chatbots with special emphasis on artificial intelligence-enabled chatgpt in medical science. Front. Artif. Intell. **6**, 1237704 (2023)
9. Chen, Y., Jensen, S., Albert, L.J., Gupta, S., Lee, T.: Artificial intelligence (AI) student assistants in the classroom: designing chatbots to support student success. Inf. Syst. Front. **25**(1), 161–182 (2023)
10. Chheang, V., et al.: Towards anatomy education with generative AI-based virtual assistants in immersive virtual reality environments. In: 2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR), pp. 21–30. IEEE (2024)
11. Cho, S.J., Tian, Y.: Why do they keep swearing? the role of outcome expectations between descriptive norms and swearing among korean youths: a test of the theory of normative social behavior. West. J. Commun. **84**(2), 227–244 (2020)
12. Dieing, T.I., Scheffler, M., Cohausz, L.: Enhancing chatbot-assisted study program orientation. In: Proceedings of DELFI Workshops 2024, pp. 10–18420. Gesellschaft für Informatik eV (2024)
13. Gallifant, J., et al.: Peer review of gpt-4 technical report and systems card. PLOS Digit. Health **3**(1), e0000417 (2024)
14. Gökçearslan, S., Tosun, C., Erdemir, Z.G.: Benefits, challenges, and methods of artificial intelligence (AI) chatbots in education: a systematic literature review. Int. J. Technol. Educ. **7**(1), 19–39 (2024)
15. Ilagan, J.B.R., Ilagan, J.R.: A prototype of a chatbot for evaluating and refining student startup ideas using a large language model (2023)

16. Kuhail, M.A., Alturki, N., Alramlawi, S., Alhejori, K.: Interacting with educational chatbots: a systematic review. Educ. Inf. Technol. **28**(1), 973–1018 (2023)
17. Labadze, L., Grigolia, M., Machaidze, L.: Role of AI chatbots in education: systematic literature review. Int. J. Educ. Technol. High. Educ. **20**(1), 56 (2023)
18. Li, C., Xing, W., Leite, W.: Building socially responsible conversational agents using big data to support online learning: a case with algebra nation. Br. J. Edu. Technol. **53**(4), 776–803 (2022)
19. Li, C., Zhu, W., Xing, W., Guo, R.: Analyzing student attention and acceptance of conversational AI for math learning: Insights from a randomized controlled trial. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, LAK 2024, pp. 836–842. ACM, New York, NY, USA (2024). https://doi.org/10.1145/3636555.3636895
20. Liu, H., Peng, H., Song, X., Xu, C., Zhang, M.: Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. Internet Interv. **27**, 100495 (2022)
21. Maiti, P., Goel, A.K.: How do students interact with an llm-powered virtual teaching assistant in different educational settings? arXiv preprint arXiv:2407.17429 (2024)
22. Mitra, C., Miroyan, M., Jain, R., Kumud, V., Ranade, G., Norouzi, N.: Retllm-e: retrieval-prompt strategy for question-answering on student discussion forums. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 23215–23223 (2024)
23. Modran, H.A., Ursuțiu, D., Samoilă, C., Gherman-Dolhăscu, E.C.: Developing a gpt chatbot model for students programming education. In: International Conference on Science and Technology Education, pp. 72–82. Springer (2024)
24. Mohamed, A.M.: Exploring the potential of an AI-based chatbot (chatgpt) in enhancing English as a foreign language (efl) teaching: perceptions of efl faculty members. Educ. Inf. Technol. **29**(3), 3195–3217 (2024)
25. Nimavat, K., Champaneria, T.: Chatbots: an overview types, architecture, tools and future possibilities. Int. J. Sci. Res. Dev **5**(7), 1019–1024 (2017)
26. Odede, J., Frommholz, I.: Jaybot–aiding university students and admission with an llm-based chatbot. In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval, pp. 391–395 (2024)
27. Okonkwo, C.W., Ade-Ibijola, A.: Chatbots applications in education: a systematic review. Comput. Educ. Artif. Intell. **2**, 100033 (2021)
28. Saha, B., Saha, U.: Enhancing international graduate student experience through AI-driven support systems: a llm and rag-based approach. In: 2024 International Conference on Data Science and Its Applications (ICoDSA), pp. 300–304. IEEE (2024)
29. Sajja, R., Sermet, Y., Cwiertny, D., Demir, I.: Platform-independent and curriculum-oriented intelligent assistant for higher education. Int. J. Educ. Technol. High. Educ. **20**(1), 42 (2023)
30. Sun, D., Boudouaia, A., Zhu, C., Li, Y.: Would chatgpt-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? an empirical study. Int. J. Educ. Technol. High. Educ. **21**(1), 14 (2024)
31. Xu, J., Ju, D., Li, M., Boureau, Y.L., Weston, J., Dinan, E.: Recipes for safety in open-domain chatbots (2021). https://arxiv.org/abs/2010.07079



Advancing AI Literacy in Medical Education: A Medical AI Competency Framework Development

Chang Cai[1], Jamie Duell[1], Daisy Minghui Chen[1], Weng Kin Ho[3], Bernett Teck Kwong Lee[1], Fang Li[2], Siyuan Liu[2], Olivia Ng[1], Vidya Sudarshan[2], Shang-Ming Zhou[4], Gaoxia Zhu[3], and Xiuyi Fan[1,2(✉)]

[1] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
xyfan@ntu.edu.sg
[2] College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore
[3] National Institute of Education, Nanyang Technological University, Singapore, Singapore
[4] Center for Health Technology, Faculty of Health, University of Plymouth, Plymouth, UK

Abstract. The integration of Artificial Intelligence (AI) into healthcare underscores the need for medical professionals to develop AI competencies. While the UNESCO AI competency framework provides general guidelines, it is broad and lacks the specificity required for medical education. This study proposed and validated an AI competency framework tailored for medical professionals, integrating UNESCO's AI framework with Miller's pyramid model. The preliminary framework was refined based on expert feedback from six hospital administrators, eight medical professionals, and ten university instructors. This study presents an AI competency framework for medical professionals, laying the groundwork for future curricula development. It offers insights and evidence to guide medical professionals' AI competency development.

Keywords: AI competency for medical professionals · AI literacy · AI education · Medical education · AI competency framework

1 Introduction

Artificial Intelligence (AI) is revolutionizing healthcare by enhancing the quality, safety, and accessibility of medical services [28]. Researchers have employed AI to detect diseases that need to be diagnosed early, such as liver, heart, skin, and alzhemier [11]. Beyond diagnostics, AI supports real-time health monitoring, informs precision medicine approaches, and facilitates personalized healthcare [10,26]. Furthermore, AI enhances healthcare management by optimizing hospital resource allocation and improving patient satisfaction [22]. However, medical professionals face challenges in integrating AI into healthcare, such as limited AI knowledge [2], low trust in AI systems [9], potential misuse [4], and difficulties

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 116–123, 2025.
https://doi.org/10.1007/978-3-031-98462-4_15

in explaining AI-driven processes [4]. This creates an urgent need for medical professionals to develop skills in AI, as ***"AI is not meant to replace medical professionals, but the ones using AI will probably replace those who don't"*** [16, p. 4]. To meet this demand, physicians need to develop both technical and non-technical AI competencies, including understanding its concepts, assessing accuracy, recognizing unintended consequences, practicing ethical and responsible AI use, and integrating AI into clinical scenarios [3,13,25].

Although numerous AI skills for medical professionals have been proposed, a dedicated AI competency framework for medical professionals remains underexplored. Global efforts, such as UNESCO's AI competency framework, provide general guidelines for AI literacy [17], but they fail to address the distinct needs of medical professionals. Unlike generic applications, applications of AI in healthcare require expertise in data governance, bias evaluation, and critical appraisal to ensure fair and responsible implementation [6,12,15,20,24].

Furthermore, emerging technologies like GenAI have further expanded AI's role in healthcare. GenAI has applications in areas including documentation automation, patient communication, and surgery support creation, which hold transformative potential [7]. However, there are challenges, such as hallucinations and source uncertainty [19,32]. Few studies address the competencies needed for responsible use of GenAI in medical practice. Therefore, an AI competency framework for medical professionals is essential for equipping them with the skills needed for AI-driven healthcare.

To address these issues, this study seeks to answer the research question: **What are the core AI competencies required for medical professionals to address the evolving demands of AI-driven healthcare?**

We propose a Medical AI Competency Framework that integrates six essential domains with Miller's four-level model. Grounded in both educational theory and healthcare-specific needs, the framework addresses the gap between general AI literacy models and the competencies required in clinical practice. It offers medical educators a developmental roadmap to systematically cultivate AI literacy among learners. This study contributes a structured and context-specific foundation for integrating AI training into medical education.

2 Theoretical Framework

UNESCO's AI competency framework provides a global reference for the core set of AI skills, which outlines four dimensions [17]:

- *Human-centred mindset*: Cultivating a human-focused view on the benefits, risks, and proportionality of AI for societal and environmental needs.
- *Ethics of AI*: Developing social and ethical abilities to navigate and apply principles regulating human behaviour across the entire life cycle of AI.
- *AI techniques and applications*: Integrating conceptual knowledge and operational skills through practical use of AI tools in real-world tasks.
- *AI system design*: Building comprehensive engineering skills for problem-solving, system architecture, and iterative optimization of AI systems.

3 Methods

3.1 Development of the AI Competency Framework

We developed our framework based on UNESCO's AI competency framework and Miller's Pyramid model. Building on UNESCO's *Human-Centred Mindset*, we introduced *Patient-Centered AI in Healthcare*, emphasizing patient-centric decision-making. Expanding on UNESCO's *Ethics of AI*, we proposed *Ethics and Transparency in Clinical AI*, highlighting ethical considerations in medical AI. From UNESCO's *AI Techniques and Applications*, we derived *Technical Proficiency in Medical AI Applications*, focusing on how clinicians utilize AI. Integrating UNESCO's *AI System Design* with an emphasis on healthcare equity [1,30], we introduced *Bias Mitigation and Health Equity in AI Design*. Additionally, given the literature's emphasis on data privacy and governance in medicine [20,24], we incorporated *Data Privacy, Security, and Compliance in Healthcare AI*. Lastly, acknowledging the growing role and challenges of GenAI in medicine [7,19,32], we proposed *Generative AI for Healthcare* to address its applications and implications in clinical practice.

Instead of using the three-level progression (understand, apply, create) from UNESCO's AI competency framework, we adopted Miller's pyramid [18] as a guiding principle. Miller's pyramid is widely used to assess medical students' competence at different levels of learning and clinical performance [27,31]. A progression pathway was established for each competency dimension, reflecting the stepwise development of competencies in medical professionals. Specifically, Miller's pyramid structures progression into four levels [18]:

- *"Know" (Knowledge)*, where students acquire basic knowledge.
- *"Know how" (Competency)*, where students know how to use the knowledge.
- *"Show how" (Performance)*, where students exhibit skills via simulations.
- *"Do" (Action)*, where students apply skills in clinical practice.

3.2 Validation and Refinement of the Framework

Framework Validation Experts. The study included diverse stakeholders from Singapore (see Table 1), employing the "triangulation of subjects" method, which gathers data from multiple participants to ensure diverse perspectives [5]. This multidisciplinary selection process ensured that the AI competency framework was rigorously evaluated from educational, clinical, and policy perspectives.

Table 1. Demographics of Framework Validation Experts.

Participant	Number	Department
Hospital administrators	6	Hospitals and government agencies
University instructors	10	A university and healthcare organization
Medical professionals	8	Hospitals

A total of 24 experts were selected, comprising six hospital administrators, eight medical professionals, and ten university instructors, each chosen based on their domain-specific qualifications:

1. **Hospital Administrators:** Selected from hospitals and government agencies, these experts held leadership positions and had experience in policy-making and management. Their macro perspectives ensured that the framework aligned with healthcare requirements.
2. **University Instructors:** Academics with at least five years of teaching experience in AI or medicine were selected. Their expertise ensured that the competency framework was grounded in sound pedagogical principles and aligned with best practices in medical and AI education.
3. **Medical Professionals:** Physicians with a minimum of five years of clinical experience and prior exposure to or knowledge of AI applications in healthcare were included. Their contributions helped refine the framework to reflect the practical AI competencies required for real-world clinical settings.

Data Collection and Analysis. Experts were asked to complete an online survey. The survey included open-ended questions on the AI competencies medical professionals should acquire and the proficiency level required for each competency (See the full questionnaire here). The survey responses were collected anonymously, ensuring participants' confidentiality. The qualitative data from the survey responses were analyzed using content analysis to identify recurring themes. Two researchers independently coded 33.33% of the survey answers to ensure inter-rater reliability, achieving a kappa coefficient of 0.81, indicating good inter-coder agreement [8]. The two researchers resolved differing results by reaching a consensus through discussion. Then, the first author independently coded the remaining data. The insights derived from this analysis informed the refinement of the Medical AI Competency Framework.

4 Results: Medical AI Competency Framework

A key result was that most experts intuitively confirmed the importance of "Application skills" ($n = 23$), "AI foundations" ($n = 19$), and "Critical reflections on AI" ($n = 16$). Thirteen experts mentioned the importance of the use of GenAI, from learning basic knowledge to integration in practice. For example, a medical professional stated *"Using ChatGPT to conceptualize brief background of certain issues."* Following suggestions from the experts for more clarity, we redefined "Application skills" and "Co-design equitable AI systems", highlighting co-create with engineers. For example, one hospital administration highlighted, *"Not to become a coder, but to get skills to be able to comprehensively lead teams and run AI projects..."* Additionally, we expanded the definition of "Real-world compliance" and emphasized the need for medical professionals to handle data appropriately under specific conditions, including sensitive and non-sensitive data. Finally, we structured the framework into six dimensions and four levels (see Fig. 1).

Aspects	Progression			
	Know	Know how	Show how	Do
Patient-Centered AI in Healthcare	**Basic understanding:** Understand the foundational impact of AI on patient rights and autonomy.	**Human advancement:** Learn how to apply AI to enhance treatment planning, automate care processes, and personalize healthcare for individual patients.	**Patient-centered decision-making:** Use AI effectively to support decision-making, ensuring patient well-being is prioritized.	**Patient-centered practice:** Integrate AI tools into clinical practice, taking responsibility for incorporating them into patient care workflows to improve care quality and outcomes.
Ethics and Transparency in Clinical AI	**Critical reflections on AI:** Comprehend the basic ethical principles guiding the use of AI in healthcare settings.	**Safe and responsible use:** Know how to use of AI in ways that ensure ethical considerations and transparency.	**Ethical decision-making:** Demonstrate ethical considerations during AI-enabled decision-making in simulated settings.	**Ethics by design:** Implement ethical standards in real-world AI applications, evaluating and improving systems for ethical compliance.
Technical Proficiency in Medical AI Applications	**AI foundations:** Acquire foundational knowledge of AI concepts and techniques.	**Healthcare AI implementation:** Learn how to apply AI methods effectively in medical contexts.	**Application skills:** Grasp the ability to perform and co-create basic AI-related programming or data analysis tasks.	**Evaluate and create AI:** Critically assess existing AI systems and develop solutions tailored to specific healthcare needs.
Data Privacy, Security, and Compliance in Healthcare AI	**Privacy awareness:** Remember regulations and the importance of safeguarding patient data.	**Regulatory adherence:** Know how to handle patient data securely, employing techniques like anonymization and governance.	**Secure data practices:** Manage patient data in simulated research projects, adhering to privacy laws.	**Real-world compliance:** Process clinical data under specific conditions, ensuring all data handling and AI applications comply with privacy and security standards in clinical and research settings.
Bias Mitigation and Health Equity in AI Design	**Foundational understanding of bias:** Identify how AI impacts health equity and its potential to either mitigate or exacerbate disparities.	**Bias detection and analysis:** Know how to identify, evaluate, and address bias in AI systems within healthcare contexts.	**Equity-oriented practices:** Select and apply AI tools that minimize bias and enhance fairness, ensuring health equity in controlled design or assessment tasks.	**Co-design equitable AI systems:** Co-create and refine AI solutions to enhance health equity, using iterative feedback to address systemic disparities.
Generative AI for Healthcare	**Generative AI basics:** Identify the basic principles, limitations, and benefits of GenAI.	**Interpretative proficiency:** Know how to interpret AI outputs, adapt data, and identify risks like hallucinations.	**Skillful development:** Use GenAI, validate outputs, and communicate effectively in clinical scenarios.	**Integration in practice:** Independently integrates, evaluates, and oversight GenAI in real clinical workflows.

Fig. 1. Medical AI Competency Framework

The dimensions are:

1. *Patient-Centered AI in Healthcare:* Using AI in healthcare to protect patient rights and prioritize their well-being.
2. *Ethics and Transparency in Clinical AI:* Applying AI responsibly and ethically in healthcare, with clear and transparent decision-making.
3. *Technical Proficiency in Medical AI Applications:* Understanding, using, and co-creating AI techniques effectively in healthcare settings.

4. *Data Privacy, Security, and Compliance in Healthcare AI:* Protecting patient data and ensuring AI systems follow privacy laws and security standards.
5. *Bias Mitigation and Health Equity in AI Design:* Identifying and reducing bias in AI systems to ensure fairness and health equity.
6. *Generative AI for Healthcare:* Understanding, using, and overseeing GenAI in clinical settings to ensure reliable integration into real workflows.

5 Discussion and Conclusion

This study presents a Medical AI Competency Framework tailored to the unique demands of healthcare. The framework comprises six domains, including patient-centered decision-making, ethical and transparent use of clinical AI, technical proficiency in medical AI applications, data governance, bias mitigation, and generative AI integration. These competencies are mapped to four levels of Miller's pyramid model: know, know how, show how, and do. Grounded in this model, the framework ensures a stepwise progression of skills, from foundational knowledge to practical application. Through expert validation across AI, clinical, educational, and policy domains, we demonstrate that the framework offers a clear developmental roadmap for developing AI literacy in medical education.

Unlike general AI literacy frameworks such as UNESCO's [17], which focus on broad AI competencies, our framework is context-specific and clinically oriented. While consistent with prior research in highlighting the importance of both technical (e.g., AI applications) and non-technical competencies (e.g., ethics) [6,12,14,21,23,29], this study integrates competencies for the responsible use of GenAI. The inclusion of GenAI competencies introduces a forward-looking perspective that remains underexplored in current research.

There are several limitations to this study. First, although the framework is well-developed, it has not been tested in actual educational or clinical settings. Future work should assess its practical effectiveness. Second, the expert panel, while interdisciplinary, was limited to Singapore-based participants, which may affect generalizability. Broader international validation will enhance its global applicability. Third, the framework currently lacks standardized assessment tools to evaluate proficiency at each level. Future work may focus on developing practical instruments to support competency-based assessment and large-scale adoption.

In conclusion, this study introduces a context-specific and educationally sound AI competency framework for medical professionals. It bridges the gap between general AI literacy models and the practical needs of clinical settings. The framework provides a structured basis for curriculum design, supporting progressive learning aligned with competency-based assessment. As AI becomes more embedded in healthcare, the framework supports the preparation of clinicians for its responsible and effective use in clinical practice. It offers a foundational step toward integrating AI training into medical education at scale.

Acknowledgments. This research is supported by the Ministry of Education, Singapore (Grant IDs: RG22/23, LKCMedicine Start up Grant).

References

1. Abràmoff, M.D., et al.: Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging Foundation, Washington, D., Maisel, W.H.: Considerations for addressing bias in artificial intelligence for health equity. NPJ Digital Med. **6**(1), 170 (2023)
2. Alanazi, A.: Clinicians' views on using artificial intelligence in healthcare: opportunities, challenges, and beyond. Cureus **15**(9) (2023)
3. Aslam, T.M., Hoyle, D.C.: Translating the machine: skills that human clinicians must develop in the era of artificial intelligence. Ophthalmol Therapy **11**(1), 69–80 (2022)
4. Buck, C., Doctor, E., Hennrich, J., Jöhnk, J., Eymann, T.: General practitioners' attitudes toward artificial intelligence-enabled systems: interview study. J. Med. Internet Res. **24**(1), e28916 (2022)
5. Bulsara, C.: Using a mixed methods approach to enhance and validate your research. Brightwater Group Res. Centre **16**, 1–82 (2015)
6. Charow, R., et al.: Artificial intelligence education programs for health care professionals: scoping review. JMIR Med. Educ. **7**(4), e31043 (2021)
7. Chen, A., Liu, L., Zhu, T.: Advancing the democratization of generative artificial intelligence in healthcare: a narrative review. J. Hospital Manag. Health Policy **8** (2024)
8. Emam, K.E.: Benchmarking Kappa: interrater agreement in software process assessments. Empir. Softw. Eng. **4**, 113–133 (1999)
9. Esmaeilzadeh, P.: Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations. Artif. Intell. Med. **151**, 102861 (2024)
10. Johnson, K.B., et al.: Precision medicine, ai, and the future of personalized health care. Clin. Transl. Sci. **14**(1), 86–93 (2021)
11. Kumar, Y., Koul, A., Singla, R., Ijaz, M.F.: Artificial Intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J. Ambient. Intell. Humaniz. Comput. **14**(7), 8459–8486 (2023)
12. Lee, J., Wu, A.S., Li, D., Kulasegaram, K.M.: Artificial Intelligence in undergraduate medical education: a scoping review. Acad. Med. **96**(11S), S62–S70 (2021)
13. Liaw, W., Kueper, J.K., Lin, S., Bazemore, A., Kakadiaris, I.: Competencies for the use of artificial intelligence in primary care. Annals Family Med. **20**(6), 559–563 (2022)
14. Ma, Y., et al.: Promoting AI competencies for medical students: A scoping review on frameworks, programs, and tools. arXiv preprint arXiv:2407.18939 (2024)
15. Mennella, C., Maniscalco, U., De Pietro, G., Esposito, M.: Ethical and regulatory challenges of ai technologies in healthcare: a narrative review. Heliyon (2024)
16. Meskó, B., Hetényi, G., Győrffy, Z.: Will Artificial Intelligence solve the human resource crisis in healthcare? BMC Health Serv. Res. **18**, 1–4 (2018)
17. Miao, F., Shiohira, K.: AI competency framework for students. UNESCO (2024)
18. Miller, G.E.: The assessment of clinical skills/competence/performance. Acad. Med. **65**(9), S63-7 (1990)
19. Moulaei, K., Yadegari, A., Baharestani, M., Farzanbakhsh, S., Sabet, B., Afrash, M.R.: Generative artificial intelligence in healthcare: a scoping review on benefits, challenges and applications. Inter. J. Med. Inform., 105474 (2024)
20. Nasir, S., Khan, R.A., Bai, S.: Ethical framework for harnessing the power of ai in healthcare and beyond. IEEE Access **12**, 31014–31035 (2024)

21. Paranjape, K., Schinkel, M., Panday, R.N., Car, J., Nanayakkara, P., et al.: Introducing artificial intelligence training in medical education. JMIR Med. Educ. **5**(2), e16048 (2019)
22. Pillai, A.S.: AI-enabled hospital management systems for modern healthcare: an analysis of system components and interdependencies. J. Adv. Analy. Healthcare Manag. **7**(1), 212–228 (2023)
23. Pupic, N., et al.: An evidence-based approach to Artificial Intelligence education for medical students: a systematic review. PLOS Digital Health **2**(11), e0000255 (2023)
24. Reddy, S., Allan, S., Coghlan, S., Cooper, P.: A governance model for the application of ai in health care. J. Am. Med. Inform. Assoc. **27**(3), 491–497 (2020)
25. Sand, M., Durán, J.M., Jongsma, K.R.: Responsibility beyond design: physicians' requirements for ethical medical AI. Bioethics **36**(2), 162–169 (2022)
26. Schork, N.J.: Artificial Intelligence and personalized medicine. Precision Med. In Cancer Therapy, 265–283 (2019)
27. Sim, J.H., Aziz, Y., Mansor, A., Vijayananthan, A., Foong, C.C., Vadivelu, J.: Students' performance in the different clinical skills assessed in osce: what does it reveal? Med. Educ. Online **20**(1), 26185 (2015)
28. Singh, P.: Transforming healthcare through AI: enhancing patient outcomes and bridging accessibility gaps. J. Artifi. Intell. Res. **4**(1), 220–232 (2024)
29. Singla, R., et al.: Developing a canadian artificial intelligence medical curriculum using a delphi study. npj Digital Med.**7**(1), 323 (2024)
30. Thomasian, N.M., Eickhoff, C., Adashi, E.Y.: Advancing health equity with artificial intelligence. J. Public Health Policy **42**(4), 602 (2021)
31. Witheridge, A., Ferns, G., Scott-Smith, W.: Revisiting miller's pyramid in medical education: the gap between traditional assessment and diagnostic reasoning. Int. J. Med. Educ. **10**, 191 (2019)
32. Zhang, P., Kamel Boulos, M.N.: Generative ai in medicine and healthcare: promises, opportunities and challenges. Future Internet **15**(9), 286 (2023)

SMARE—Structure Matching and Recognition Engine for Hand-Drawn Chemical Formulas

Mitra Purandare[1(✉)], Tobias Rothlin[1], Frieder Loch[1], Johannes Huwer[2,3], and Lars-Jochen Thoms[2,3]

[1] Department of Computer Science, Eastern Switzerland University of Applied Sciences, Rapperswil, Switzerland
{mitra.purandare,tobias.rothlin,frieder.loch}@ost.ch
[2] Chair of Science Education, Thurgau University of Teacher Education, Kreuzlingen, Switzerland
{johannes.huwer,lars.thoms}@phtg.ch
[3] Chair of Science Education, University of Konstanz, Konstanz, Germany

Abstract. Expressing chemical compounds in various representations is challenging. This is especially true for novices, since the task demands extensive domain-specific knowledge and spatial visualization skills. To address this challenge, we propose SMARE, our Structure Matching and Recognition Engine for chemical formulas. It interprets hand-drawn molecular structures and identifies and highlights errors and thereby is a fundamental component of educational applications. SMARE leverages a YOLO (You Only Look Once) model to recognize fundamental entities in chemical structures such as atoms or bonds. The dataset for training, validating, and testing the model consists of 1,844 hand-drawn chemical molecular images collected from students. SMARE processes the identified entities to construct an abstract molecular graph. The engine compares the identified molecular graph against a database of known molecules and detects errors such as incorrect bonding, and valency violations. Our fine-tuned YOLO model achieves an accuracy of 93.8% in recognizing chemical entities in hand-drawn molecules. SMARE was tested on 7,909 hand-drawn chemical structures from 519 school students under real-world conditions, successfully identifying numerous errors in their chemical drawings. This demonstrates effectiveness of SMARE as a powerful and practical tool for chemistry education.

Keywords: Chemical education · Image recognition · YOLO · Error analysis · Spatial visualization

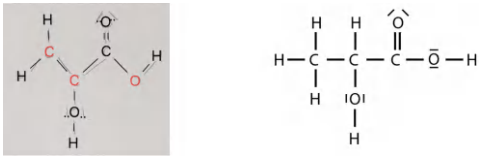
1 Introduction

Chemical compounds can be represented in various ways, such as molecular formulas, skeletal formulas, or perspective drawings, each emphasizing different aspects of chemical bonding. Understanding the representations, choosing the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 124–132, 2025.
https://doi.org/10.1007/978-3-031-98462-4_16

appropriate one for a given purpose, and writing chemical formulas require specific skills. While experts can use and mentally translate between these formats, novices must first learn the underlying rules.

Translating between representations can be practiced with a dedicated application [19]. Such an application needs an algorithm to identify hand-drawn chemical formulas and detect errors (see Fig. 1). The task is challenging due to the diversity of student illustrations and handwriting styles. This paper presents a system that identifies and analyzes hand-drawn molecules to detect and highlight errors when drawing Lewis structures.



(a) Hand-drawn (b) Correct

Fig. 1. (a) Hand-drawn Lewis structure of R-lactic acid with overlay of recognized entities and errors (red, left to right): Missing bonds and (H) atoms at the first and second carbon atoms, two missing lone electron pairs at the oxygen atom. (b) Correct drawing of the R-lactic acid on the right. (Color figure online)

Figure 2 depicts our solution. The contributions are threefold: (1) a dataset of 1,844 hand-drawn chemical molecular images collected from students in Switzerland and Germany [20], (2) YOLOv8 model pipeline to detect molecular entities, which runs locally on consumer devices without cloud processing, and (3) an error detection algorithm to identify errors in the hand-drawn chemical formulas.

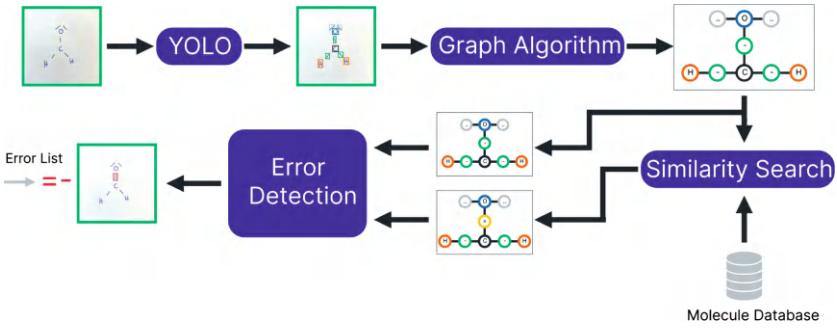


Fig. 2. SMARE for hand-written molecular recognition and error analysis.

2 Related Work

Many students struggle to reason about abstract representations of scientific concepts that require a high level of spatial visualization [14]. While experts can mentally translate between multiple representations [6,8,15], novices must first develop the necessary mental models [1,15,16], often supported by interactive applications [7,21].

Deep neural networks can recognize printed or handwritten chemical structures [2,4,9,11,12,23–25], and even take the context into account [10]. Large databases with (partly already decoded) images of chemical formulas are available (e.g., CASIA-CSDB [3]). Most of the above methods convert a molecular image to SMILES [22] or InChI [5] format which assume correct molecular structures. They lack the ability to encode errors like missing atoms, incorrect bonds, or lone pair errors explicitly. Since SMARE focuses on identifying and correcting structural errors, a molecular graph representation is necessary.

Molscribe [11] and Molminer [23] generate molecular graphs from images but assume error-free input. Molscribe employs a complex encoder-decoder architecture trained on correct drawings. Molminer [23] employs YOLOv5 for distance-based graph construction without error detection.

3 SMARE: Methods and Approach

SMARE must recognize hand-drawn chemical formulas despite variations in handwriting and overlaps between atoms and bonds. Errors occur when the detected structure deviates from the expected molecule. We apply YOLO (You Only Look Once) for real-time recognition. Our approach has three steps: Object Detection, Graph Construction, and Graph Matching with Error Detection.

SMARE is designed for handheld devices. Local processing enhances privacy by not having to transmit images for analysis and ensuring independence from internet connection. This limits AI model and error detection choices, requiring efficiency. To maintain a smooth user experience, SMARE imposes an upper limit of 50 ms for end-to-end processing, from image capture to error display.

Object Detection. We train a YOLOv8 model to detect individual chemical components in an image, such as atoms, bonds, or free electrons. Our full training dataset comprising 1,844 hand-drawn chemical molecules is openly available [20].

These curated examples reflect hand-writing variations and practical learning scenarios. Additionally, 32 examples were incorporated into the training data to address images devoid of molecular structures such as scratched-out drawings. Data augmentation such as rotation, scaling, noise addition, and contrast adjustments are performed in Roboflow. The model is fine-tuned using the usual 80-20 training-validation split. The natural imbalance of elements cannot be mitigated as it reflects the inherent bias present in real molecules.

Graph Construction. The output of the YOLO Model comprises all relevant bounding boxes and their class designations. The graph construction method

selects a random bond as the initial point of reference and establishes an edge between its two nearest atoms. This process is applied iteratively to all the bonds to construct the core structure of the graph. The free electrons and charges are then associated with their proximal atoms.

Graph Matching and Error Detection. If the reference molecule is known, no matching is required but the reference still needs to be aligned with the hand-drawn structure for error detection. Aligning the two graphs even if they differ in orientation and positioning is crucial. Figure 3 shows two molecules to be compared. The hand-written molecules in our dataset are small, averaging only 14 nodes with a maximum of 39 nodes, justifying a brute-force approach shown in Fig. 4. A random starting point is selected from the reference graph, and every atom in the hand-drawn graph is tested as a potential starting point. The alignment with the fewest errors is chosen. Due to molecular symmetry, multiple optimal alignments may exist. The alignment with the minimal number of errors is selected as a chosen alignment for displaying errors (see Fig. 5).

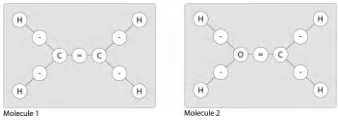


Fig. 3. Compare molecular graphs: hand-drawn (left) vs reference (right).



(a) Alignment 1 (b) Alignment 2

Fig. 4. Different starting points in Fig. 3 result in different path alignments. Paths in the hand-drawn molecule starting with hydrogen (H) are shown on the left. Paths in the reference molecule with different starting points are shown on the right. Each path in the hand-drawn molecule is compared to the paths of the reference graphs. Error scores are mismatches on each path.

If the reference molecule is not known a priori, i.e., the learner scans an arbitrary hand-written molecule, a matching reference needs to be identified. The unknown reference molecule must be carefully identified resulting in a minimal number of errors to not overwhelm the learner. A simple histogram matching algorithm can be employed to reduce the number of reference candidates.

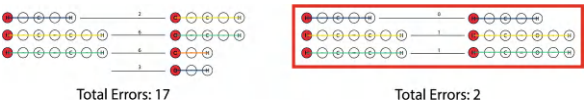


Fig. 5. From the alignments in Fig. 4, the paths with least errors are chosen and the final error count is generated. The alignment with the least count is chosen.

Comparing the number of atoms and bond types in the molecules eliminates positional information. A molecule's histogram need not always be unique. To address this issue, error detection is applied to all the reference candidates. The reference with the least error score is chosen to display errors.

4 Evaluation

To evaluate the generalization capability of the YOLO model, a test set of 100 hand-drawn molecules created by an independent individual is used.

Laboratory Testing. The confusion matrix reveals that the model detected the correct atoms with an accuracy of above 90%. There are minor instances of incorrect atom classification. The majority of the remaining 10% errors are associated with the background class, indicating that the model detects elements that are

Table 1. Normalized confusion matrices for electrons, charges, and bonds

Electrons ground truth column-wise and prediction row-wise

	FreeElectronPair	SingleFreeElectron	Background
FreeElectronPair	0.99	0.09	0.67
SingleFreeElectron	0.00	0.82	0.33
Background	0.01	0.09	0.00

Charges ground truth column-wise and prediction row-wise

	Negative	Positive	Background
Negative	0.78	0.00	0.00
Positive	0.11	1.00	1.00
Background	0.11	0.00	0.00

Bonds ground truth column-wise and prediction row-wise

	DashedTri.	DoubleB.	FullTri.	SingleB.	TripleB.	Background
DashedTri.	1.00	0.00	0.00	0.00	0.00	0.05
DoubleB.	0.00	0.98	0.00	0.00	0.00	0.00
FullTri.	0.00	0.00	0.83	0.00	0.04	0.05
SingleB.	0.00	0.00	0.00	0.98	0.00	0.65
TripleB.	0.00	0.00	0.00	0.00	0.92	0.25
Background	0.00	0.02	0.17	0.02	0.04	0.00

not actually present in the image. Users rarely observe this phenomenon, because these extraneous elements get discarded during graph-construction.

Tables 1 and 2 show the confusion matrices for electrons, charges, bonds, and atoms, showing high overall accuracy. The non-zero entries in the last column in Table 2 show that background may get misidentified as an element. The *Negative* label, representing a negative charge, is the least frequent in the training data and also has the worst test performance. The training data distribution explains why underrepresented classes perform worse. Despite the issues, models exhibit high weighted average accuracy of 93.8%. The weights reflect the prevalence of chemical entities in the training data.

Field Testing. To assess SMARE under real-life conditions, 519 school students tested SMARE via the iOS app OrChemSTAR [17], generating 7,909 hand-drawn scans [18]. Without prior knowledge of drawing structural formulas or organic chemistry, the students learned only through the app. Therefore, Table 3 shows a high error rate. The error detection of the application helped the students refine their structural drawing skills in further exercises.

Table 2. Normalized Confusion Matrix for Atoms (BG = Background).

	Br	C	Cl	F	H	I	N	O	S	BG
Br	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13
C	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.15
Cl	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.04
F	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.02
H	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.28
I	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.11
N	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.02
O	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.98	0.00	0.23
S	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.02
BG	0.10	0.00	0.00	0.00	0.01	0.04	0.00	0.00	0.00	0.00

Table 3. Frequency (in %) of classes of errors in 7,909 scans by school students.

Class	Freq.	Class	Freq.
`Octet rule violated`	50.21	`Missing Bond`	49.18
`Missing Atom`	49.18	`Incorrect Atom`	28.68
`Incorrect Bond`	22.49	`Excess Bond`	21.76
`Excess Atom`	21.58	`Excess Electron`	7.75
`Missing Electron`	7.50	`Detected Natta Projection`	2.52
`Excess Charge`	1.01	`Missing Charge`	0.10
`No Errors found`	27.42		

5 Conclusion and Future Work

SMARE analyzes hand-drawn molecules accurately, achieving 93.8% accuracy in detecting chemical entities. Its robust architecture allows integration of new molecules without retraining. Evaluations with many students confirm its effectiveness to detect errors in chemical drawings in real-world conditions.

While graph construction from YOLO output is efficient, moving beyond a greedy proximity-based algorithm could improve reliability. The error detection algorithm performs well on the molecule sizes that are typically encountered in school exercises. However, the scalability to larger molecules is limited. The role of maximum subgraph isomorphism [13] in error detection requires further exploration for error detection in larger structures.

Acknowledgements. This work is part of the project "OrChemSTAR - Organic Chemistry Science Teaching and Learning with Augmented Reality", funded by the Swiss National Science Foundation (CRSK-1_221108).

Disclosure Statement. The authors declare that they have no competing interests to declare that are relevant to the content of this article.

References

1. Ainsworth, S.: DeFT: a conceptual framework for considering learning with multiple representations. Learn. Instr. **16**(3), 183–198 (2006). https://doi.org/10.1016/j.learninstruc.2006.03.001
2. Bukhari, S.S., Iftikhar, Z., Dengel, A.: Chemical structure recognition (CSR) system: automatic analysis of 2D chemical structures in document images. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, pp. 1262–1267. IEEE (2019). https://doi.org/10.1109/ICDAR.2019.00-41
3. Ding, L., Zhao, M., Yin, F., Zeng, S., Liu, C.L.: A large-scale database for chemical structure recognition and preliminary evaluation. In: 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, pp. 1464–1470. IEEE (2022). https://doi.org/10.1109/ICPR56361.2022.9956654
4. Hagag, A., Omara, I., Alfarra, A.N.K., Mekawy, F.: Handwritten chemical formulas classification model using deep transfer convolutional neural networks. In: 2021 International Conference on Electronic Engineering (ICEEM), Menouf, Egypt, pp. 1–6. IEEE (2021). https://doi.org/10.1109/ICEEM52022.2021.9480627
5. Heller, S.R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D.: InChI, the IUPAC international chemical identifier. J. Cheminformatics **7**(1), 1–34 (2015). https://doi.org/10.1186/s13321-015-0068-4
6. Kohl, P.B., Finkelstein, N.D.: Patterns of multiple representation use by experts and novices during physics problem solving. Phys. Rev. Spec. Top. Phys. Educ. Res. **4**(1), 010111 (2008). https://doi.org/10.1103/PhysRevSTPER.4.010111

7. Kozma, R., Russell, J.: Students becoming chemists: developing representationl competence. In: Gilbert, J.K. (ed.) Visualization in Science Education, pp. 121–145. Springer, Dordrecht (2005)
8. Kozma, R.B.: Use of multiple representations by experts and novices. In: Van Meter, P., List, A., Lombardi, D., Kendeou, P. (eds.) Handbook of Learning from Multiple Representations and Perspectives, 1 edn. Routledge, New York (2020). https://doi.org/10.4324/9780429443961
9. Li, Y., Chen, G., Li, X.: Automated recognition of chemical molecule images based on an improved TNT model. Appl. Sci. **12**(2), 680 (2022). https://doi.org/10.3390/app12020680
10. Musazade, F., Jamalova, N., Hasanov, J.: Review of techniques and models used in optical chemical structure recognition in images and scanned documents. J. Cheminformatics **14**(1), 61 (2022). https://doi.org/10.1186/s13321-022-00642-3
11. Qian, Y., Guo, J., Tu, Z., Li, Z., Coley, C.W., Barzilay, R.: Molscribe: robust molecular structure recognition with image-to-graph generation. J. Chem. Inf. Model. **63**(7), 1925–1934 (2023)
12. Rajan, K., Zielesny, A., Steinbeck, C.: DECIMER: towards deep learning for chemical image recognition. J. Cheminformatics **12**(1), 1–9 (2020). https://doi.org/10.1186/s13321-020-00469-w
13. Raymond, J.W., Willett, P.: Maximum common subgraph isomorphism algorithms for the matching of chemical structures. J. Comput. Aided Mol. Des. **16**(7), 521–533 (2002). https://doi.org/10.1023/A:1021271615909
14. Sahin, D., Yilmaz, R.M.: The effect of augmented reality technology on middle school students' achievements and attitudes towards science education. Comput. Educ. **144**, 103710 (2020). https://doi.org/10.1016/j.compedu.2019.103710
15. Schnotz, W., Bannert, M.: Construction and interference in learning from multiple representation. Learn. Instr. **13**(2), 141–156 (2003)
16. Sunyono, S., Leny, Y., Muslimin, I.: Supporting students in learning with multiple representation to improve student mental models on atomic structure concepts. Sci. Educ. Int. **26**(2), 104–125 (2015)
17. Thoms, L.J.: OrChemSTAR [iOS application] (2024). https://apps.apple.com/de/app/orchemstar/id6636548074
18. Thoms, L.J.: Hand-drawn chemical formulas captured during field testing of the orchemstar app with and without annotations gained using the structure matching and recognition engine SMARE (2025). https://doi.org/10.5281/zenodo.15191840
19. Thoms, L.J., Huwer, J.: Das projekt OrChemSTAR – strukturformeln durch augmented reality zeichnen lernen [the OrChemSTAR project – learning to draw structural formulas using augmented reality]. In: Huwer, J., Wilke, T., Banerji, A. (eds.) Progress in Digitalisation in Chemistry Education 2024, pp. 113–118. Waxmann (2025). https://doi.org/10.31244/9783818850425
20. Thoms, L.J., Rothlin, T.: Hand-drawn chemical formulas used to train the structure matching and recognition engine SMARE (2024). https://doi.org/10.5281/zenodo.15191704
21. van der Meij, J., de Jong, T.: Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. Learn. Instr. **16**(3), 199–212 (2006)
22. Weininger, D.: SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. **28**(1), 31–36 (1988). https://doi.org/10.1021/ci00057a005
23. Xu, Y., et al.: Molminer: you only look once for chemical structure recognition. J. Chem. Inf. Model. **62**(22), 5321–5328 (2022)

24. Xu, Z., Li, J., Yang, Z., Li, S., Li, H.: SwinOCSR: end-to-end optical chemical structure recognition using a swin transformer. J. Cheminformatics **14**(1), 41 (2022). https://doi.org/10.1186/s13321-022-00624-5
25. Zhang, T., Wang, Y., Jin, X., Gu, Z., Zhang, X., He, B.: An auto-grading oriented approach for off-line handwritten organic cyclic compound structure formulas recognition. Comput. Model. Eng. Sci. **135**(3), 2267–2285 (2023). https://doi.org/10.32604/cmes.2023.023229

An Emergent Bottom-Up Categorization of Students' LLMs Usage in an Undergraduate Research Course

Ivan Orozco Vasquez[1(✉)], Romina Mahinpei[1,2], Noureddine Elouazizi[1], and Cristina Conati[1]

[1] University of British Columbia, Vancouver, BC, Canada
iorozco@student.ubc.ca, rm9688@princeton.edu, elouazizin@latentsense.ai, conati@cs.ubc.ca
[2] Princeton University, Princeton, NJ, USA

Abstract. The rise of large language models (LLMs) in education has sparked debate over how they should be used to foster learning. While prior research focuses on problem-solving domains, we investigate how students use LLMs in a seminar-style research methods course in which one of the main activities is critical analyses of research papers. Through assignment-level reports, we derive an emergent bottom-up categorization of LLMs usage by students, revealing diverse strategies from clarifying concepts to generating ideas and generating answers. Our findings highlight both LLMs productive uses and potential over-reliance, emphasizing the need for clear guidelines to handle with care the integration of LLMs in education (This work falls under and complies with University's policies 81 and LR11 regarding ownership and usage of course material for pedagogical quality assurance.)

Keyword: Students Usage of LLMs · Emergent Categorization · Integration of LLMs

1 Introduction

Large language models (LLMs), such as ChatGPT and Claude, are increasingly being used across a wide range of domains, including education. Witnessing the growing use of LLMs among students, the educational community has expressed divergent views on their role in learning [3, 4, 6]. Some support the full integration of LLMs into courses and advocate for the benefits that LLMs offer to students, such as instant and personalized feedback [3, 4]. Others counterargue that LLMs will hinder students' abilities and prefer to ban the usage of LLMs in courses [6]. Given this divide, understanding how students actually use LLMs and how they should be used to foster learning is crucial. Recent surveys summarizing the state-of-the-art LLMs usage in education further highlight the need for the exploration of students' experiences with LLMs across different educational domains, emphasizing the need to investigate how diverse student groups engage with this technology [1, 2].

I. O. Vasquez and R. Mahinpei—Equal contributions.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 133–139, 2025.
https://doi.org/10.1007/978-3-031-98462-4_17

However, much of the existing research has focused on problem-solving domains, particularly programming [3–5, 10–13]. In this paper, we extend and contribute to this growing body of work by examining students' LLMs usage as it emerges in a course that relies on a very different set of learning activities and skills than problem-solving. Namely, we focus on a third-year undergraduate course where students learn how to read, reason about, and critique research papers. The course is part of a multidisciplinary program offered collaboratively by our institution's Computer Science, Linguistics, Philosophy, and Psychology departments. Unlike the courses in prior studies, this course involves a student population with very diverse backgrounds given that the students come from four different departments.

In this paper, we present a fine-grained analysis of emerging LLMs usage patterns based on assignment-level data collected throughout the semester that included student reports of how they used LLMs, as required by university policy. As part of our exploration, we propose an emergent categorization of students' interactions with LLMs. Prior studies have introduced LLMs usage frameworks through a prescribed top-down approach, wherein the authors predefined the categories based on theoretical insights or reviews of the existing literature [7–9]. Our categorization is different. We derived it from a bottom-up, data-driven analysis of student usage patterns, as they emerged in a natural educational setting.

2 Literature Review

LLMs-powered AI chatbots are increasingly used in education, with some educators fully integrating them into courses, where students generally find them beneficial [3–5]. For example, a CS1 course allowed unrestricted LLMs usage, and most students reported positive learning experiences [3]. Similarly, a software engineering course used LLMs for code explanations, which students found helpful [4]. However, some educators are resistant, with concerns about AI-assisted cheating leading to calls for restrictions or bans [6]. This divide highlights the need to understand how students actually use LLMs [1, 2]. Most studies on LLMs usage in education focus on problem-solving domains, showing that students find LLMs useful for programming tasks such as generating code and debugging [3–5]. To expand this research beyond problem-solving, we investigate LLMs usage in a third-year course focused on reading and analyzing research papers. This shifts the focus from programming to activities involving critical reading, comprehension, and synthesis of academic literature.

We also propose an emergent categorization of students' interactions with LLMs, informed by our data. Previous frameworks categorized LLMs usage through top-down approaches based on theoretical models or literature reviews [7–9]. For instance, one framework outlined seven AI integration methods in education using predefined categories [7], while another created a taxonomy of human-LLMs interaction modes from HCI literature [8]. A third study proposed a taxonomy for designing prompts but did not examine student-generated prompts or real-world educational applications [9]. In contrast, our work focuses on student interactions with LLMs in a specific educational setting, using a bottom-up approach to categorize emergent usage patterns. This method offers insights into how students organically engage with LLMs, providing a

nuanced understanding of AI's role in supporting learning. By directly observing students' behavior, our study bridges gaps left by prescriptive taxonomies, contributing a novel perspective on LLMs integration in academic contexts.

3 Course Structure

Our exploratory study was conducted with 37 undergraduate students in a third-year undergraduate course offered at a public North American institution. The course's objective was to teach students research methods in the areas of Cognitive Systems, covering the skills and principles common to all forms of research, such as critical thinking and communication. As part of the course, students were assessed using 14 individual take-home assignments and three individual in-class midterms, with the take-home assignments being graded for completion while the in-class midterms were graded for quality. The take-home assignments required students to read an assigned research paper on the general topic of human-AI interaction and complete two tasks:

1. **Critical Summary:** Summarize the assigned paper, focusing on the motivations of the work, the proposed solution, how the solution was evaluated, the contributions of the work, and specific aspects related to the details of the human-AI interaction.
2. **Discussion Points:** Brainstorm discussion points identifying and discussing weaknesses in the presented research, relating the research to general issues in the field, making connections with other readings, and proposing alternative approaches.

Students were permitted to use LLMs for their take-home assignments, provided they reported their usage by submitting reports detailing their LLMs usage for each relevant submission. This reporting, mandated by institutional policies, required students to document the specific LLMs used, the input prompts, and the generated output. As for the in-class midterm, students were required to work on tasks similar to their assignments (i.e. Critical Summary and Discussion Points for an assigned paper) without any form of assistance.

4 Categorization of Students' LLMs Usage

Given the LLMs usage data collected throughout the semester, we developed a categorization using an emergent bottom-up approach. To create our categorization, one author initially reviewed student LLMs submissions and developed categories based on observed usage patterns. Independently, another author categorized the same submissions using this initial classification while considering modifications, such as adding new categories. We calculated percentage agreement (94%) and Cohen's Kappa (0.92) to measure consistency between the two authors. Afterward, the two authors reviewed their classifications together, resolving discrepancies and refining category descriptions as needed.

During this refinement process, disagreements arose regarding the broad coverage of some classifications, leading to multiple modifications to narrow their scope. For example, students' use of LLMs to request information was initially coded as one category. However, upon closer examination, two distinct categories emerged: "Confirming Ideas"

and "Clarifying Concepts." The distinction lies in the source of the information: "Confirming Ideas" involves discussing a student's own ideas, whereas "Clarifying Concepts" refers to seeking factual information relating to the lecture material.

A similar refinement was needed for the category "Summarizing", originally introduced to capture any LLMs usage in which students asked the LLMs to summarize the paper. As we progressed with the data analysis, we noticed that some students submitted the summaries generated by the LLMs as their final answers, whereas others would use them as their basis to generate the critical summary and discussion points requested. Therefore, we created two different categories to capture these behaviours. Instances where a student asked the LLMs to summarize a paper and directly submitted it as their summary, were categorized as "Generating Answers." However, instances where the LLMs-generated summary was further elaborated by the student were categorized as "Summarizing". This distinction required considering both the student's initial request to the LLMs as well as how they used it.

After multiple iterations and refinements, we finalized our categorization into seven categories, as shown in Table 1. We note that although our categories are distinct, they are not mutually exclusive. Three submissions fit into multiple categories either due to prompts aligning with more than one classification or because students engaged in multiple exchanges with the LLMs, incorporating different types of text generation within a single submission.

Table 1. Our categorization of LLMs usage based on students' reports.

Category	Description: Students…
IL: Improving Language	Write text and ask the LLMs to revise it
PB: Paragraph from Bullets	Write input bullet points and ask the LLMs to make it into a paragraph
CC: Clarifying Concepts	Ask LLMs to clarify concepts from the course
CI: Confirming Ideas	Propose an idea and ask LLMs if it is feasible
GI: Generating Ideas	Ask LLMs for ideas on what to write
S: Summarizing	Ask LLMs to summarize the entire paper
GA: Generating Answers	Ask LLMs to generate the answer from scratch

5 Students' LLMs Usage Trends

Fewer than 30% of students (10 out of 37) used LLMs throughout the course. Figure 1 (left) displays the overall breakdown of LLMs usage by the categories proposed in Table 1 throughout the semester. Although the true usage may be slightly higher given that there may be students who did not declare their LLMs usage due to negligence, it is unlikely for the true usage to be significantly different given that the students were permitted the use of AI without penalty as long as they reported their usage.

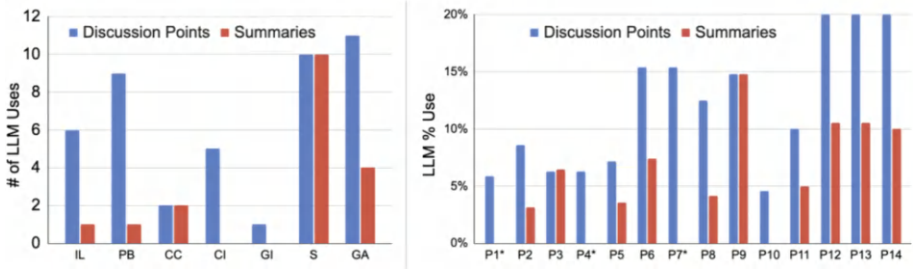


Fig. 1. LLMs usage breakdown by the proposed categories in Table 1 (left) and the percentage of students that used LLMs for the summaries and discussion points tasks in each assignment (right). The starred assignments in the right figure are those that only required discussion points (i.e., no critical summary).

Figure 1 (right) shows the percentage of students who used LLMs for the two tasks assigned in each assignment – discussion points and critical summary. A notable observation includes the increasing LLMs usage for later assignments, especially for the discussion points task. This could be the result of students becoming busier with work from other courses.

In addition to the aggregated views, Fig. 2 provides a temporal view of LLMs usage broken down by the proposed categories at the level of each assignment (x-axis) and each student (y-axis) who reported using LLMs throughout the course. The rows, representing students, are ordered from highest (bottom) to lowest (top) usage.

Observing Fig. 2, we note that two submissions from Student 33 and one submission from Student 34 were assigned multiple categories. These LLM interactions were ones where a student would ask the LLM multiple questions in a single session, with some questions belonging to one category and the remaining questions to another. We also observe that students tend to be consistent in the way that they use LLMs. For example, Student 2 solely used LLMs for summarizing and Student 34 mainly used LLMs to generate paragraphs from their own bullet points.

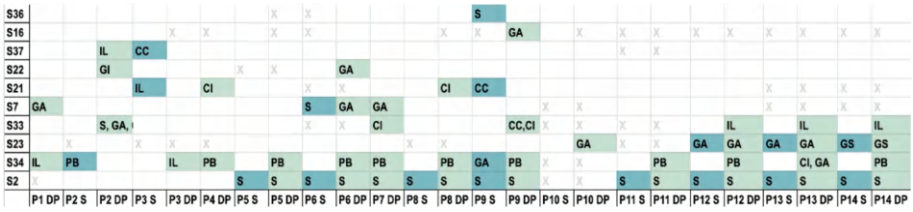


Fig. 2. LLM usage categories across time for each student who used LLMs at least once across assignments. Submissions from students who did not complete an assignment are marked with 'X'. Submissions from students who did not use LLMs are colored white, those who used LLMs for discussion points (DP) are in light green, and those who used LLMs for critical summaries (S) are in teal.

Among the students who used LLMs, two distinct groups emerged: those who used them occasionally (1–2 times, top 4 rows) and those who used them regularly (remaining 6 rows). Informal feedback from some of these students revealed the following.

For students who occasionally used LLMs, one reported doing so to help them during busy work periods of the semester. For another, the occasional use was born out of wanting to assess the helpfulness of LLMs at the beginning of the course, but then they decided to stop using it for assignments because they realized that it wasn't helping them learn as much as they wanted. One of the students who used LLMs regularly during the second part of the course expressed their concern around how using LLMs at the beginning of the course would intrude with their learning, but they felt more comfortable using them once they felt they had acquired the knowledge needed to generate good critical summaries and discussion points. The sentiment that using LLMs would interfere with one's learning was also echoed by one of the students who did not use LLMs at all.

6 Conclusion

Our study explores how students use LLMs in a research-methods critical thinking course, revealing varied usage patterns with implications for learning, assessment, and instructional design. Some students used LLMs for support, such as clarifying concepts or refining work, while others relied on them for idea generation or direct answers, raising concerns about balancing efficiency with effective learning. These findings highlight the need for clear guidelines to help students use LLMs effectively without fostering over-reliance. A limitation of our study is the reliance on data from a single course, which may affect the generalizability of our findings. With only 37 data points, some trends may not apply to broader populations. Additionally, students may have underreported LLMs usage due to negligence. Future work includes collecting data from subsequent courses to validate our findings and exploring grade incentives to encourage accurate reporting. We also plan to expand this research across diverse courses and institutions to refine best practices for integrating LLMs in education.

Acknowledgments. This work was funded by the UBC Skylight grant PM015111, falls under and complies with University's policies 81 and LR11 regarding ownership and usage of course material for pedagogical quality assurance.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Labadze, L., Grigolia, M., Machaidze, L.: Role of AI chatbots in education: systematic literature review. Inter. J. Educ. Technol. Higher Educ. **20**(56) (2023). https://doi.org/10.1186/s41239-023-00426-1
2. Memarian, B., Doleck, T.: ChatGPT in education: Methods, potentials, and limitations. Comput. Human Behav. Artifi. Hum. **1**(2) (2023). https://doi.org/10.1016/j.chbah.2023.100022

3. Vadaparty, A., et al.: CS1-LLM: Integrating LLMs into CS1 Instruction. In: Proceedings of the 2024 on Innovation and Technology in Computer Science Education, V.1 (2024). https://doi.org/10.1145/3649217.3653584
4. MacNeil, S., et al.: Experiences from using code explanations generated by large language models in a web software development E-Book. In: Proceedings of the 54th ACM Technical Symposium on Computer Science Education, V.1 (2023). https://doi.org/10.1145/3545945.3569785
5. Rasnayaka, S., Wang, G., Shariffdeen, R., Iyer, G.N.: An empirical study on usage and perceptions of LLMs in a software engineering project. In: Proceedings of the 1st International Workshop on Large Language Models for Code (2024). https://doi.org/10.1145/3643795.3648379
6. Lau, S., Guo, P.: From "Ban it till we understand it" to "resistance is futile": how university programming instructors plan to adapt as more students use AI code generation and explanation tools such as ChatGPT and GitHub copilot. In: Proceedings of the 2023 ACM Conference on International Computing Education Research V.1 (2023). https://doi.org/10.1145/3568813.3600138
7. Mollick, E.R., Mollick, L.: Assigning AI: seven approaches for students, with prompts. The Wharton School Research Paper (2023). https://doi.org/10.2139/ssrn.4475995
8. Gao, J., Gebreegziabher, S.A., Choo, K.T.W., Li, T.J.J., Perrault, S.T., Malone, T.W.: A taxonomy for human-LLM interaction modes: an initial exploration. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (2024). https://doi.org/10.1145/3613905.3650786
9. Karmaker Santu, S.K., Feng, D.: TELeR: A general taxonomy of llm prompts for benchmarking complex tasks. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp. 14197–14203 (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.946
10. Bull, C., Kharrufa, A.: Generative artificial intelligence assistants in software development education: a vision for integrating generative artificial intelligence into educational practice, not instinctively defending against it. IEEE Software (2024)
11. Kazemitabaar, M., Hou, X., Henley, A., Ericson, B.J., Weintrop, D., Grossman, T.: How novices use LLM-based code generators to solve CS1 coding tasks in a self-paced learning environment. In: Proceedings of the 23rd Koli Calling International Conference on Computing Education Research (2023). https://doi.org/10.1145/3631802.3631806
12. Ghimire, A., Edwards, J.: Coding with AI: how are tools like ChatGPT being used by students in foundational programming courses. Artifi. Intell. Educ. (2024). https://doi.org/10.1007/978-3-031-64299-9_20
13. Brender, J., El-Hamamsy, L., Mondada, F., Bumbacher, E. Who's helping who? when students use chatgpt to engage in practice lab sessions. Artificial Intelligence in Education, pp. 235–249. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-64302-6_17



Large Language Models for Zero-Shot Exercise Recommendation in Adaptive Learning

Tengju Li[1], Cunling Bian[1], Kaiquan Chen[1], and Weigang Lu[1,2(✉)]

[1] Department of Educational Technology, Ocean University of China, Qingdao 266100, China
luweigang@ouc.edu.cn

[2] College of Computer Science and Technology, Ocean University of China, Qingdao 266100, China

Abstract. Personalized exercise recommendation is crucial for effective adaptive learning systems. However, existing data-driven approaches are often limited by their reliance on extensive student-exercise interaction data, which is frequently unavailable or expensive to acquire. To address this challenge, we propose a zero-shot exercise recommendation framework powered by Large Language Models (LLMs). Inspired by human educators, our three-stage framework utilizes LLMs to predict knowledge states, identify weak concepts, and recommend relevant exercises without requiring domain-specific training. Experiments on a real-world educational dataset show that our zero-shot approach achieves comparable performance to data-driven methods, highlighting its potential for practical and scalable adaptive learning, particularly in data-limited scenarios.

Keywords: Adaptive learning · Recommendation Systems · Large Language Models · Zero-Shot Learning

1 Introduction

Adaptive learning systems aim to offer personalized learning experiences tailored to the unique needs of individual learners [3]. A crucial component of these systems is learning path recommendation, which helps to suggest personalized learning materials (such as exercises) based on the learner's current state [15]. The rapid growth of online learning platforms has resulted in an overwhelming abundance of learning materials. To prevent information overload and guide learners effectively, personalized learning path recommendation is essential [5].

To enable adaptive learning, a variety of data-driven artificial intelligence algorithms, such as collaborative filtering [9,13], neural networks [7,18], and reinforcement learning [2,6,10,16], have been successfully applied. These methods learn to predict optimal learning paths from historical student-exercise

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 140–147, 2025.
https://doi.org/10.1007/978-3-031-98462-4_18

interaction data. However, the reliance on large, high-quality datasets of student exercise behaviors poses significant challenges [14]. Collecting and curating these datasets is often expensive and time-consuming, requiring substantial resources and expertise. Furthermore, in many real-world educational scenarios, student exercise data is often noisy, incomplete, or insufficient for robust training, particularly in niche subject areas [20]. This problem significantly hinders the widespread adoption of these methods, as their offline performance often fails to translate to real-world educational settings.

To address this, we propose leveraging Large Language Models (LLMs) for zero-shot exercise recommendation. LLMs' remarkable reasoning and generalization abilities, along with their capacity for natural language understanding and knowledge representation, make them well-suited for this task [19]. By carefully crafting prompts, LLMs can analyze a learner's knowledge state and recommend exercises without domain-specific training data [21], reducing the cost and complexity of building adaptive learning systems. Inspired by the way human teachers guide students, we introduce a three-stage exercise recommendation framework. In this framework, LLMs continuously track the learner's knowledge state, identify weak knowledge concepts, and generate exercise recommendations. Experiments on a real-world educational dataset show our zero-shot method provides effective and coherent learning paths, comparable to data-driven methods, with significant potential for real-world applications.

2 Problem Statement

We formally define the exercise recommendation problem as follows: Given a student's interaction history $\mathcal{X}_t = \{\langle q_1, r_1\rangle, \langle q_2, r_2\rangle, \ldots, \langle q_t, r_t\rangle\}$, where $\langle q_i, r_i\rangle$ denotes the i-th interaction with exercise q_i and the corresponding result $r_i \in \{\text{success}, \text{failure}\}$, the system aims to recommend the next exercise q_{t+1} from an exercise bank $\mathcal{E}$. The system also has access to exercise metadata $\mathcal{A}$, a domain knowledge graph $\mathcal{G} = (\mathcal{C}, \mathcal{R})$ defining concepts $\mathcal{C}$ and prerequisite relations $\mathcal{R}$, and a target concept set $\mathcal{C}' = \{c_1, c_2, \ldots, c_n\}$ for mastery. The objective is to iteratively select exercises $q_{t+1}, q_{t+2}, \ldots, q_{t+m}$ to construct an adaptive learning path $\mathcal{P} = [q_{t+1}, q_{t+2}, \ldots, q_{t+m}]$ that efficiently guides the student to master all concepts in $\mathcal{C}'$ with a minimal number of exercises.

3 Three-Stage Exercise Recommendation Framework Based on LLMs

In contrast to data-driven methods heavily reliant on extensive historical interaction data, experienced educators offer personalized guidance by leveraging their expertise to analyze exercise content and discern crucial information, such as knowledge concepts and difficulty levels [17]. This enables rapid assessment of student mastery, identification of learning gaps, and selection of suitable subsequent exercises. Inspired by this process, we introduce a novel three-stage framework for zero-shot exercise recommendation, powered by LLMs. This framework

emulates the expert decision-making of human educators, facilitating personalized exercise recommendations without the need for domain-specific training data. As illustrated in Fig. 1, the framework decomposes the exercise recommendation task into three stages, each strategically employing LLM capabilities to maximize recommendation effectiveness and personalization.

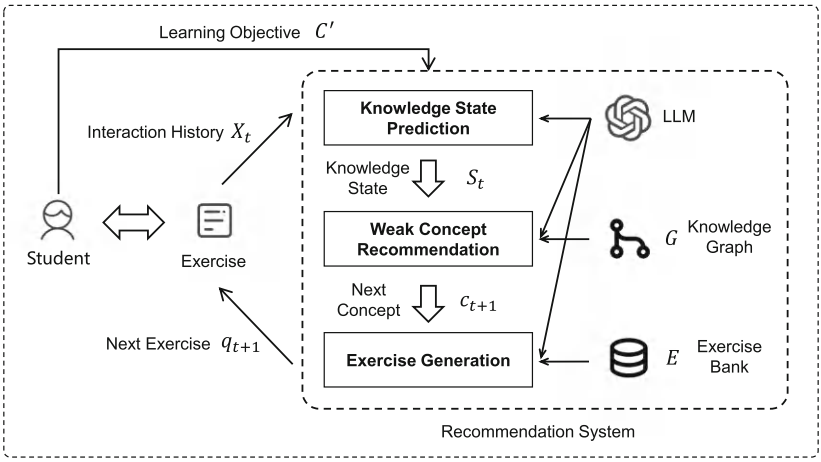


Fig. 1. Three-stage exercise recommendation framework based on LLMs.

Step 1: Knowledge State Prediction. The knowledge state prediction stage infers a student's proficiency across knowledge concepts from their interaction history. Unlike knowledge tracing, which focuses on predicting future performance, this stage emphasizes evaluating current concept mastery. To facilitate LLM processing, the exercise sequence $\mathcal{X}_t$ and metadata $\mathcal{A}$ are transformed into a textual representation $\mathcal{X}'_t$, enriching exercise information to optimize LLM reasoning. This transformation is formalized as:

$$\mathcal{X}_t, \mathcal{A} \rightarrow \mathcal{X}'_t, \tag{1}$$

where $\mathcal{X}'_t = \{\langle q_1, a_1, r_1\rangle, \langle q_2, a_2, r_2\rangle, \ldots, \langle q_t, a_t, r_t\rangle\}$, with a_t denoting supplementary exercise information (e.g., exercise text, associated concepts), informed by pedagogical principles and prior research [8]. Subsequently, the LLM, M_{KT}, analyzes $\mathcal{X}'_t$ and the target concept set $\mathcal{C}'$ to generate the knowledge state $\mathcal{S}_t$ at time t:

$$\mathcal{S}_t = M_{KT}(\mathcal{X}'_t, \mathcal{C}'), \tag{2}$$

where $M_{KT}(\cdot) = \arg\max_\omega P(\omega \mid \cdot, \omega_{KT_prompts})$. The knowledge state $\mathcal{S}_t = \{\langle c_1, l_1\rangle, \langle c_2, l_2\rangle, \ldots, \langle c_i, l_i\rangle\}$ quantifies the learner's mastery $l_i \in \{0, 1, 2, 3, 4, 5\}$ for each concept c_i on a discrete 0–5 scale. $\omega_{KT_prompts}$ represents the prompts guiding the LLM for knowledge state prediction.

Step 2: Weak Concept Recommendation. The weak concept recommendation stage identifies the most essential concept for the learner's next practice. Based on the knowledge state $\mathcal{S}_t$, the model decides whether to recommend reviewing previously practiced concepts or progressing to new, subsequent ones. To inform this decision, the domain knowledge graph $\mathcal{G}$ is integrated, providing conceptual relationships. The recommended concept c_{t+1} is then generated by the LLM, M_{CR}:

$$c_{t+1} = M_{CR}(\mathcal{X}'_t, \mathcal{S}_t, \mathcal{C}', \mathcal{G}), \tag{3}$$

where $M_{CR}(\cdot) = \arg\max_{\omega} P(\omega \mid \cdot, \omega_{CR_prompts})$, and $\omega_{CR_prompts}$ represents prompts for guiding weak concept recommendation.

Step 3: Exercise Generation. The exercise generation stage selects the optimal exercise q_{t+1} from the exercise bank $\mathcal{E}$, informed by the current knowledge state and the recommended concept c_{t+1}. Given the potential scale of $\mathcal{E}$, directly processing all exercises with the LLM would be computationally expensive and exceed input token limits. Therefore, we initially filter $\mathcal{E}$ based on c_{t+1} to create a candidate set $\mathcal{E}_{c_{t+1}}$ containing exercises related to c_{t+1}:

$$c_{t+1}, \mathcal{E} \rightarrow \mathcal{E}_{c_{t+1}}. \tag{4}$$

Subsequently, the LLM, M_{ER}, recommends an exercise by selecting q_{t+1} from $\mathcal{E}_{c_{t+1}}$:

$$q_{t+1} = M_{ER}(\mathcal{X}'_t, \mathcal{S}_t, \mathcal{E}_{c_{t+1}}), \tag{5}$$

where $M_{ER}(\cdot) = \arg\max_{\omega} P(\omega \mid \cdot, \omega_{ER_prompts})$, and $\omega_{ER_prompts}$ represents prompts guiding exercise recommendation. Figure 2 illustrates examples of the prompts used in the three-stage framework.

4 Experiments

4.1 Experimental Settings

To evaluate the effectiveness of the proposed method, we conducted a series of simulation experiments. We began by training a virtual student simulator based on the Deep Knowledge Tracing (DKT) model [1]. This simulator was trained using data from the "Integer Addition and Subtraction" unit of the real-world educational dataset XES3G5M [11], a subset comprising 81 exercises, 12 knowledge concepts, and 77,422 student interaction records. The simulator was designed to emulate student knowledge acquisition, responding to exercise recommendations based on its dynamically updated knowledge state [6,10,16]. Student knowledge level was quantified by the average mastery probability p output by the simulator, with higher p values indicating greater mastery. Consequently, effective recommendation strategies should demonstrably accelerate the increase of p. Our proposed zero-shot approach was compared against the following baseline methods:

```
KT_system_prompt =
You are an experienced mathematics educator tasked
with evaluating a student's mathematical proficiency.
### Task:
Analyze the student's **Exercise Records** to assess
their mastery level for each concept within the
**Target Concept Set**.
### Input Format:
The input is divided into two sections: **Exercise
Records** and **Target Concept Set**.
**1. Exercise Records:**
Exercises are listed in the order they were completed
by the student. Each exercise is enclosed within
`<Exercise>` tags:
<Exercise>
<Exercise id>: [Exercise ID]
<Content>: [Exercise description]
<Knowledge Concepts>: [List of knowledge concepts
covered in the exercise, e.g., ["Concept A", "Concept
B"]]
<Is_Correct>: [1 if correct, 0 if incorrect]
</Exercise>
**2. Target Concept Set:**
A list of concepts the student is expected to master:
[concept_1, concept_2, ..., concept_n]
### Output Format:
For each concept in the **Target Concept Set**,
provide a mastery score on a scale of 0 to 5,
indicating the student's current understanding.
Provide your results in JSON format, mapping each
concept to its mastery score. Output *only* the JSON,
with no additional explanations or commentary.
Example JSON Output:
{"concept_1": 4, "concept_2": 1, "concept_n": 0}
```

```
CR_system_prompt =
You are an expert learning path planner in
mathematics education. Your task is to recommend the
most appropriate next concept for a student to learn,
aiming to optimize their learning progression towards
mastering a target set of concepts.
### Task:
Recommend the **Next Concept to Learn** for a student,
by considering their **Exercise Records**, **Current
Knowledge State**, **Target Concept Set**, and
**Knowledge Graph**.
### Input Format:
The input is divided into four sections: **Exercise
Records**, **Current Knowledge State**, **Target
Concept Set**, and **Knowledge Graph**.
**1. Exercise Records:**
 (Same format as in KT_system_prompt)
**2. Current Knowledge State:**
 (Same format as in KT_system_prompt)
**3. Target Concept Set:**
 (Same format as in KT_system_prompt)
**4. Knowledge Graph:**
Represents prerequisite relationships between
concepts.
Example:
{"Basic Addition": [],
 "Basic Subtraction": ["Basic Addition"],
 "Multiplication": ["Basic Addition", "Basic
Subtraction"]}
This example indicates "Basic Subtraction" requires
"Basic Addition", and "Multiplication" requires both.
### Output Format:
Recommend **one** concept from the **Target Concept
Set** as the most suitable next concept for the
student to learn. Provide your result in JSON format,
outputting *only* the JSON, with no additional
explanations or commentary.
Example JSON Output:
{"Recommended Concept": "Basic Subtraction"}
```

```
ER_system_prompt =
You are an experienced mathematics educator, highly
skilled in selecting practice exercises that will
maximize their knowledge gain across all target
concepts.
### Task:
Recommend the **Single Best Exercise** from a
comprehensive **Exercise Bank** that will **Maximize
the Student's Overall Knowledge State**.  Your goal
is to identify the exercise that, upon successful
completion by the student, will lead to the greatest
improvement in their understanding across the entire
**Target Concept Set**.
### Input Format:
The input is divided into three sections: **Exercise
Records**, **Current Knowledge State**, and
**Exercise Bank**.
**1. Exercise Records:**
 (Same format as in KT_system_prompt)
**2. Current Knowledge State:**
 (Same format as in KT_system_prompt)
**3. Exercise Bank:**
A set of all available exercises. Each exercise is
enclosed within `<ExerciseBankItem>` tags:
<ExerciseBankItem>
<Exercise id>: [Exercise ID]
<Content>: [Exercise description]
<Knowledge Concepts>: [List of knowledge concepts
covered in the exercise, e.g., ["Concept A", "Concept
B"]]
</ExerciseBankItem>
### Output Format:
Recommend **one** exercise from the **Exercise Bank**
that you judge will best maximize the student's
overall Knowledge State. Provide your result in JSON
format, outputting *only* the JSON, with no
additional explanations or commentary.
Example JSON Output:
{"Recommended Exercise ID": "31"}
```

Fig. 2. Example prompts for the knowledge state prediction, weak concept recommendation, and exercise generation stages.

- Random: Exercises were selected randomly from the exercise bank.
- DKT-50% [4]: Based on the DKT model, this strategy recommended exercises for which the predicted mastery probability was approximately 50%.
- Monte Carlo Search (MCS) [12]: Based on the DKT model, MCS was employed to search learning paths maximizing the mastery probability.
- Reinforcement Learning (RL) [2]: This approach utilized deep reinforcement learning to train an agent via the student simulator, dynamically optimizing recommendation strategies based on historical interaction data.

Excluding the Random Recommendation, all data-driven baseline methods underwent training and fine-tuning on 50% of the XES3G5M dataset. Our zero-shot approach considered three prominent LLMs: GPT-4, Gemini-2, and GLM-4, accessed via their respective APIs.

4.2 Performance Comparison

We set the learning path length to 10. Each recommendation method was executed five times, and the mean results are reported in Table 1. The results clearly demonstrate that both data-driven methods (DKT-50%, MCS, RL) and LLM-based methods (Gemini-2, GPT-4, GLM-4) significantly improved students' final knowledge levels, outperforming the random recommendation approach. This highlights the effectiveness of personalized recommendations. Notably, within the three-stage recommendation framework, Gemini-2 and GPT-4 performed exceptionally well. Gemini-2 achieved the highest final knowledge level, surpassing data-driven methods like RL and MCS. This suggests that LLM-based zero-shot methods can recommend effective learning paths that lead to significant knowledge gains.

Table 1. Final knowledge level results.

Method	Random	DKT-50%	MCS	RL	Gemini-2	GPT-4	GLM-4
Final Level	0.7482	0.77636	0.8054	0.8151	**0.8219**	0.8189	0.7993

Figure 3 further illustrates the performance changes as the learning path length increases. Gemini-2 and RL consistently outperformed other methods. While RL showed rapid knowledge improvement between path lengths 2–4, its performance eventually plateaued. Gemini-2, however, maintained a sustained knowledge increase, ultimately achieving superior results. Figure 4 presents a case study of learning paths generated by Gemini-2 and RL. RL tended to recommend exercises within mastered knowledge concepts, likely to maintain a high knowledge level. In contrast, Gemini-2 focused on weak concepts, guiding learners toward deeper engagement and more effective knowledge growth. This demonstrates that the three-stage recommendation framework enables more reasonable learning paths, fostering continuous knowledge advancement.

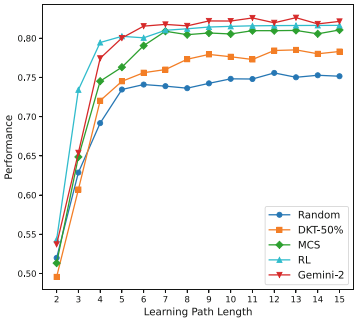


Fig. 3. Performance comparison at different path lengths.

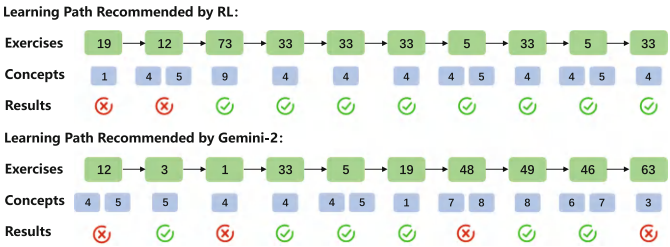


Fig. 4. A case study of learning paths generated by RL and Gemini-2.

5 Conclusion

This paper proposed a novel three-stage framework leveraging LLMs for zero-shot exercise recommendation in adaptive learning systems. Our approach addresses the critical challenge of data scarcity that often hinders the deployment of traditional data-driven recommendation methods. By emulating the decision-making process of experienced human teachers, our framework decomposes the

exercise recommendation task into knowledge state prediction, weak concept recommendation, and exercise generation, each powered by LLMs. Experimental results on the XES3G5M dataset demonstrated that our zero-shot approach achieves performance comparable to state-of-the-art data-driven methods. This demonstrates that LLMs, with appropriate prompts, can effectively analyze learner interactions and create effective learning paths without extensive student data, highlighting their potential for practical, cost-effective adaptive learning systems, particularly in data-limited scenarios.

Acknowledgements. This work has been supported by National Natural Science Foundation of China (No. 62277045) and Natural Science Foundation of Shandong Province (No. ZR 2021MF011).

References

1. Abdelrahman, G., Wang, Q., Nunes, B.: Knowledge tracing: a survey. ACM Comput. Surv. **55**(11), 1–37 (2023)
2. Ai, F., et al.: Concept-aware deep knowledge tracing and exercise recommendation in an online learning system. In: Proceedings of the 12th International Conference on Educational Data Mining (2019)
3. Chen, Y., Li, X., Liu, J., Ying, Z.: Recommendation system for adaptive learning. Appl. Psychol. Meas. **42**(1), 24–41 (2018)
4. Chounta, I.A., McLaren, B.M., Albacete, P., Jordan, P., Katz, S.: Modeling the zone of proximal development with a computational approach. In: Proceedings of the 10th International Conference on Educational Data Mining, pp. 56–57 (2017)
5. Dietrich, J., Greiner, F., Weber-Liel, D., Berweger, B., Kämpfe, N., Kracke, B.: Does an individualized learning design improve university student online learning? A randomized field experiment. Comput. Hum. Behav. **122**, 106819 (2021)
6. Huang, Z., et al.: Exploring multi-objective exercise recommendations in online education systems. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1261–1270 (2019)
7. Jiang, W., Pardos, Z.A., Wei, Q.: Goal-based course recommendation. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge, pp. 36–45 (2019)
8. Li, H., et al.: Explainable few-shot knowledge tracing. arXiv preprint arXiv:2405.14391 (2024)
9. Lin, C.F., Yeh, Y.C., Hung, Y.H., Chang, R.I.: Data mining for providing a personalized learning path in creativity: an application of decision trees. Comput. Educ. **68**, 199–210 (2013)
10. Liu, Q., et al.: Exploiting cognitive structure for adaptive learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 627–635 (2019)
11. Liu, Z., et al.: Xes3g5m: a knowledge tracing benchmark dataset with auxiliary information. In: Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 32958–32970 (2023)
12. Metropolis, N., Ulam, S.: The Monte Carlo method. J. Am. Stat. Assoc. **44**(247), 335–341 (1949)

13. Nabizadeh, A.H., Goncalves, D., Gama, S., Jorge, J., Rafsanjani, H.N.: Adaptive learning path recommender approach using auxiliary learning objects. Comput. Educ. **147**, 103777 (2020)
14. Nabizadeh, A.H., Leal, J.P., Rafsanjani, H.N., Shah, R.R.: Learning path personalization and recommendation methods: a survey of the state-of-the-art. Expert Syst. Appl. **159**, 113596 (2020)
15. Siren, A., Tzerpos, V.: Automatic learning path creation using OER: a systematic literature mapping. IEEE Trans. Learn. Technol. **15**(4), 493–507 (2022)
16. Tang, X., Chen, Y., Li, X., Liu, J., Ying, Z.: A reinforcement learning approach to personalized learning recommendation systems. Br. J. Math. Stat. Psychol. **72**(1), 108–135 (2019)
17. Wan, H., Zhong, Z., Tang, L., Gao, X.: Pedagogical interventions in SPOCs: learning behavior dashboards and knowledge tracing support exercise recommendation. IEEE Trans. Learn. Technol. **16**(3), 431–442 (2023)
18. Wu, Z., Li, M., Tang, Y., Liang, Q.: Exercise recommendation based on knowledge concept prediction. Knowl.-Based Syst. **210**, 106481 (2020)
19. Xu, L., et al.: Prompting large language models for recommender systems: a comprehensive framework and empirical analysis. arXiv preprint arXiv:2401.04997 (2024)
20. Yu, J., et al.: MoocRadar: a fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in MOOCs. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2924–2934 (2023)
21. Zhang, J., Xie, R., Hou, Y., Zhao, X., Lin, L., Wen, J.R.: Recommendation as instruction following: a large language model empowered recommendation approach. ACM Trans. Inf. Syst. (2023)

Learning by Teaching: Enhancing Music Learning Through LLM-Based Teachable Agents

Lingxi Jin[1], Baicheng Lin[2], Mengze Hong[3], Hyo-Jeong So[1(✉)], and Kun Zhang[4]

[1] Department of Educational Technology, Ewha Womans University, Seoul, South Korea
{jinlingxi,hyojeongso}@ewha.ac.kr
[2] Department of Education, Sejong University, Seoul, South Korea
linbaicheng@sju.ac.kr
[3] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
mengze.hong@connect.polyu.hk
[4] School of Engineering Mathematics and Technology, University of the West of England, Bristol, UK
kun2.zhang@live.uwe.ac.uk

Abstract. This study presents Chat Melody, an LLM-based multimodal teachable agent designed to enhance music theory learning through the Learning-by-Teaching (LBT) approach. By simulating a novice learner, Chat Melody engages students in structured dialogue that prompts explanation, error identification, and conceptual refinement. The system integrates a multimodal interface combining visual notation, auditory feedback, and AI-generated explanations. LBT is operationalized through a structured prompting strategy based on the help-seeking model, guiding progressive reasoning instead of offering direct answers. In a quasi-experimental study with 28 students, *Chat Melody* was compared to traditional teacher-guided instruction. Results showed that students using the teachable agent demonstrated greater improvements in knowledge construction. Additionally, responses to the Situational Motivation Scale (SIMS) indicated higher intrinsic motivation and perceived value in music learning. These findings underscore the potential of LLM-based teachable agents to support meaningful learning and cognitive engagement in AI-assisted music education.

Keywords: Large Language Models (LLMs) · Teachable Agent · Learning by Teaching (LBT) · Knowledge Construction · Music Education

1 Introduction

Large Language Models (LLMs) are increasingly integrated into educational environments, offering dynamic learning experiences through contextualized dialogue, adaptive feedback, and personalized instruction [12, 19]. These systems have demonstrated potential in supporting self-directed learning [11], problem-solving [16], and knowledge construction [14]. However, empirical studies suggest that students often exhibit

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 148–155, 2025.
https://doi.org/10.1007/978-3-031-98462-4_19

passive reliance on AI-generated feedback, leading to shallow engagement and limited metacognitive development [6, 18, 19]. When AI functions primarily as an information provider, learners may uncritically accept responses, hindering higher-order thinking. Addressing this challenge requires the design of AI-enhanced learning environments that promote deeper cognitive engagement and meaningful knowledge construction. A promising pedagogical approach to foster such engagement is Learning-by-Teaching (LBT), wherein students deepen understanding by teaching others. Research in cognitive science and educational psychology has shown that teaching enhances conceptual clarity, encourages reflective thinking, and promotes metacognitive regulation [5, 11, 16]. Within this framework, Teachable Agents (TAs)—AI-driven virtual learners that simulate novice students—have proven effective in domains such as mathematics, language learning, and programming [2, 4, 11, 16, 20]. Unlike traditional rule-based tutors, TAs prompt students to articulate, justify, and revise their understanding, supporting active learning and metacognitive development [17]. Recent advances in LLM-powered conversational AI have significantly expanded the potential of teachable agents. Traditional TAs, constrained by predefined rules and rigid interaction structures, often struggle to support open-ended reasoning and critical inquiry, particularly in fields like music education that demand interpretative flexibility [15]. In contrast, LLM-driven TAs leverage generative models to facilitate adaptive, context-aware, and fluid dialogues [1, 9, 11, 16]. By engaging learners in iterative explanation and reflection, LLM-powered TAs enhance the effectiveness of LBT environments, offering a more personalized and cognitively demanding learning experience [4]. Despite their growing use in programming and mathematics education, the application of LLM-powered teachable agents in music education remains underexplored. Prior studies have shown that such agents can enhance learning outcomes and self-regulation. For example, [4] implemented a ChatGPT-based TA using Gall's help-seeking model [7], ensuring structured cognitive progression. Similarly, [11] demonstrated that concealing the agent's knowledge state prompted learners to actively elicit and refine concepts through dialogue. In [16], an LLM-based TA was successfully used to support middle school students in mathematical concept learning by encouraging explanation and self-monitoring.

Building on these findings, this study introduces *Chat Melody*, an LLM-based tutoring system that integrates the LBT approach to support music theory learning. By combining LLM-driven dialogue with structured pedagogical scaffolding, this study explores how teachable agents can enhance learners' motivation and promote durable knowledge construction in music education. The research questions of this study are as follows: **RQ1**: What impact does the teachable agent (*Chat Melody*) have on students' knowledge construction in music learning? **RQ2**: How does the teachable agent influence students' motivation in music learning?

2 The Design of an LLM-Based Teachable Agents

Our research team developed an LLM-powered teachable agent [13] to facilitate structured interactions in music learning through adaptive AI prompts (See Fig. 1). Grounded in the Learning-by-Teaching (LBT) approach, the system repositions AI from a static knowledge provider into a teachable agent that prompts learners to articulate, refine, and critically assess their understanding of music theory.

To ensure structured learning interactions, we employed a prompt-based design aligned with the five-stage help-seeking process model [7]: (1) awareness of the need for help, (2) decision to seek help, (3) identification of a potential help source, (4) strategies to elicit help, and (5) reactions to help-seeking attempts. Rather than providing direct answers, the agent simulates a novice music student to engage learners in structured reasoning, prompting them to explain concepts, justify interpretations, and refine understanding. For example, when analyzing the harmonic structure in Measure 5 of Mozart's Piano Sonata No. 5 in G major, K. 283 (Movement 2), the agent may ask: "I see E, G, C, and G in the right hand—what does this chord represent?" This initiates the first stage of help-seeking, prompting the learner to identify the notes as a C major chord (IV in G major). The agent then asks about the G in the bass, guiding the learner to recognize the chord as a cadential six-four (IV6/4) resolving to V (D major), and ultimately to I (G major). The conversation continues with prompts such as, "What happens in the next beat?", reinforcing the functional role of harmonic progression (IV6/4 → V → I) and deepening conceptual understanding. Through explanation and guided validation, the system transforms the LLM into a cognitive partner, supporting active knowledge construction.

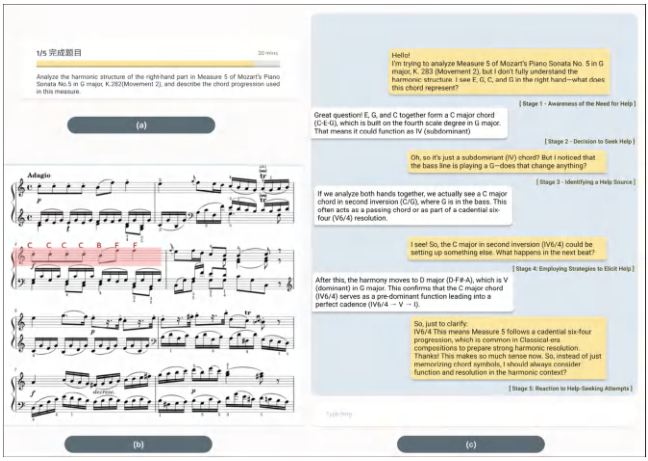


Fig. 1. Interface design of the LLM-powered Teachable Agent. The system comprises three key interactive components: (a) the problem statement area, which presents structured music analysis tasks; (b) the interactive music sheet display, which supports note-clicking and auditory feedback to connect notation with perception; (c) and the teachable agent, powered by GPT-4, which engages learners in structured dialogue as a novice learner.

Although LLMs such as GPT-4 occasionally produce inaccurate responses [10], we do not treat these errors as system flaws. Instead, following [4], we conceptualize them as learning opportunities, viewing these inaccuracies as cognitive triggers that mirror the reasoning struggles of novice learners. This perspective encourages reflection and supports knowledge construction. For instance, when the agent mislabels a harmonic function—such as misidentifying a cadential six-four as a tonic chord—the student must

diagnose the mistake, justify the correct interpretation, and explain their reasoning. This aligns with the principles of error-driven learning, in which understanding is reinforced through the detection and correction of conceptual errors. Unlike traditional AI tutors that prioritize correctness, this method promotes higher-order thinking by requiring students to critically evaluate AI-generated responses.

3 Method

A) **Participants.** A total of 28 Chinese university students (18 females, 10 males) participated in the study, ranging in age from 19 to 29 years ($M = 24.29$, $SD = 2.62$). All had prior instrumental learning experience and self-reported a foundational understanding of music theory. Since the length of instrumental training does not necessarily reflect theory proficiency [3, 20], participants completed a 15-min pre-test on notation, scales, rhythm, chords, and harmonic structures. The test, developed by an experienced music educator, confirmed no significant difference between the experimental group ($n = 16$) and the control group ($n = 16$) at baseline ($t = 0.86$, $p = .396$). During the study, four control group participants withdrew before completing the post-test, resulting in a final sample of 12 in that group. The experimental group received instruction via a teachable agent, while the control group received teacher-guided online instruction.

B) **Experiment Procedures and Learning Task.** The study was conducted entirely online and consisted of four phases: **Phase 1 (Pre-Test)***:* Before the intervention, all participants completed a pre-test assessing foundational music theory knowledge, including notation, scales, and harmonic structures. **Phase 2 (Instructional Phase):** All Participants then engaged in a two-day Zoom-based instructional phase. Day 1 focused on music notation and interpretation, covering note values, rhythm, articulation, scales, and key signatures, reinforced through guided exercises and listening activities. Day 2 addressed musical structures, particularly sonata form, using Mozart's Piano Sonata No. 5 in G Major, K. 283 (Movement 2) as the core material. Learners analyzed the exposition, development, and recapitulation, examining thematic development, modulation, and harmonic progression. **Phase 3 (Post-Test):** Following instruction, participants completed a 60-min analysis task requiring them to apply learned concepts to Mozart's piece. The experimental group interacted with Chat Melody, engaging in structured dialogue with the LLM-based teachable agent to explain, justify, and refine their analysis. The control group completed the same task independently, using only instructional materials and recorded lectures. All participants shared their screens via Zoom during the task, monitored by a researcher to prevent the use of external resources. Immediately after the analysis task, participants completed the Situational Motivation Scale (SIMS) to evaluate the impact of instructional methods on motivation. **Phase 4 (Delayed Test):** Two weeks later, a delayed test was administered to assess long-term knowledge retention.

C) **Music Theory Knowledge: Pre-test, Post- test, and Delayed test.** We adopted pre-tests, post-tests, and delayed tests to assess participants' progress in music theory knowledge. The pre-test primarily consisted of multiple-choice questions, whereas the post-test and delayed test included multiple-choice, short-answer, analysis, and

comprehensive questions. The format remained consistent across all assessments, with variations limited to question content, answer choices, and numerical values. Each test had a total score of 50 points, with raw scores converted into standardized Z-scores on a 100-point scale. The test was developed by a faculty member in music education and a doctoral student specializing in music education assessment, both of whom have extensive teaching experience. In this study, knowledge construction is understood as a process in which individuals progressively deepen their understanding of music theory through active cognitive engagement, transitioning from basic information acquisition to more advanced conceptual comprehension [3, 18]. The delayed test evaluates long term knowledge retention [14, 18]. By administering these tests, researchers were able to more accurately assess how learners acquire and retain music theory knowledge.

D) **Motivation Scale.** We used the Situational Motivation Scale (SIMS) proposed by [8], and adapted it to the context of music theory learning. This scale on the seven-point Likert scale evaluates four dimensions of motivation: (a) intrinsic motivation that reflects personal interest in learning music (e.g., "Because I think learning music theory is interesting"); (b) identified regulation that occurs when learners recognize the value of music education (e.g., "I am learning music because I believe it is important for me"); (c) external regulation that involves external pressures such as expectations or requirements (e.g., "I am learning music because I am supposed to"); and (d) amotivation that represents a lack of motivation or perceived relevance (e.g., "I don't see the point of learning music theory"). This scale has been shown to have high reliability and validity in previous studies [8].

4 Results

Q1: The Impact of Chat Melody on Knowledge Construction. To assess participants' progress in music theory knowledge, evaluations were con ducted at three distinct intervals: a pre-test (before instructional intervention), a post test (immediately after the intervention), and a delayed test (two weeks later). Two ANCOVAs were performed to examine variations across these time points, thereby evaluating learners' knowledge retention and construction. The first ANCOVA treated post-test scores as the dependent variable with pre-test scores as the covariate to determine differences in music theory understanding between the experimental group (EG) and control group (CG) following the instructional intervention. The second ANCOVA used delayed test scores as the dependent variable and post-test scores as the covariate to measure knowledge retention after a two-week period. Before conducting ANCOVA, Levene's test was performed to verify the homogeneity of variances, with the results presented in Table 1. Both post-test and delayed test *p*-values were above 0.05, ensuring the validity of the equal variance assumption and confirming ANCOVA's suitability for analysis. To verify group equivalence in prior music theory knowledge, an independent t-test was conducted on pre-test scores. The control group had a mean of 51.89 (SD = 8.33), and the experimental group had a mean of 48.58 (SD = 11.14). The result ($t = 0.86$, $p > .05$) indicated no significant difference between the groups, confirming comparable prior knowledge before instruction (see Table 2).

Table 1. The Levene's test results for knowledge construction.

Variable	Levene's test	
	F	*P*
Post-test for music theory knowledge	.59	.450
Delayed test for music theory knowledge	1.04	.317

Table 2. t-Test result of the pre-test scores.

		N	Mean	S.D	*t*	*p*
Pre-test	EG	16	48.58	11.14	.86	.369
	CG	12	51.89	8.33		

Table 3 presents the descriptive statistics and ANCOVA results for post-test and delayed test scores, highlighting the impact of *Chat Melody* on learners' knowledge retention and construction in music theory. The adjusted mean post-test scores were 59.65 (SE = 1.69) for the experimental group (EG) and 46.94 (SE = 2.00) for the control group (CG), with a significant difference ($F = 19.98$, $p < .001$), indicating greater improvement in the EG. A second ANCOVA on delayed test scores, using post-test scores as the covariate, showed adjusted means of 52.19 (SE =.93) for the EG and 45.83 (SE = 1.09) for the CG, also yielding a significant difference ($F = 17.65$, $p < .001$). These results suggest that learners supported by the teachable agent demonstrated stronger knowledge retention and construction than those who received teacher-guided instruction alone.

Table 3. Descriptive data and ANCOVA of the post-test results and delayed-test results.

Group		N	Mean	S.D	Adjusted Mean	S.E	*F*	*p*	η^2
Post-test	EG	16	57.30	6.72	59.65	1.69	19.98	<.001	.35
	CG	12	50.07	8.30	46.94	2.00			
Delayed-test	EG	16	55.25	6.88	52.19	.93	17.75	<.001	.11
	CG	12	41.75	9.55	45.83	1.09			

Q2: Influence of the Teachable Agent on Learners' Motivation. The results of the motivational scale data (see Table 4) indicate that the use of Chat Melody in music theory learning had a significant impact on learners' intrinsic motivation and identified regulation, while no significant differences were found in external regulation or amotivation between the experimental and control groups. Specifically, learners who engaged with the teachable agent reported significantly higher intrinsic motivation ($t = 2.55$, $p < .05$, Cohen's d =.97) and identified regulation ($t = 2.81$, $p < .01$, Cohen's d = 1.07) compared to those who received teacher-guided online instruction. In contrast, the findings

revealed no significant effect of Chat Melody on external regulation (t =.61, p =.546, Cohen's d =.23) or amotivation (t = −0.77, p =.446, Cohen's d = -0.30).

Table 4. Independent Sample t test for motivation.

Group		Mean	S.D.	S.E.	*T*	*p*	*Effect Size*
Intrinsic motivation	EG	25.56	3.67	.92	2.55	.017	.97
	CG	21.83	4.04	1.17			
Identified regulation	EG	25.38	3.24	.81	2.81	.009	1.07
	CG	21.25	4.54	1.31			
External regulation	EG	16.13	4.91	1.23	0.61	.546	.23
	CG	15.00	4.67	1.35			
Amotivation	EG	10.25	5.05	1.26	-0.77	.446	-0.30
	CG	11.75	5.12	1.48			

5 Discussion and Conclusion

This study investigated how an LLM-powered teachable agent, *Chat Melody*, supports music theory learning through the Learning-by-Teaching (LBT) approach. Results showed that learners using Chat Melody achieved significantly greater post-test and delayed-test gains than those receiving teacher-guided instruction, indicating improved knowledge acquisition and sustained knowledge construction (η^2 =.35; η^2 =.11). The agent also enhanced intrinsic motivation and identified regulation, suggesting that simulating a novice learner encouraged students to take ownership of learning and engage more deeply. No significant effects were found for external regulation or amotivation, highlighting the agent's specific impact on self-directed motivation. While limited by a small sample and lack of interaction data, the study offers evidence that LLM-based teachable agents can promote meaningful, self-motivated learning in music education. Future research should incorporate qualitative methods and broader populations to explore the cognitive and emotional dynamics of human-AI collaboration in learning contexts.

References

1. Baranwal, D.: A systematic review of exploring the potential of teachable agents in english learning. Pedagogical Res. **7**(1) (2022)
2. Biswas, G., Katzlberger, T., Bransford, J., Schwartz, D.: Extending intelligent learning environments with teachable agents to enhance learning. Artifi. Intell. Educ., 389–397 (2001)
3. Brochard, R., Dufour, A., Després, O.: Effect of musical expertise on visuospatial abilities: Evidence from reaction times and mental imagery. Brain Cogn. **54**(2), 103–109 (2004)

4. Chen, A., Wei, Y., Le, H., Zhang, Y.: Learning-by-teaching with ChatGPT: The effect of teachable ChatGPT agent on programming education (2024). http://arxiv.org/abs/2412.15226,
5. Duran, D.: Learning-by-teaching. evidence and implications as a pedagogical mechanism. Innovat. Educ. Teaching Inter. **54**(5), 476–484 (2017)
6. Fan, Y., et al.: Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. British J. Educ. Technol., 13544 (2024)
7. Gall, S.: Help-seeking: An understudied problem-solving skill in children. Dev. Rev. **1**(3), 224–246 (1981)
8. Guay, F., Vallerand, R.J., Blanchard, C.: On the assessment of situational intrinsic and extrinsic motivation: the situational motivation scale (SIMS). Motivation and Emotion (2001)
9. Hayashi, Y., Shimojo, S., Kawamura, T.: Scripted interventions versus reciprocal teaching in collaborative learning: a comparison of pedagogical and teachable agents using a cognitive architecture. Learn. Instr. **96**, 102057 (2025)
10. Hong, M., Zhang, C., Chen, C., Lian, R., Jiang, D.: Dialogue Language Model with Large Scale Persona Data Engineering (2024). https://arxiv.org/abs/2412.09034,
11. Jin, H., Lee, S., Shin, H., Kim, J.: Teach AI how to code: using large language models as teachable agents for programming education. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–28. ACM, Honolulu HI USA (2024)
12. Jin, L., Kim, K., So, H.J., Lee, G.: What do university students say about ChatGPT? a topic modeling of perception on GenAI in academic writing. In: Proceedings of the International Conference on Computers in Education. APSCE, Manila, Philippines (2024)
13. Jin, L., Lin, B., Hong, M., Zhang, K., So, H.J.: Exploring the impact of an LLM-Powered teachable agent on learning gains and cognitive load in music education. In: CHI 2025 Workshop on Augmented Educators and AI: Shaping the Future of Human and AI Cooperation in Learning, Yokohama, Japan, arXiv preprint arXiv:2504.00636 (2025)
14. Lee, H.-Y., Chen, P.-H., Wang, W.-S., Huang, Y.-M., Wu, T.-T.: Empowering ChatGPT with guidance mechanism in blended learning: effect of self-regulated learning, higher-order thinking skills, and knowledge construction. Inter. J. Educ. Technol. Higher Educ. **21**(1), 16 (2024)
15. Matsuda, N., Cohen, W.W., Koedinger, K.R.: Teaching the Teacher: Tutoring SimStudent Leads to More Effective Cognitive Tutor Authoring. Inter. J. Artifi. Intell. Educ. **25**, 1–34 (2015)
16. Song, Y., Kim, J., Liu, Z., Li, C., Xing, W.: Students' perceived roles, opportunities, and challenges of a generative AI-powered teachable agent: a case of middle school math class. J. Res. Technol. Educ. 1–19 (2025)
17. Stewart, L., Walsh, V., Frith, U.: Reading music modifies spatial mapping in pianists. Perception Psychophys. **66**(2), 183–195 (2004)
18. Wu, T., Lee, H., Li, P., Huang, C., Huang, Y.: Promoting self-regulation progress and knowledge construction in blended learning via ChatGPT-based learning aid. J. Educ. Comput. Res. **61**(8), 3–31 (2024)
19. Yan, L., Greiff, S., Teuber, Z., Gašević, D.: Promises and challenges of generative artificial intelligence for human learning. Nat. Hum. Behav. **8**(10), 1839–1850 (2024)
20. Zhao, G., Shen, Z.: Learning-by-Teaching: Designing teachable agents with intrinsic motivation. J. Educ. Technol. Soc. **15**(4), 62–74 (2012)



Searching for Transactivity in a Collaborative Chat: Analysis of Novelty and Reference with GenAI

David M. Otten(✉), Pantelis M. Papadopoulos, Maryam A. Haeri, and Maaike D. Endedijk

University of Twente, Drienerlolaan 5, 7522NB Enschede, The Netherlands
{d.m.otten,p.m.papadopoulos,m.amirhaeri,m.d.endedijk}@utwente.nl

Abstract. Transactivity, or building on the contribution of a learning partner, is an essential part of collaboration. While previous studies often emphasize the context of knowledge co-construction, less attention has been given to the analysis of transactivity independent of task. Recent research opted for perceiving transactivity as having two core elements: novelty and reference. In this study, we have constructed a model operationalizing both elements as scales. Subsequently, we explored the use of ChatGPT for classifying novelty and reference, achieving acceptable inter-rater reliability with human raters. Our dataset consisted of 21 collaborative dialogues of a Computational Thinking assignment in dyads. Results indicated that reference was more continuously present while novelty appeared in peaks. Transactivity, likewise, appeared mostly in isolated peaks. Regarding dyadic collaboration, novelty was also found to be more unevenly distributed than reference. This implied that novelty relied more frequently on one person. Our recommendation to instructional designers is to focus primarily on scaffolding for novelty, preferably tailored to individual participants.

Keyword: Transactivity · Collaborative Learning · GenAI. Classification

1 Introduction

Transactivity is generally defined as building on the contribution of a (learning) partner and is considered a crucial part of collaboration [1, 2]. Various perspectives and operationalizations have been used over the years, mostly encompassing features such as confirmations, elaborations, co-construction of knowledge, and the interlinking of ideas (e.g., [2–6). Although sequential analysis on the evolution of transactivity has been conducted, this often incorporated such features (e.g., [7]). The internal elements of transactivity have been studied less, independent of task or outcome [8].

In this study, we utilize the model by Vogel and Weinberger, who propose two core elements for transactivity: *novelty* and *reference* [8]. Novelty relates to introducing new information while reference means referring back to what a partner already contributed. We align with the perspective that analyzing both elements separately can inform how a dialogue is balanced or evolves. It was also suggested that both elements may vary over

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 156–163, 2025.
https://doi.org/10.1007/978-3-031-98462-4_20

time during collaboration [8]. Novelty might be prominent early in a dialogue where new ideas are introduced, while referring back to what has been discussed may be more common later in the dialogue [2, 8].

In addition, it was proposed that a better understanding of transactivity can simultaneously increase understanding of collaboration [5]. Peers tend to be unequal in various features, such as commitment, expertise, and general collaborative skills (e.g., [4, 6, 9–11]). It therefore needs to be assessed if transactivity is also dependent on a specific learner, or prevalent in all collaborators. More research on the interplay of novelty and reference, and their balance between participants, will provide instructional designers with more insights for adaptive support to foster a transactive dialogue.

For this study, we have constructed scales for novelty and reference. Using scales may provide better insights into how transactivity evolves and how the two elements interact. We have used these scales as a measure of transactivity, which we perceived as a function of novelty and reference. Furthermore, we have employed ChatGPT (using the GTP-4o model) for the classification of novelty and reference. In addition to contributing empirical data on the use of sub-scales for transactivity, this study also exemplifies the use of GenAI for classification tasks. We aim to answer the following research questions:

- RQ1: How does transactivity, as depicted by a function of novelty and reference, evolve in collaborative dialogues?
- RQ2: How are novelty, reference, and transactivity distributed within dyads in collaborative dialogues?

2 Method

2.1 Novelty and Reference Scales

Table 1 shows the scales of novelty and reference in our model. For novelty, introducing a new concept *and* linking it to previous ones is considered the highest level of novelty. Similarly, we distinguish between a simple reference, such as verbatim or via a pronoun (e.g., "*That* was a good idea"), and a reference that rephrases or elaborates on what has been said, suggesting that the discussants have a better grasp of the dialogue and can reference using their own vocabulary.

2.2 Data Preprocessing

The data was collected within the colMOOC project [12] in which bachelor students participated in an in-class course session on Computational Thinking. The first two hours consisted of a typical lecture while the third hour involved an online assignment in which randomly assigned dyads had to apply computational thinking and define an algorithmic decision process for choosing a study program. Although the project involved a conversational agent, the agent was not present during the sessions used in our data. Several steps of preprocessing were undertaken before conducting the analysis: 1) deleting messages consisting of streaks of emojis or exclamations; 2) concatenating consecutive chat messages into speaking turns; and 3) pruning starting/ending parts of the dialogue that were off-topic or nonsensical. This resulted in a new dataset based on speaking turns. The data consisted of 21 chat logs ($n = 21$) with the number of speaking turns ranging from 10 to 45 ($M = 26.57$; $SD = 8.79$).

Table 1. Definitions of the levels of novelty, reference, and transactivity.

Novelty Level	Definition
0 absent	No new concept is being mentioned in the message
1 low	A new concept is introduced, but is not linked to what has been discussed previously *OR* a link between previously discussed concepts is created or modified
2 high	A new concept is introduced *AND* directly linked to previously discussed concepts

Reference Level	Definition
0 absent	No reference to concepts discussed previously
1 low	A reference to previously discussed concepts is made verbatim or via a substitute word (e.g., pronoun, adjective, etc.)
2 high	A reference to previously discussed concepts is made by rephrasing or further elaboration

Transactivity Level	Definition
0 absent	Absent novelty or reference (i.e., 0–0, 0–1, 1–0)
1 low	Low novelty and reference (i.e., 1–1)
2 medium	High novelty and low reference *OR* vice versa (i.e., 2–1, 1–2)
3 high	High novelty and reference (2–2)

2.3 Data Analysis

Before conducting our analysis, we examined the outputs of ChatGPT (using GTP-4o) and two rater and examined the test-retest reliability of ChatGPT outputs. Specifically, two raters coded two dialogues. Weighted Cohen's kappa for novelty (κ =.69) and reference (κ =.66) was moderate and acceptable. We experimented with three different prompting techniques, zero-shot, chain-of-thought, and tree of thoughts, on these two dialogues. Comparing ChatGPT output with raters led to weighted Cohen's kappa's for novelty (κ =.54) and reference (κ =.57) which were also deemed acceptable. Test-retest reliability for novelty (κ =.85) and reference (κ =.87) was high. We prompted ChatGPT to include an explanation for each classification. This can help improve performance (e.g., [13, 14]). Also, it enhances opportunities for human assessment of LLM outputs, which remains needed for classification tasks (e.g., [15]).

ChatGPT outputs for novelty and reference were used to calculate the level of transactivity (Table 1). For RQ1, we analyzed the average distribution of the three variables per dialogue, calculated the correlation between novelty and reference, and examined transition states using Markov chains showing also novelty, reference, and transactivity in an example dialogue. For RQ2, we calculated the Coefficient of Variation (CV) between dyads for novelty, reference, and transactivity and assessed the correlations between novelty and reference.

3 Results

3.1 Distribution and Correlation

Figure 1 depicts the average distributions per dialogue for all the levels of novelty, reference, and transactivity. The figure indicates that *absent* was the most frequent level for novelty and transactivity, and *low* for reference. Spearman's correlation between novelty and reference ($\rho = .37$, $p < .01$) indicated a moderate relationship.

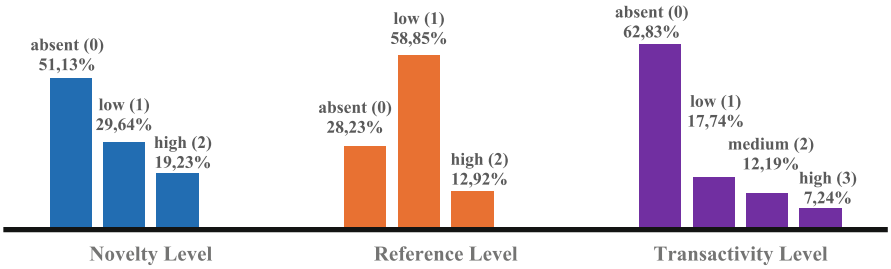


Fig. 1. Average levels of novelty, reference, and transactivity in the dialogues.

3.2 Transitions

We analyzed the transitions of novelty and reference between consecutive speaking turns. Figure 2 depicts a Markov chain diagram illustrating that staying in, or returning to, *absent* novelty and *low* reference have the highest probabilities. We calculated the steady-state probabilities of *absent* (.50), *low* (.33), and *high* novelty (.17), and of *absent* (.23), *low* (.65), and *high* reference (.12) Results show that *absent* novelty and *low* reference are the most likely levels for a dialogue to be in.

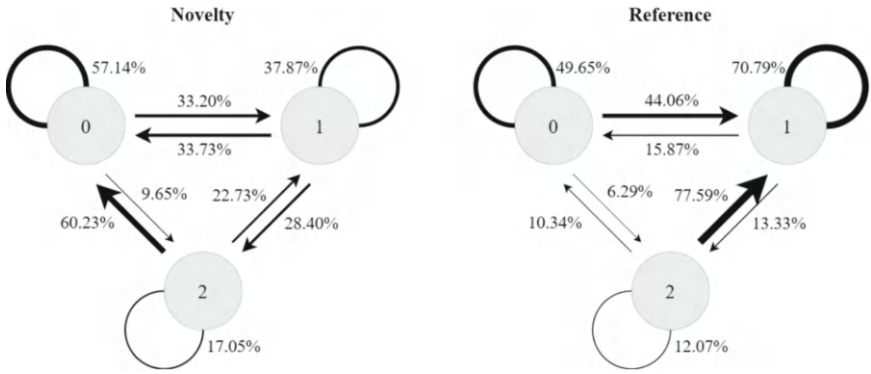


Fig. 2. Markov chain for transitions between levels of novelty and reference.

Figure 3a depicts the Markov chain for transitions between levels of transactivity. A transition to *absent* transactivity was the most common. Altogether, this underlines that our measures of transactivity mostly appear in isolated peaks. Figure 3b shows the evolution of novelty, reference, and transactivity during one example dialogue. Noteworthy are alternating occurrences of novelty and reference at the beginning, followed by multiple peaks of transactivity. Figure 3a and 3b underline that streaks of multiple consecutive transactive messages are rare.

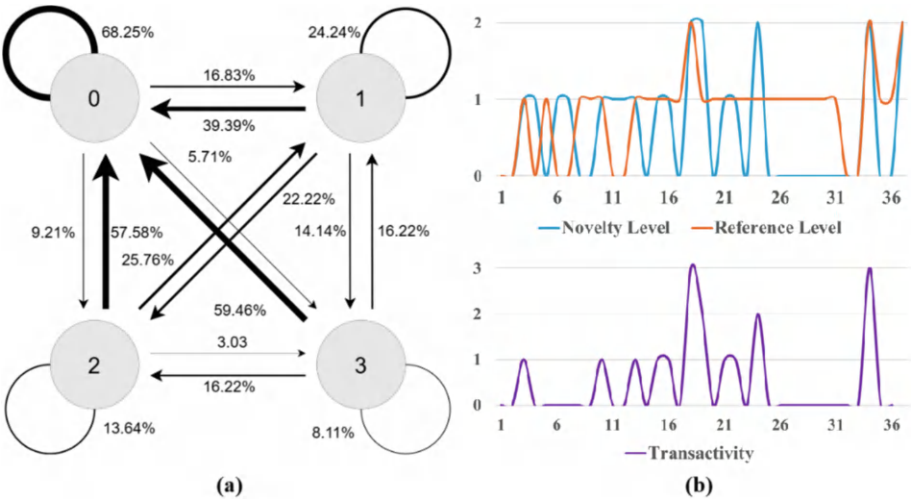


Fig. 3. a) Markov chain for transitions between levels of transactivity; b) evolution of novelty, reference, and transactivity during one of the dialogues.

3.3 Balance Within Dyads

To assess the balance within dyads, we calculated the Coefficient of Variation (*CV*) for novelty, reference, and transactivity for all 21 dyads (Table 2). Novelty has a higher mean CV than reference, indicating that novelty more frequently relies on one person. Also, the mean CVs for novelty and reference were unrelated (ρ = -.04, p =.84) indicating that imbalanced contributions in novelty and reference may occur independently.

Table 2. CV values of novelty, reference, and transactivity.

	M	*SD*	*Range*
Novelty (0–2)	0.31	0.34	0.00 – 1.10
Reference (0–2)	0.16	0.16	0.00 – 0.71
Transactivity (0–3)	0.45	0.38	1.0 – 1.41

4 Discussion and Conclusion

To answer "*How does transactivity, as depicted by a function of novelty and reference, evolve in collaborative dialogues*?" (RQ1), we found that *absent* novelty and *low* reference were the most frequently occurring levels. This means that a) novelty tends to be absent during large parts of dialogue and b) a low level of reference is constantly present throughout the dialogues. Also, transactivity mostly occurred in isolated peaks. We found that buildups or consecutive sequences were less prevalent. As for "*How are novelty, reference, and transactivity distributed within dyads in collaborative dialogues*?" (RQ2), we found higher mean CVs for novelty and transactivity within the dyads. Novelty was, thus, more likely than reference to rely on one person. Novelty and reference were moderately related, while their CVs were unrelated. Together, this indicates that novelty and reference have a separate contribution to transactivity. Lastly, transactivity was depicted as a function of novelty and reference. A causal relationship between the two elements and transactivity can therefore not be assessed.

Reference is needed for a dialogue to be interconnected. In other words, a continuous appearance of reference is expected to keep a dialogue coherent and on-topic. Furthermore, our measurement of the reference level of a turn considered all the previous turns. Participants consequently have increasingly more to refer to as a dialogue progresses. Novelty turned out to appear less and also more infrequent. It should be noted that the collaborative sessions were short and consisted of a concise task. This might have provoked discussants to engage in consensus-building relatively quickly, taking fewer opportunities for considering multiple new ideas. This builds on previous research suggesting students tend to agree or build on other's contributions, regardless if they perceive them as correct [7]. Furthermore, previous work assumed that participants in collaborative sessions can especially benefit from scaffolding novelty [8]. Our results corroborate this assumption by finding novelty to occur less than reference and to be more imbalanced within the dyads. Together, this hints that novelty might need the most support since striving for consensus may occur more naturally.

In conclusion, we present empirical evidence that our model, based on Vogel and Weinberger [8], may not be balanced. We found that reference is higher, not by chance, but by design. This questions whether both elements contribute to transactivity at the same level. Novelty mirrored the volatile character of transactivity more than reference and was also more unevenly distributed within the dyads. Instructional designers should give priority to scaffolding novelty in ways that are individually tailored, to foster a transactive dialogue. Last but not least, we found GenAI useful for the classification of novelty and reference. Short chats may be challenging since they provide an LLM with a limited context. We tackled this by prompting explanations, thereby enhancing accuracy and enabling human validation [14, 15]. Also, both novelty and reference are not common terms. This means LLM-based technologies might have difficulties since researchers also use differing definitions. Despite this, the reliability was already useful, opening potential for further refinement in future research.

Acknowledgments. This research was partly funded by (a) the Dutch Nationale Wetenschaps Organisatie (NWO) as part of the CLIC-IT project (File no. KICH1.HC01.20.006), and (b) the

Erasmus+ Programme of the European Union as part of the Learning to Be Green project (Grant Agreement no. 101139849).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Teasley, S.D.: Talking about reasoning: how important is the peer in peer collaboration?. In: Discourse, tools and reasoning Essays on situated cognition (pp. 361–384). Springer, Berlin (1997). https://doi.org/10.1007/978-3-662-03362-3_16
2. Weinberger, A., Fischer, F.: A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. Comput. Educ. **46**(1), 71–95 (2006). https://doi.org/10.1016/j.compedu.2005.04.003
3. Noroozi, O., Biemans, H.J., Weinberger, A., Mulder, M., Chizari, M.: Scripting for construction of a transactive memory system in multidisciplinary CSCL environments. Learn. Instr. **25**, 1–12 (2013). https://doi.org/10.1016/j.learninstruc.2012.10.002
4. Howley, I.K., Rosé, C.P.: Towards careful practices for automated linguistic analysis of group learning. J. Learn. Analy. **3**(3), 239–262 (2016). https://doi.org/10.18608/jla.2016.33.12
5. Gätje, O., Jurkowski, S.: When students interlink ideas in peer learning: linguistic characteristics of transactivity in argumentative discourse. Inter. J. Educ. Res. Open **2**, 100065 (2021). https://doi.org/10.1016/j.ijedro.2021.100065
6. Popov, V., van Leeuwen, A., Buis, S.C.: Are you with me or not? temporal synchronicity and transactivity during CSCL. J. Comput. Assist. Learn. **33**(5), 424–442 (2017). https://doi.org/10.1111/jcal.12185
7. Nemeth, L., Blumenfeld, T., Denn, A.K., Hirstein, A., Lipowsky, F.: An exploratory analysis of transactive interaction patterns in cooperative learning using sequential analysis. Educ. Sci. **13**(8), 790 (2023). https://doi.org/10.3390/educsci13080790
8. Vogel, F., et al.: Transactivity and knowledge co-construction in collaborative problem solving. In: Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning-CSCL 2023, pp. 337–346. International Society of the Learning Sciences (2023). https://doi.org/10.22318/cscl2023.646214
9. Pauli, R., Mohiyeddini, C., Bray, D., Michie, F., Street, B.: Individual differences in negative group work experiences in collaborative student learning. Educ. Psychol. **28**(1), 47–58 (2008). https://doi.org/10.1080/01443410701413746
10. Le, H., Janssen, J., Wubbels, T.: Collaborative learning practices: teacher and student perceived obstacles to effective student collaboration. Camb. J. Educ. **48**(1), 103–122 (2018). https://doi.org/10.1080/0305764X.2016.1259389
11. de Araujo, A., Martens, J., Papadopoulos, P.M.: Enhancing student dialogue productivity with learning analytics and fuzzy rules. In: International Conference on Artificial Intelligence in Education, pp. 397–404. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-64299-9_36
12. Demetriadis, S., et al.: Conversational agents in MOOCs: reflections on first outcomes of the colMOOC project. Intell. Syst. Learn. Data Analy. Online Educ. xxxvii-lxxiv. (2021). https://doi.org/10.1016/B978-0-12-823410-5.00001-2
13. Ashwin, J., Chhabra, A., Rao, V.: Using large language models for qualitative analysis can introduce serious bias. arXiv preprint arXiv:2309.17147. https://doi.org/10.1596/1813-9450-10597(2023).

14. Wu, X., et al.: Usable XAI: 10 strategies towards exploiting explainability in the LLM era. arXiv preprint arXiv:2403.08946. https://doi.org/10.48550/arXiv.2403.08946. (2024).
15. Barany, A., et al.: ChatGPT for education research: exploring the potential of large language models for qualitative codebook development. In International Conference on Artificial Intelligence in Education, pp. 134–149. Springer Nature Switzerland, Cham (2024).. https://doi.org/10.1007/978-3-031-64299-9_10



Risk or Promise: Educators' Perspectives on Student Use of ChatGPT

Tami Seifert(✉) and Taly Tzamir-Levi

Kibbutzim College of Education, Tel-Aviv, Israel
tami.seifert@smkb.ac.il

Abstract. Teachers' attitudes significantly impact technology integration in teaching and learning and are seen as the main barriers to successful integration in education. This study examines the attitudes of 211 school teachers and higher education lecturers towards using ChatGPT in Israel's education system, before official guidelines were established. The research explores educators' perspectives and future usage intentions. Significant differences emerged between school teachers and lecturers: Teachers had more positive views on "pedagogical benefit" and "learners' enjoyment" aspects, lecturers tended to support the application's "usage value". Women supported the need to restrict implementation more strongly than men, while men highlighted pedagogical benefits more than women. A significant positive correlation was found between positive implementation attitudes and future usage intentions. A significant negative correlation was found between the perceived usage risks and intended future use. This study contributes to understanding educators' perspectives on AI integration in education, highlighting the complex interplay of factors influencing technology adoption in teaching and learning.

Keywords: ChatGPT · Artificial Intelligence · Teachers' Attitude Toward Computers (TAC) · Technology Implementation

1 Introduction

ChatGPT is an innovative AI application developed by OpenAI. Its launch in November 2022 sparked a public debate about the significance and implications of AI use in education. The study's point of departure is that a close link exists between incorporating new technology into education and educators' attitudes toward technology. Teachers' attitudes are significant as they determine their collaboration and willingness to integrate technology into their teaching [1].

The present study examines the beliefs and perceptions of school teachers and higher-education lecturers about ChatGPT. To this end, it explores their stands about the application, their intention to use it in the future, and the connection between the two. It also aims to pinpoint differences between the stands of teachers and lecturers against their background variants, including gender, age, experience, and the extent and nature of their ChatGPT use. This contributes uniquely to the literature by examining distinctions between K-12 and higher education contexts [2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 164–170, 2025.
https://doi.org/10.1007/978-3-031-98462-4_21

At the focus of the study are the stands of "early adopters" [3], introduced to the application shortly after its launching before determining a clear policy for its future use. Studies emphasize that ChatGPT raises concerns about academic integrity [4] while also offering benefits like immediate feedback and personalized learning [5]. These paradoxical aspects highlight the need for responsible implementation [6] and the development of AI literacy among educators and students [7].

2 Methodology

2.1 Research Questions

Based on the identified literature gaps regarding educators' perceptions of ChatGPT and their potential impact on implementation decisions, this study addresses four critical questions:

1. Do the participants' background characteristics of gender and professional status affect the differences in their stands toward ChatGPT?
2. Do the participants' background characteristics (age, years of teaching experience, previous use of ChatGPT) affect their stands toward ChatGPT?
3. Do the participants' stands on ChatGPT predict their intentions to use it in the future?
4. Which stands predict future use intentions, and do any background variants also predict such intentions?

2.2 Research Population

The studied cohort consisted of 211 Israeli school teachers and higher education lecturers. 167 (79%) were women and 44 (21%) were men. Of the participants, 117 (55.5%) were school teachers (54 elementary school teachers (25.5 %) and 63 (30%) high school teachers. 94 (44.5%) were academia lecturers. Their average age was 48.5 (SD=9.8), and their average teaching experience was 14 years (SD=9.2). 193 (91%) were familiar with ChatGPT. 143 (73%) practiced it unguided, and 53 (27%) were assisted by a friend or another information source. On filling out the questionnaires, 187 (89%) used ChatGPT for various purposes, while 24 (11%) did not use it.

2.3 Research Procedure

The quantitative research was conducted in May 2023, about six months after ChatGPT was made public. Information was gathered through a voluntary, anonymous online questionnaire. A call to fill out the questionnaire appeared on various teacher and lecturer platforms and WhatsApp interest groups discussing ChatGPT in teaching and learning. It was also circulated among faculty members of Kibbutzim College of Education. The cohort is, therefore, a non-representative sample. Data analysis used SPSS, including prevalence tests, T-tests for independent samples, Pearson correlation, and linear regression.

2.4 Research Tools

The research questionnaire was based on an adapted version of the Teachers' Attitude Toward Computers questionnaire (TAC) [8], which has been extensively validated in previous studies. The questionnaire consisted of three parts: **Personal details,** including age, gender, status (teacher/lecturer), and teaching experience. **Acquaintance and use of ChatGPT**: Self-report questions about the extent of acquaintance with the application, current use, and intentions to use it in the future. **Stands regarding ChatGPT use**: Questions on six aspects defined as stand variables rated on a 1–5 Likert scale, 1 standing for "Not at all" and 5 – "To a great extent."

Figure 1 visualizes the nature and trends of the stands' rating. The rating is close in two main areas: Usage value and usage limitation (both ethical perspectives) and teachers' and students' enjoyment. The pedagogical benefit was rated lower than other aspects, while future damage risk is rated lowest.

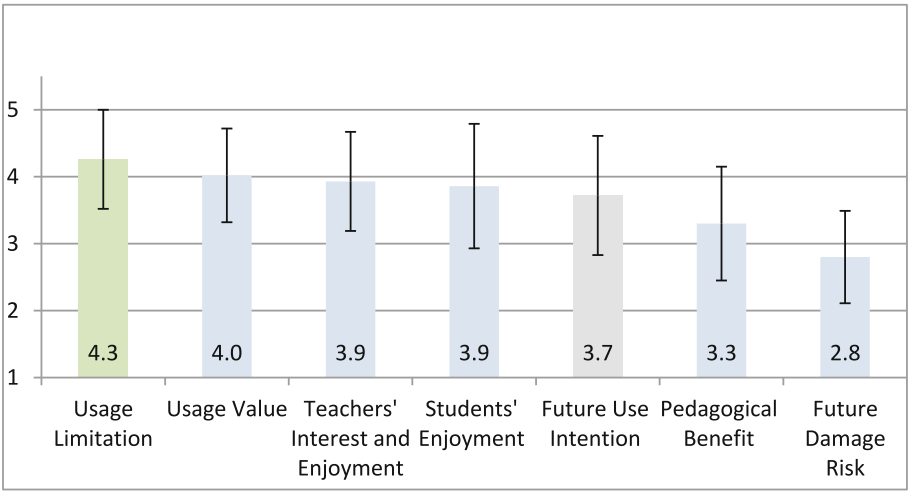


Fig. 1. Research participants' stands regarding ChatGPT use.

3 Findings

The first research question aimed to find out whether the participants' background characteristics, gender, and professional status affected their stands. A T-test for independent samples examined this point. Table 1 presents the gender and status (teachers vs. lecturers) average stand differences regarding ChatGPT.

Table 1 shows a significant relation ($p < 0.5$) between the gender variable and the stands about pedagogical benefit and usage limitation.

Men held more positive stands ($M = 3.52$) than women ($M = 3.29$) about ChatGPT use in the pedagogical benefit aspect ($t(76) = 1.69$, $p < 0.5$). Women supported usage limitation ($M = 4.34$) more than men ($M = 3.97$) ($t(56) = -2.87$, $p < .001$).

Table 1. Gender- and status-related differences in stands about using ChatGPT: Averages and SD.

Stands	Gender			Status		
	Woman (N = 167	Man (N = 44		Teacher (N = 117)	Lecturer (N = 94	
	M	M	t	M	M	t
Pedagogical benefit	3.29 (.87)	3.52 (.75)	1.69*	3.50 (.79)	(.88)	−3.06**
Usage value	3.97 (.77)	4.19 (.88)	1.54	3.93 (.82)	4.12 (.76)	1.68*
Usage limitation	4.34 (.68)	3.97 (.89)	−2.51**	4.27 (.73)	4.25 (.75)	−2.03
Enjoyment & interest: teachers	3.90 (.83)	4.05 (.76)	1.13	4.00 (.75)	3.84 (.83)	−1.41
Student enjoyment	3.87 (.92)	3.80 (1.00)	−.472	3.97 (.95)	3.72 (.89)	−1.89*
Usage risks	2.79 (.70)	2.79 (.64)	−.016	2.79 (.73)	2.78 (.65)	−.143

* $p < 0.5$ ** $p < 0.01$ *** $p < 0.001$

Regarding professional roles, there was a significant statistical difference ($p < .001$) between the stands of teachers and lecturers about the application's pedagogical benefit ($M = 3.50$ vs. $M = 3.14$ respectively) ($t(196) = -3.13$, $p < .001$). In contrast, lecturers attached higher value to the ethical value of the application's use ($M = 4.12$) than teachers ($M = 3.93$) ($t(207) = -.296$, $p < .01$). Teachers also perceived the application as enjoyable for the students ($M = 3.97$) more than lecturers ($M = 3.72$) ($t(204) = -1.89$,$p < 0.1$).

The second research question aimed to reveal whether the participants' background characteristics (age, teaching experience, extent of ChatGPT use) interconnect, affecting their stands about ChatGPT. To this end, we performed a Pearson correlations test that examined the connections between age, teaching experience, and stands about ChatGPT.

The findings revealed that age negatively correlated with perceived pedagogical benefit ($r(209) = -.153$, $p < .001$) and teachers' enjoyment ($r(209) = -.164$, $p < .05$). The extent of use positively correlated with usage value ($r(209) = .280$, $p < .001$) and enjoyment ($r(209) = 341$, $p < .001$), and negatively with perceived risk ($r(209) = -.186$, $p < .05$). No correlation emerged between teaching experience and stands. The third research question was whether the participants' stands about ChatGPT predicted its future use intentions. Correlation analysis showed that future use intention was positively associated with stands on pedagogical benefit ($r(209) = .675$, $p < .001$), usage value ($r(209) = .639$, $p < .001$), teachers' enjoyment and interest ($r(209) = .742$, $p < .001$), and students' enjoyment ($r(209) = .257$, $p < .001$). It negatively correlated with perception of future damage risk.

The fourth research question aimed at finding out which stands, and background variables predict future use of the application. To test the prediction model, we conducted a stepwise linear regression test, Table 2 presents the linear regression model results that found the model statistically significant.

Table 2. Stepwise linear regression predicting future use intention.

	Step 1		Step 2		Step 3		Step 4	
	B	β (t)	B	β (t)	B	β (t)	B	β (t)
Teachers' enjoyment and interest	.804	.742*** (16.00)	.615	.568*** (12.40)	.526	.486*** (10.25)	.525	485*** (10.35)
Usage value			.416	.373*** (4.94)	.290	.260*** (5.14)	.282	.252*** (5.05)
Pedagogical benefit					.255	.243*** (4.52)	.249	.238*** (4.48)
Gender							.207	.094** (2.45)
F	256		201		153		119	
R^2(Adjust R^2)	.55 (.54)		.66 (.65)		.69 (.68)		.70 (.69)	

* $p < 0.5$ ** $p < 0.01$ *** $p < 0.001$

Table 2 shows that the model is statistically significant and predicts 70% of future use intentions ($R2 = .70$). The β values indicate that the variables are sound predictors for the dependent variable "future use intention." In the present study, the variables explaining the variance in ChatGPT future use intention are, by descending order, teachers' enjoyment and interest, usage value, pedagogical benefit, and the gender background variable. The model's steps indicate that gender has the lowest influence.

4 Discussion

4.1 A Subsection Sample

This study examined educators' perspectives on ChatGPT in teaching and learning, considering six aspects: three positive (pedagogical value, usage value, enjoyment) and two negative (usage limitation, perceived risk). The findings revealed diverse opinions among participants, reflecting the complex nature of AI integration in education, which aligns with [6]'s (2023) observations on the advantages and disadvantages of ChatGPT in education.

The analysis revealed several significant findings. "Usage limitation" received the highest support, while "damage risk" was rated lowest, consistent with [4]'s (2024) call for balancing potential advantages with safeguarding academic integrity. Gender

differences emerged prominently in this study, with women favoring usage limitations more than men, who emphasized pedagogical benefits, supporting earlier findings on gender differences in technology adoption [9]. These gender-based differences likely reflect broader patterns in risk assessment and pedagogical priorities that should inform differentiated professional development approaches.

Teachers rated pedagogical benefits and student enjoyment higher than lecturers, echoing [5]'s (2022) observation of teachers being more receptive to AI integration. This finding challenges assumptions about higher education leading technology adoption and may reflect the different institutional pressures and assessment paradigms between K-12 and higher education. K-12 teachers may perceive greater flexibility in assessment design that accommodates AI tools, while university lecturers face more rigid academic integrity concerns. Age also played a role, as older participants found less pedagogical benefit and enjoyment in ChatGPT use, though recent studies suggest this digital gap may be narrowing [10]. The age-related findings align with technology acceptance models that emphasize perceived ease of use and usefulness as key adoption factors [11]. The negative correlation between age and perceived benefits suggests that professional development should be tailored to address age-specific concerns and learning needs.

The strong prediction model, explaining 70% of future use intentions, represents a significant contribution to understanding AI adoption in education. The primacy of teacher enjoyment and interest as predictors suggests that affective factors may outweigh rational assessments of utility in technology adoption decisions - a finding that extends previous technology acceptance frameworks [9, 11] and highlights the importance of creating positive initial experiences with AI tools.

5 Conclusions and Recommendations

This research contributes significant insights to understanding the complex attitudes of educators towards ChatGPT in educational settings. Our findings reveal that educators' perceptions are shaped by multiple intersecting factors including professional role, gender, age, and prior experience with the technology. The identified predictive relationship between attitudes and intended adoption (explaining 70% of variance) provides a powerful framework for developing targeted interventions to support AI integration in education.

Based on our findings, we recommend developing comprehensive training programs on responsible ChatGPT use [12] and establishing controlled practice environments for educators [13]. Additionally, institutions should implement clear policies that balance innovation with academic integrity [4]. These policies should move beyond binary "allow or prohibit" approaches to establish nuanced guidelines that support educators' professional judgment and epistemic agency in determining appropriate contexts for AI use. Educational leaders should recognize that educators' support for usage limitations coupled with low risk perception suggests receptiveness to structured implementation rather than restriction. Finally, promoting AI literacy among educators and students is crucial to foster critical thinking and evaluation skills addressing potential adoption barriers [14].

A limitation of this study is its non-representative sampling approach, which may have attracted participants with pre-existing interest in AI. Future research should employ

stratified random sampling to capture perspectives across the full spectrum of technology adopters. Additionally, qualitative investigations into the specific nature of educators' concerns would complement our quantitative findings and inform more targeted interventions.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Islahi, F.: Exploring teacher attitude towards information technology with a gender perspective. Contemp. Educ. Technol. **10**(1), 37–54 (2019)
2. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education–where are the educators? Int. J. Educ. Technol. High. Educ. **16**(1), 1–27 (2019)
3. Rogers, E.M.: Diffusion of innovations. Free Press, New York (2003)
4. Susnjak, T., McIntosh, T.R.: ChatGPT: The end of online exam integrity? Educ. Sci. **14**(6), 656 (2024)
5. Zhai, X.: ChatGPT: Artificial intelligence for education. supporting instructional decision making: the potential of an automatically scored three-dimensional assessment system. SSRN J., 1–18 (2022). https://doi.org/10.2139/ssrn.4312418
6. Mhlanga, D.: Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. In: FinTech and Artificial Intelligence for Sustainable Development: The Role of Smart Technologies in Achieving Development Goals, pp. 387–409. Springer, Cham (2023)
7. Long, D., Magerko, B.: What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2020)
8. Christensen, R.W., Knezek, G.A.: Construct validity for the teachers' attitudes toward computers questionnaire. J. Comput. Teach. Educ. **25**(4), 143–155 (2009)
9. Hung, Y.W., Hsu, Y.S.: Examining teachers' CBT use in the classroom: a study in secondary schools in Taiwan. J. Educ. Technol. Soc. **10**(3), 233–246 (2007)
10. Soomro, K.A., Kale, U., Curtis, R., Akcaoglu, M., Bernstein, M.: Digital divide among higher education faculty. Int. J. Educ. Technol. High. Educ. **17**, 1–16 (2020)
11. Teo, T.: Modeling technology acceptance among pre-service teachers. In: Technology Acceptance in Education, pp. 79–99. SensePublishers, Rotterdam (2011)
12. Zimmerman, M.: Teaching AI: Exploring new frontiers for learning. Int. Soc. Technol. Educ. (2018)
13. Tondeur, J., Van Braak, J., Ertmer, P.A., Ottenbreit-Leftwich, A.: Understanding the relationship between teachers' pedagogical beliefs and technology use in education: a systematic review of qualitative evidence. Educ. Technol. Res. Dev. **65**, 555–575 (2017)
14. Romero Alonso, R., Riquelme Plaza, I., Halal Orfali, C.: Barriers in teacher perception about the use of technology for evaluation in higher education. Digital Educ. Rev. **35**, 170–185 (2019). https://doi.org/10.1344/der.2019.35.170-185

Bridging the AI Adoption Gap: Designing an Interactive Pedagogical Agent for Higher Education Instructors

Si Chen[1(✉)], Reid Metoyer[2], Khiem Le[1], Adam Acunin[1], Izzy Molnar[1], Alex Ambrose[1], James Lang[1], Nitesh Chawla[1], and Ronald Metoyer[1]

[1] University of Notre Dame, Notre Dame, IN, USA
{schen34,kle3,aacunin2,imolnar,gambrose,jlang4,nchawla,rmetoyer}@nd.edu
[2] Columbia University, New York, NY, USA
ram2299@columbia.edu

Abstract. Despite their pivotal role in integrating AI into education, instructors' adoption of AI-powered tools remains inconsistent, with limited research on designing AI tools for broader adoption. This study employs a human-centered design and qualitative approach to investigate the design of interactive pedagogical agents that provide suggestions on instructors' questions. **First**, a formative study examined current pedagogical support strategies through interviews with five pedagogy experts. **Second**, a participatory design session engaged ten pedagogy experts in reviewing a storyboard illustrating chatbot features tailored to instructors with varying AI literacy and attitudes. Experts also evaluated the quality of LLM-generated suggestions based on instructors' common teaching challenges. Findings highlight the need for chatbot interactions that foster trust, particularly for AI-conservative instructors, by providing social transparency in peer usage and greater flexibility in how much or how little instructors engage with the system. Additionally, we propose design recommendations for enhancing AI-generated teaching suggestions, such as assessing prior teaching experiences. This work underscores the urgency of supporting AI-conservative instructors as AI literacy and attitudes intertwine, posing risks of pedagogical divides and diminished student learning opportunities if left unaddressed.

Keywords: AI Adoption · Pedagogical Support · Human-centered Design

1 Introduction

AI is reshaping education, but its success depends not only on technical advancement, but also on human-centered design that supports instructors' needs, expertise, and agency. While AI tools have been widely explored for student

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 171–178, 2025.
https://doi.org/10.1007/978-3-031-98462-4_22

learning—adaptive platforms, AI tutors, automated assessments [1,2]—less work has focused on empowering instructors in instructional design and pedagogy.

Instructor adoption of AI varies widely. Research shows that AI literacy—the ability to evaluate and use AI—and attitudes toward AI strongly influence adoption [3,4]. Confidence in using AI predicts adoption, while concerns over pedagogical autonomy, reliability, and ethics often foster skepticism [5]. Institutional support, such as faculty workshops, facilitates AI adoption [3,4]; without it, instructors risk disengagement, widening pedagogical divides [6], particularly in low-resourced contexts [7].

Advances in large language models (LLMs), such as GPT-4 [8], enable real-time, adaptive support [9,10], opening new possibilities for instructor-facing AI tools. LLMs have been applied for automated feedback [11] and content evaluation [10,12], yet their role in supporting instructor decision-making remains underexplored. Unlike traditional task-specific AI, LLMs allow conversational engagement, offering a pathway to scaffold gradual AI adoption.

Prior research highlights the need for iterative refinement [9,13], trust-building mechanisms [2], and participatory design with instructors [14,15] to align AI tools with pedagogical goals. However, most work has focused on student-facing tools or instructors already comfortable with AI.

This study explores how LLM-powered pedagogical agents—chatbots that generate teaching suggestions in response to instructors' questions—can support broader AI adoption. Here, "pedagogical agents" refer to screen-based learning partners [16], distinct from AI planning agents [17]. We investigate: *How can interactive pedagogical agents be designed to promote broader AI adoption?*

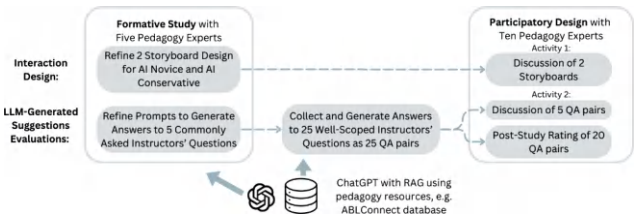


Fig. 1. Two-Phase Study Design with Pedagogy Experts

We conducted a two-phase study (Fig. 1). A formative study first explored pedagogy experts' experiences supporting AI adoption and providing teaching suggestions. Insights informed a participatory design phase, including storyboarding interaction flows and expert evaluation of LLM-generated suggestions. Our approach integrates Retrieval-Augmented Generation (RAG) [18] to ground suggestions in expert knowledge. Findings provide design principles to support AI-conservative instructors—those with low AI literacy and negative attitudes—through flexible, transparent, and scaffolded engagement.

2 Formative Study with Pedagogy Experts

We gathered insights from five pedagogy experts in instructional design and faculty development, affiliated with various AI-related education teams at the same institution. They shared perspectives on barriers to AI adoption, which were further explored through individual follow-ups. All experts consented to note-taking. Thematic analysis [19] of detailed notes, using open-ended coding, identified patterns in support strategies, adoption challenges, and AI tool expectations. Three key insights informed the next phase:

Current AI Tools Lack Support for Varied AI Literacy/Attitudes. LLM tools like ChatGPT don't address diverse instructor needs. While some instructors experiment confidently, AI-conservatives stick to traditional methods. Faculty workshops exist but require significant effort, limiting access. Experts emphasized the need for built-in scaffolding to support gradual AI adoption. *Insight 1:* We developed two storyboards to explore interactions for instructors with low AI literacy: one for AI-conservatives (negative attitude) and one for AI-novices (positive attitude), capturing a spectrum of user perspectives.

F2F Support is Effective but High-Effort. Consulting human experts or joining faculty groups offers strong pedagogical support but requires time many instructors lack. Some also avoid it due to perceptions of weakness. Still, F2F consultations follow a clear, effective process: identifying challenges, giving examples, and refining ideas. *Insight 2:* This F2F structure informed our chatbot interaction flow. The next phase's suggestion-generation module draws from ABLConnect [20], a curated database of active learning practices by Harvard's Derek Bok Center. *Insight 3:* Experts must validate LLM-generated suggestions. Experts were unsure if LLM outputs align with real teaching needs.

3 Participatory Design with Pedagogy Experts

We conducted participatory design sessions with ten pedagogy experts experienced in faculty development and instructional design. Participants held roles such as learning designers, education researchers, teaching tech specialists, and teaching professors, with experience ranging from two to twenty years. One participant in this phase also took part in the previous session. Three described themselves as AI-conservatives, while the others were positive or neutral.

3.1 Procedure

Sessions were conducted either remotely via Zoom or in person, each lasting approximately one hour. Each session began with an introduction to the chatbot's purpose, followed by two activities, as illustrated in the right box of Fig. 1. Activity 1 lasted 30 min, and Activity 2 took 15 min, followed by a post-study rating task that required an additional 20 min that participants completed on their own.

Activity 1: Storyboard Review. Storyboarding is a common technique in human-centered design for demonstrating system interfaces and contexts of use [21]. Participants reviewed two storyboards—one for AI-Novices in Fig. 2 and one for AI-conservative instructors in Fig. 3 —assessing user experience and feature usefulness. Their feedback included areas for refinement and how different instructors might engage with the tool.

Activity 2: LLM-Generated Suggestion Evaluations. After the storyboard review, experts evaluated LLM-generated teaching suggestions for relevance, accuracy, and pedagogical alignment. Experts compared outputs, selected preferred responses, and discussed refinements. We collected 25 well-scoped instructor questions from five pedagogy experts, addressing the tendency for overly broad queries (discussed in Findings). The first 5 QA pairs were immediately evaluated by all ten experts. The remaining 20 questions were used in a comparative study, where experts rated two responses per question—one from our prototype and one from default ChatGPT—yielding 40 total responses. Each response was rated on a 10-point scale across four dimensions: Relevance, Actionability, Alignment with effective teaching, and Reliability.

Storyboard Design Details. We designed two storyboards: one for instructors with low AI literacy (AI-Novices) and another for AI-Conservatives, who also hold negative attitudes. Informed by formative study insights, the AI-Novice design emphasized simplicity and gradual feature introduction, while the AI-Conservative design focused on transparency, offering detailed explanations, multiple options, and alignment with institutional values. Both were refined through multiple expert feedback rounds to mirror real-world interaction flows.

Suggestion Generation Details. The chatbot aims to help instructors leverage teaching and learning expertise by offering actionable suggestions based on their goals. We conducted six weekly sessions with two pedagogy experts to refine output generation. Initial trials with ChatGPT, using simple instructions on five expert-pooled questions, revealed vague, template-like responses. To improve



Fig. 2. AI-Novice storyboard illustrating instructor-AI interaction.



Fig. 3. AI-Conservative storyboard illustrating instructor-AI interaction.

specificity, we applied In-Context Learning [22], providing expert-written examples. Evaluation showed the need for a standardized example database to ensure grounding in expert knowledge. We constructed a database of 115 active learning examples from ABLConnect [20], after filtering irrelevant entries. When a user submits a question, the system retrieves the three most relevant examples, following the concept of RAG [18], to guide GPT-4's output and reduce hallucination risks. Finally, a carefully designed instruction consisting of retrieved examples is constructed for feeding into the model, i.e., GPT-4, to generate suggestions.

Analysis. Audio-recorded storyboard and QA discussions were transcribed and analyzed using thematic analysis [19]. Two researchers with higher education teaching experience independently coded transcripts, resolving discrepancies through discussion; a third researcher contributed to theme discussions only. Analysis began after the fourth session and continued iteratively until thematic saturation, with ten participants in total. Post-study ratings of 20 QA pairs were averaged and are presented in Fig. 4. Sample quotes for each theme are in the appendix.

3.2 Findings 1 Designing Interactions for AI-Conservatives

Easing AI Engagement Through Incremental Input. AI-Conservatives often feel overwhelmed by upfront information demands. Experts recommended a flexible, step-by-step input process, letting users decide how much to share. When input is minimal, the chatbot can offer general suggestions and invite users to refine. This mirrors expert-led consultations, which avoid requiring detailed forms.

Scaffolding Adoption with Low-Effort, Progressive Features. Hesitancy often stems from deeper fears about AI disrupting teaching norms. Experts advised showcasing AI's range of capabilities in ways that scale with instructor comfort. For example, start with lightweight tips (e.g., a 5-min activity), then offer deeper options (e.g., slide generation). Gradual exposure, as shown in storyboard panels 2–5 for AI-Novice, helps build comfort over time.

Building Trust via Peer Validation. AI-Conservatives are more open when seeing strategies used by peers. The chatbot should offer testimonials and discipline-specific examples, not generic advice. For instance, instead of "Flipped classrooms work," it could say, "A colleague in your department uses flipped learning successfully. Want to see how?" Peer-supported entry points are more realistic than starting solo (see storyboard panel 1 for AI-Conservatives).

Fostering Collaboration Through Conversational Design. Experts stressed that AI-Conservatives prefer collaborative over prescriptive dialogue. The chatbot should offer options with a suggestive tone ("Some instructors use peer learning—interested?"), allow adjustments, and explain reasoning ("Why this suggestion?"). Options to view alternatives or flag responses (as in storyboard panels 3–4 for AI-Conservatives) support deeper, trust-building engagement.

Addressing Data Privacy with Transparency. Privacy concerns around student data and materials are a key barrier. Proactively, the chatbot should state clear policies ("Your info isn't stored or shared") and allow anonymous input. Reactively, it should explain how suggestions were generated and offer options to delete learned data (see storyboard panel 6 for AI-Conservatives).

3.3 Findings 2 Expert Evaluation of LLM-Generated Suggestions

Figure 4 presents the averaged rating scores of ten experts on the generated suggestions from our prototype compared to ChatGPT across 20 questions (Rated blindly, with the question sequence randomized, on 1–10 scale). Experts consistently prefer our prototype's suggestions over ChatGPT's, rating them about one point higher across all dimensions. Although demonstrated potential, the prototype still needs improvement.

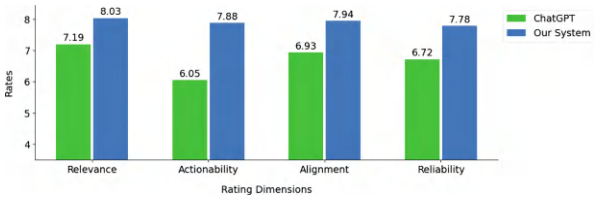


Fig. 4. Averaged expert ratings for suggestions from our prototype and ChatGPT

Ways to Improve Suggestion Generation Process for AI-Conservative Instructors. Instructors with higher AI literacy can quickly refine prompts and adapt "good enough" responses into practical strategies, aided by their better understanding of LLM capabilities. In contrast, AI-Conservatives—characterized by lower literacy and negative attitudes—often feel overwhelmed by the uncertain and extensive input needed for useful responses, leading to early frustration and disengagement. Instead of requiring all details upfront, the chatbot should guide instructors through a multi-turn scaffolding process that refines queries, incorporates context, and draws on teaching experience. Tailoring suggestions to discipline-specific practices—especially in STEM—and considering class size, course format, student level, and course type can further improve relevance. Experts also observed that some AI outputs merged unrelated suggestions without clear rationale; while blending ideas is not inherently problematic, starting with one evidence-based, goal-oriented suggestion is preferred.

4 Conclusion

This study explored the design of an LLM-powered interactive pedagogical agent to generate teaching suggestions for instructors with limited AI literacy. Find-

ings highlight the importance of social transparency and incremental information collection in instructor-AI interaction design to foster AI adoption, particularly among AI-conservative instructors. Through a human-centered approach, this work demonstrates how pedagogical agents can help bridge gaps in faculty development. Future research should investigate how human-in-the-loop AI tools can further refine interaction design and optimize LLM-generated suggestions to better support a wide range of instructors. Link to Appendix.

References

1. Cukurova, M., Miao, X., Brooker, R.: Adoption of artificial intelligence in schools: unveiling factors influencing teachers' engagement. In: International Conference on Artificial Intelligence in Education, pp. 151–163. Springer (2023)
2. Smolansky, A., Cram, A., Raduescu, C., Zeivots, S., Huber, E., Kizilcec, R.F.: Educator and student perspectives on the impact of generative AI on assessments in higher education. In: Proceedings of the Tenth ACM Conference on Learning@ Scale, pp. 378–382. ACM (2023)
3. Nourani, M., Hashky, A., Ragan, E.D.: User profiling in human-AI design: an empirical case study of anchoring bias, individual differences, and AI attitudes. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 12, pp. 137–146. AAAI (2024)
4. Ayanwale, M.A., Sanusi, I.T., Adelana, O.P., Aruleba, K.D., Oyelere, S.S.: Teachers' readiness and intention to teach artificial intelligence in schools. In: Computers and Education: Artificial Intelligence, vol. 3, pp. 100099. Elsevier (2022)
5. Long, D., Magerko, B.: What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–16. ACM (2020)
6. Wang, C., Boerman, S.C., Kroon, A.C., Möller, J., de Vreese, C.H.: The Artificial Intelligence Divide: Who is the Most Vulnerable? New Media & Society, p. 14614448241232344. SAGE Publications, London (2024)
7. Yadav, M., Singh, S.K., Sharma, P., Hung, T.H.: AI and lifelong learning in developing countries: investigate the potential of AI to support lifelong learning and development in low-resource contexts. In: Impacts of Generative AI on the Future of Research and Education, pp. 49–72. IGI Global (2025)
8. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
9. Kim, T.S., Lee, Y., Shin, J., Kim, Y.-H., Kim, J.: Evallm: interactive evaluation of large language model prompts on user-defined criteria. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–21. ACM (2024)
10. Lu, X., Wang, X.: Generative students: using LLM-simulated student profiles to support question item evaluation. In: Proceedings of the Eleventh ACM Conference on Learning@ Scale, pp. 16–27. ACM (2024)
11. Yan, L., et al.: Practical and ethical challenges of large language models in education: a systematic scoping review. Br. J. Educ. Technol. **55**(1), 90–112 (2024)
12. Moore, S., Nguyen, H.A., Bier, N., Domadia, T., Stamper, J.: Assessing the quality of student-generated short answer questions using GPT-3. In: European Conference on Technology Enhanced Learning, pp. 243–257. Springer (2022)
13. Arawjo, I., Swoopes, C., Vaithilingam, P., Wattenberg, M., Glassman, E.L.: ChainForge: a visual toolkit for prompt engineering and LLM hypothesis testing. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–18 (2024)

14. Lin, P., Van Brummelen, J.: Engaging teachers to co-design integrated AI curriculum for K-12 classrooms. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2021)
15. Holstein, K., McLaren, B.M., Aleven, V.: Co-designing a real-time classroom orchestration tool to support teacher-AI complementarity. Grantee Submission, ERIC (2019)
16. Clarebout, G., Heidig (née Domagk), S.: Pedagogical agents. In: Seel, N.M. (ed.) Encyclopedia of the Sciences of Learning. Springer, Boston (2012). https://doi.org/10.1007/978-1-4419-1428-6_942
17. Wang, L., et al.: A survey on large language model based autonomous agents. Front. Comput. Sci. **18**(6), 186345 (2024)
18. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474 (2020)
19. Braun, V., Clarke, V.: Thematic analysis. In: Encyclopedia of Quality of Life and Well-Being Research, pp. 7187–7193. Springer (2024)
20. ABLConnect. ABLConnect: Active Learning at Harvard. Harvard University. https://ablconnect.harvard.edu/. Accessed 17 Feb 2025
21. Truong, K.N., Hayes, G.R., Abowd, G.D.: Storyboarding: an empirical determination of best practices and effective guidelines. In: Proceedings of the 6th Conference on Designing Interactive Systems, pp. 12–21. ACM (2006)
22. Brown, T., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)

LipMOOC: A Modular Synthetic MOOC for Lipreading

Snigdha Agarwal[1(✉)], Vinay Namboodiri[2], and C. V. Jawahar[1]

[1] International Institute of Information Technology Hyderabad, Hyderabad 500032, Telangana, India
snigdha.a@research.iiit.ac.in, jawahar@iiit.ac.in
[2] University of Bath, Claverton Down, Bath BA2 7AY, UK
vpn22@bath.ac.uk

Abstract. For people with hearing impairments, lipreading is a crucial skill, as it bridges the gap between spoken language and understanding, improving the ability to communicate effectively. However, the creation of lipreading resources is pervasive and labor-intensive, leading to a scarcity of high-quality, structured, and widely accessible educational materials. To address this, we developed LipMOOC, a Massive Open Online Course (MOOC) tailored for lipreading education. Our approach leverages synthetically generated content, reducing the dependence on human-recorded videos and building an accessible and effective learning option. Using the state-of-the-art models, we created a comprehensive lipreading course and conducted a study to evaluate its effectiveness. We statistically proved that (1) Synthetic videos can replace human-recorded videos without affecting the learning outcomes, and (2) Learner performance on a structured course, i.e., LipMOOC, is much better than a standard course. A total of 120 participants with hearing impairment and no prior lipreading experience volunteered. These results highlight the potential of AI-driven content generation and scientific instructional design.

Keywords: MOOC · AI Teaching · Inclusion · Learning Pedagogy

1 Introduction

According to the World Health Organization (WHO) [9], more than 430 million people worldwide experience hearing loss, a number expected to rise to over 700 million by 2050. Despite its importance, lipreading education still faces significant challenges. There are platforms like *lipreading.org*[1] and *lipreadingpractice*[2] which provide basic online resources for lipreading. However, most available learning materials require a fee, are unstructured, or do not include the diversity of speech patterns, accents, and ethnic representations required for thorough

[1] https://www.lipreading.org/.
[2] https://lipreadingpractice.co.uk/.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 179–186, 2025.
https://doi.org/10.1007/978-3-031-98462-4_23

learning. To overcome these constraints, we created an AIED system that can offer structured, scalable, and accessible learning opportunities. Moreover, such systems can mass-produce content that can be made widely available, ensuring equal accessibility to individuals from different socio-economic backgrounds.

To leverage benefits of AI-driven technology for societal benefit, we made the following contributions through this work:

- A beginner-friendly and self-paced lipreading course (LipMOOC) with structured design & high-quality content, to help improve learning outcomes.
- We evaluated LipMOOC with 120 participants, showing that its varied question type and modular design significantly improve learning outcomes.

We use statistical analysis to prove two key hypotheses.

(H1) Lipreading performance remains consistent whether users view real or synthetic videos generated through our pipeline.
(H2) A structured course design, as implemented in LipMOOC, leads to better learning outcomes compared to traditional lip-reading courses.

2 Related Work

Lipreading education has been a longstanding area of research, with various methodologies explored to enhance learning effectiveness such as, video-based lessons [6], and systematic lipreading training [4,13]. AI-driven solutions like, deep learning-based assistive technology to improve audio-visual speech recognition [8], voice-based smart devices [2], generating real-time synthetic talking faces [10] have been proposed.

Agarwal et al., 2023 [1] introduced the use of synthetic talking heads for lipreading MOOCs, but their work primarily serves as a proof-of-concept rather than a structured educational framework. Furthermore, our work extends beyond proof-of-concept by conducting empirical assessments to validate the effectiveness of structured AI-driven lipreading training.

3 Methods and Data

3.1 LipMOOC

We developed a MOOC to teach lipreading by initiating the learning process at the phoneme level. This foundational approach was particularly chosen to ensure that learners focus on the smallest units of speech, making the task of lipreading more manageable and less overwhelming. Thus, learners can develop a strong understanding of basic visual patterns of speech, which serves as a critical building block for more complex lipreading tasks. The course content, featuring videos with no audio, focuses on different phonemes and how to spot and articulate them along with examples. The question types are - Single Word (SW), Phrase Level (PL), Fill in the Blanks (FIB), and Sentence Level (SL) items

The course is divided into three distinct levels: Beginner, Intermediate, and Expert. Each level contain same ten lessons with increasing question difficulty to ensure progressive skill development. This structured approach is designed to ensure a clear learning path and more personalized pacing. For every lesson covered, there are 4 practice questions of each type $10 \times 4 = 40$ at each level. The course has a total of 120 practice questions ($40 \times 3 = 120$). This approach ensured that learners remained engaged and motivated, avoiding potential frustration or discouragement by offering progressively more demanding tasks. This was hosted on a website[3] and supported various devices PC, laptop, mobile, etc.

3.2 Lipreading Data Generation

This pipeline (1) Uses face images from Synthetic Faces High Quality (SFHQ) [3], encompassing a wide range of ethnicities, pose, etc. (2) Processes the images in which lips are fully visible and filter out the rest (side view angles). (3) Generates a synthetic speech utterance from the provided text, adding buffers at the start and end. (4) Selects the face and the utterance and gives it as input to a talking head-generation model, EchoMimic [5], that generates synthetic talking heads using audio and a reference face, keeping the background intact.

Text-to-Speech System. For our experiments we used Bark[4] to generate speech. The TTS model used in this pipeline is a configurable plug-and-play module which can be replaced by any TTS of choice.

Synthetic Talking Head Videos. We tested a lot of talking head generation models [5,11,12,14–17] and consequently, we adopt EchoMimic [5] in our pipeline.

The images from SFHQ [3] are used to ensure that we do not use real individual faces without permission as we will use it as content for our LipMOOC on lipreading. This approach helps maintain ethical standards and ensures that we comply with intellectual property laws. Images with face and lips clearly visible with a front view angle are then filtered from the dataset. This image is passed as a reference image along with the generated audio from Bark to our pipeline. Thus, creating realistic-looking synthetic talking faces.

3.3 Test Design

Experiment 1. For our first experiment, the primary goal was to test **(H1)**. For this study 50 (22 female and 28 male) participants with varying degrees of hearing loss (40 db to >90 db) volunteered. The average age of participants was 25 years, ranging from 13 to 35 years. Participants included eighth-grade students to Master's degree students. This validation seeks to establish that

[3] https://ilocr.iiit.ac.in/lipreading.
[4] https://github.com/suno-ai/bark.

synthetic videos maintain a comparable level of clarity and effectiveness, ensuring their utility in lipreading tasks without compromising the accuracy of visual information. Each participant participates in two quizzes corresponding to two datasets made for (i) Real and (ii) Synthetic talking head videos. In total every user attempted 2 quizzes hosted on Google Forms, and participants reported taking the quiz from various devices PC, laptop, mobile, etc.

Data Analysis Plan. To validate **(H1)** the difference in the mean score should be statistically insignificant. For that evidence of the null hypothesis is required, just the absence of evidence is not enough to support the hypothesis **(H1)**. For that we can conduct a Bayesian Equivalence Analysis using the **Bayesian Estimation Supersedes the t-test (BEST)** [7] and **95% HDI** to help validate potential of our pipeline to develop video content for a large-scale lipreading platform.

Experiment 2. Our second experiment aimed to validate **H2**. Our study used three different lipreading training courses, each designed to investigate specific aspects of lipreading instruction: (1) Standard Course (Lipreading.org) (SC): We utilized the established "Lipreading with Alphabet" course from Lipreading.org as our control condition. (2) Non-Modular Synthetic Course (NMS): The content for the LipMOOC we developed was presented in a nonmodular structure. (3) Modular Synthetic Course (MSC) i.e. LipMOOC. All three courses focused on teaching lipreading through a phonetic structure approach. To ensure comparability, we maintained consistency in the core lesson content across all courses, with variations primarily in the practice pattern (NMS) and course structure (MSC) to further assess the effects of these changes in learning. A total of 120 (52 female and 68 male) participants with moderate hearing loss (40–60 dB) to profound hearing loss (>90 dB) volunteered in this study. None of the participants had any previous experience in lipreading. All of them came from a low-income background. The study was conducted among sixth to eighth-grade school students following a special educational curriculum made for students with special needs. The age range of participants was 12 to 17 years. SC was conducted on the lipreading.org website, NMS was hosted on articulate 360 platform, and for MSC a course website[5] was created. All these courses supported various devices PC, laptop, mobile, etc. Implementation characteristics for Experiment 2 are shown in Table 1.

Data Analysis Plan. Every participant took the same pre-test, which was used as the deciding factor to assign participants to different groups. This would avoid giving any unfair advantage to one group and prevent any imbalance as each group starts with a similar skill level. By ensuring homogeneous groups, we reduce variance and increase the ability to detect meaningful effects. To validate **(H2)**, we performed two-way mixed ANOVA to determine the effect, one for

[5] https://ilocr.iiit.ac.in/lipreading.

Table 1. Summary of implementation characteristics across courses. In Question type *Varied* refers to all SW, PL, FIB, and SL question type and *Single* refers to only SW question type.

	Standard Course (SC)	Non-Modular Synthetic Course (NMS)	Modular Synthetic Course (MSC)
Video Content	Human recorded	Synthetic	Synthetic
Duration	18 days	18 days	18 days
No. of students	38	42	40
Question Type	*Single*	*Varied*	*Varied*
Structure Type	Non Modular	Non Modular	Modular
No. of Questions	110	120	120

Question type (Single Vs. Varied) and another for Structure type (Non-Modular vs Modular).

4 Results

To validate this **(H1)**, we report the BEST statistics on Real and Synthetic videos for all four types of lipreading question patterns in Table 2. We also report the t-statistic and p-value using the standard two-tailed t-test. From Table 2, it is clear that the BEST statistic lies within the acceptable 95% HDI for all four question types indicating that the difference in the scores between the two groups is statistically insignificant. This suggests that our pipeline is a viable alternative to the existing manually curated talking-head videos.

Table 2. We perform Bayesian Equivalence Analysis using BEST to calculate the 95% HDI range. The p-value and t-statistic were calculated from a standard t-test for comparison.

Measure	Mean	MGD	P-value	T-statistic	95% HDI
SW	0.9333	0.941	0.1019	1.6621	[−0.1488, 2.0155]
FIB	0.6000	0.624	0.2651	1.1253	[−0.4275, 1.6275]
PL	0.6667	0.701	0.2399	1.1874	[−0.4153, 1.7487]
SL	0.5000	0.513	0.2876	1.0732	[−0.3978, 1.3978]

Table 3 shows descriptive statistics from Experiment 2, suggesting both question and structure types positively affect learning outcomes. To rigorously test (H2), we conducted two-way mixed ANOVAs.

Table 3. Experiment 2 descriptive statistics with averages for pre-test and post-test scores.

Group	Course	Questions Type	Structure Type	Pre-test	Post-test
Group 1	SC	Simple	Non-Modular	7.7	9
Group 2	NMS	Varied	Non-Modular	8.025	11.35
Group 3	MSC	Varied	Modular	7.8	14.58

Question Type. A two-way mixed ANOVA (Table 4) between Groups 1 and 2 examined the impact of Question Type on post-test scores. Results showed a highly significant main effect, $F(1, 78) = 143.37, p = 2.39 \times 10^{-19}$, with large effect size ($\eta_p^2 = 0.648$), indicating that varied questions lead to greater gains. A strong interaction effect ($F(1, 78) = 134.95, p = 1.09 \times 10^{-18}, \eta_p^2 = 0.634$) showed improvement over time depended on question type.

Table 4. Two way mixed ANOVA Analysis of Question Type and test score between Group 1 and 2.

Source of Variation	DF	F-Statistic)	p-unc	np2
Between-subjects *questions type*	78	143.37	2.39e-19	0.648
Within-subjects *time*	78	17.82	6.50e-05	0.186
Interaction *questions type:time*	78	134.95	1.09e-18	0.634

Structure Type. A separate ANOVA (Table 5) between Groups 2 and 3 assessed the impact of Structure Type. There was a significant main effect, $F(1, 78) = 95.59, p = 3.39 \times 10^{-15}$, with large effect size ($\eta_p^2 = 0.551$), showing structure influences learning. The interaction effect ($F(1, 78) = 102.41, p = 7.43 \times 10^{-16}, \eta_p^2 = 0.568$) revealed that modular formats enhance learning more effectively over time. Thus, validating **(H2)** that a synthetically generated lipreading course with varied question types and a modular structure yields superior learning outcomes compared to a standard lipreading course.

Table 5. Two way mixed ANOVA Analysis of Structure Type and test score between Group 2 and 3.

Source of Variation	DF	F-Statistic	p-unc	np2
Between-subjects *structure type*	78	95.59	3.39e-15	0.551
Within-subjects *time*	78	760.84	5.54e-42	0.907
Interaction *structure type:time*	78	102.41	7.43e-16	0.568

5 Conclusion

This work introduces LipMOOC, a scalable AIED system designed to teach lipreading through a structured and self-paced approach. We demonstrate that lipreading can be effectively taught with AI-generated visual speech, making the learning process more accessible. Our empirical evaluation with 120 participants confirms that structured lipreading training leads to improved learning outcomes and that synthetic talking faces can serve as a viable alternative to real-world speakers. LipMOOC has the potential to bridge accessibility gaps and transform the way individuals with hearing impairments develop and refine their communication skills.

References

1. Agarwal, A., Sen, B., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: Towards MOOCs for lipreading: using synthetic talking heads to train humans in lipreading at scale. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2217–2226 (2023)
2. Battina, D.S., Surya, L.: Innovative study of an AI voice-based smart device to assist deaf people in understanding and responding to their body language. SSRN Electron. J. **9**, 816–822 (2021)
3. Beniaguev, D.: Synthetic faces high quality - text 2 image (SFHQ-T2I) dataset (2024). https://doi.org/10.34740/kaggle/dsv/9548853. https://github.com/SelfishGene/SFHQ-T2I-dataset
4. Bernstein, L.E., Jordan, N., Auer, E.T., Eberhardt, S.P.: Lipreading: a review of its continuing importance for speech recognition with an acquired hearing loss and possibilities for effective training. Am. J. Audiol. **31**(2), 453–469 (2022)
5. Chen, Z., Cao, J., Chen, Z., Li, Y., Ma, C.: Echomimic: lifelike audio-driven portrait animations through editable landmark conditions. arXiv preprint arXiv:2407.08136 (2024)
6. Dodd, B., Plant, G., Gregory, M.: Teaching lip-reading: the efficacy of lessons on video. Br. J. Audiol. **23**(3), 229–238 (1989)
7. Kruschke, J.K.: Bayesian estimation supersedes the t test. J. Exp. Psychol. Gen. **142**(2), 573 (2013)

8. Kumar, L.A., Renuka, D.K., Rose, S.L., Wartana, I.M., et al.: Deep learning based assistive technology on audio visual speech recognition for hearing impaired. Int. J. Cogn. Comput. Eng. **3**, 24–30 (2022)
9. Olusanya, B.O., Davis, A.C., Hoffman, H.J.: Hearing loss grades and the international classification of functioning, disability and health. Bull. World Health Organ. **97**(10), 725 (2019)
10. Palanivel, N., Balamurugan, G., Kumar, S.S., Hassan, S., Thirumurugan, C., Thulasinathan, R.: Visual recognition system for hearing impairment using internet of things. In: AIP Conference Proceedings, vol. 3075. AIP Publishing (2024)
11. Peng, Z., et al.: Synctalk: the devil is in the synchronization for talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 666–676 (2024)
12. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 484–492 (2020)
13. Santana, R.S., Menezes, E.C., Ralin, V., Givigi, S., Batorowicz, B.: Lipreading as a communication strategy to enhance speech recognition in individuals with hearing impairment: a scoping review. Disabil. Rehabil. Assist. Technol. 1–12 (2025)
14. Song, L., Liu, P., Yin, G., Xu, C.: Adaptive super resolution for one-shot talking-head generation. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4115–4119 (2024). https://doi.org/10.1109/ICASSP48485.2024.10446837
15. Xu, M., et al.: Hallo: hierarchical audio-driven visual synthesis for portrait image animation. arXiv preprint arXiv:2406.08801 (2024)
16. Ye, Z., et al.: Real3d-portrait: one-shot realistic 3D talking portrait synthesis. arXiv preprint arXiv:2401.08503 (2024)
17. Zhang, Y., et al.: Musetalk: real-time high quality lip synchronization with latent space inpainting. arXiv preprint arXiv:2410.10122 (2024)

LLM-Driven Personalized Answer Generation and Evaluation

Mohammadreza Molavi(✉), Mohammadreza Tavakoli, Mohammad Moein, Abdolali Faraji, and Gábor Kismihók

Leibniz Information Centre for Science and Technology (TIB), Hanover, Germany
{mohammadreza.molavi,reza.tavakoli,mohammad.moein,
abdolali.faraji,gabor.kismihok}@tib.eu

Abstract. Online learning has experienced rapid growth due to its flexibility and accessibility. Personalization, adapted to the needs of individual learners, is crucial for enhancing the learning experience, particularly in online settings. A key aspect of personalization is providing learners with answers customized to their specific questions. This paper therefore explores the potential of *Large Language Models* (*LLMs*) to generate personalized answers to learners' questions, thereby enhancing engagement and reducing the workload on educators. To evaluate the effectiveness of LLMs in this context, we conducted a comprehensive study using the *StackExchange* platform in two distinct areas: language learning and programming. We developed a framework and a dataset for validating automatically generated personalized answers. Subsequently, we generated personalized answers using different strategies, including 0-shot, 1-shot, and few-shot scenarios. The generated answers were evaluated using three methods: 1. *BERTScore*, 2. LLM evaluation, and 3. human evaluation. Our findings indicated that providing LLMs with examples of desired answers (from the learner or similar learners) can significantly enhance the LLMs' ability to tailor responses to individual learners' needs.

Keywords: Question Answering · Personalized Education · LLMs

1 Introduction

Online learning has transformed traditional education, offering flexibility and accessibility to learners worldwide [5]. Its importance was underscored during the COVID-19 pandemic, which accelerated its adoption as a vital alternative to in-person learning [1,20]. Personalization is an important element in online learning, improving efficiency and engagement [4,8], and enabling learners to receive an education tailored to their individual needs [5,17]. A key aspect of personalization is providing tailored answers to learners' questions, offering individualized guidance that enhances their learning [7,10].

Accordingly, efforts have been made to provide personalized answers to learners' questions [19,22]. Yang et al. [19] developed *YA-TA*, a virtual teaching assistant aimed at providing personalized support in large classes. It used a framework

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 187–194, 2025.
https://doi.org/10.1007/978-3-031-98462-4_24

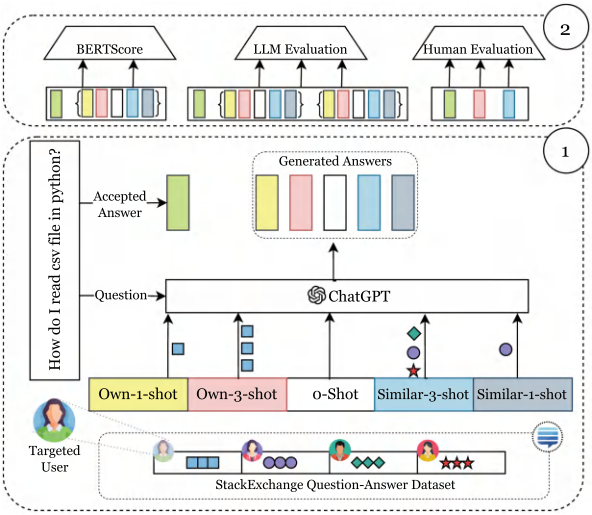


Fig. 1. The proposed pipeline for personalized answer generation consists of two phases: **(1)** scenario creation based on contextual answers. **(2)** evaluation using BERTScore, LLM evaluation, and human evaluation.

that combines instructor and learner knowledge to generate relevant responses. Classroom experiments showed the framework's effectiveness in delivering contextually accurate support. Moreover, Zhang [22] proposed an intelligent Q&A system for online education platforms using *Natural Language Processing* (*NLP*). The system employed techniques like named entity recognition, sentiment analysis, and semantic understanding to effectively analyze and respond to user queries.

The emergence of *Large Language Models* (*LLMs*) has stimulated research into automatic question answering [6,15,16]. For instance, Tan et al. [16] evaluated LLMs on 190,000 complex Knowledge-Based (KB) questions using the *CheckList* framework, identifying limitations in handling real-world KB queries. As another example, Hasan et al. [6] analyzed GPT-3's impact on online learning communities like *StackOverflow*, examining linguistic strengths, problem-solving accuracy, and effects on user participation. Despite these efforts, the development of methods specifically customized to personalized answering remains an underexplored area. These methods not only enhance the learning experience by providing personalization [10] but also significantly reduce the time burden on educators tasked with answering learners' questions [7].

Therefore, this paper aims to (1) develop a framework for evaluating personalized answer generation and (2) investigate the performance of different strategies, such as 0-shot, 1-shot, and few-shot scenarios, in generating personalized answers. Figure 1 depicts our pipeline, which includes the following steps[1]:

1. ***Collecting a dataset.*** Our study required a dataset where learners explicitly expressed their satisfaction with answers. To address this, we utilized the StackExchange platform, where users pose questions and indicate whether an answer was accepted (desired) by them or not.
2. ***Generating various answers.*** We developed different methods to investigate how personalization could improve the baseline answers (0-shot scenario) which were generated by an LLM.
3. ***Evaluating the generated answers.*** To assess the quality of generated answers, we employed a three-phase evaluation approach: 1) calculating BERTScore similarity to desired answers, 2) using an LLM to evaluate the generated answers, and 3) obtaining human expert blind decisions on the similarity of the generated answers to the desired ones.

2 Method

This study focuses on two key educational domains: English language learning and programming [2,3,9,13]. For these domains, we collected the necessary dataset from the StackExchange platform and developed strategies to generate various personalized responses.

Dataset. For creating our dataset, we relied on *StackExchange*[2]. While StackExchange is not a traditional learning platform, it is a recognized educational technology platform and plays a significant role in lifelong, informal, workplace, and problem-based learning scenarios [11,18]. We used a publicly available dataset on *BigQuery*[3], which included a StackOverflow dump up to the end of 2022. To ensure diverse programming topics (e.g., backend, frontend, data science), we selected all questions from 2022 tagged with Python or JavaScript. We focused on users who had asked at least four questions to enable 0-shot, 1-shot, and few-shot evaluations. For these users, we collected all questions with both an accepted answer and at least one other response. The accepted answers served as benchmarks for the users' preferred outcomes. This strategy enabled us to evaluate how effectively our generated answers aligned with learners' needs, as users understood the accepted answer, knew how to apply it, and preferred it over other alternatives. For the English language, due to the limited availability of questions, we applied the same strategy to collect questions posted from 2018 onward using the StackExchange Data Explorer tool. The dataset specifications are as follows: Python programming included *553* users and *3,379* questions,

[1] Supplementary materials including implementations are available here: https://gitlab.com/edoer-research/aied2025_llm_answer_generation.

[2] https://data.stackexchange.com/.

[3] https://bit.ly/stackdump.

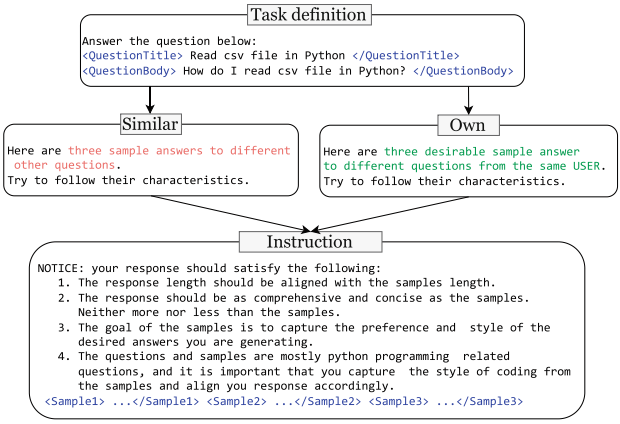


Fig. 2. Our 3-shot prompt scenarios for Python questions, which used sample accepted answers from similar users (similar scenario) and the same users (own scenario).

JavaScript programming had *276* users and *1,630* questions, and English learning comprised *341* users and *1,564* questions.

Personalized Answer Generation. We explored personalized answer generation using well-established prompting scenarios by providing GPT-4o with zero, one, or three sample desired answers as input, which are detailed below:

(1) ***0-shot.*** We did not provide any sample answers in this scenario, establishing a baseline for comparison with personalized scenarios.
(2) ***Own-1-shot.*** We enriched the prompt with the accepted answer from the user's previous question.
(3) ***Own-3-shot.*** Similar to the previous scenario, we incorporated the accepted answers from the user's three previous questions into the prompt.
(4) ***Similar-1-shot.*** To mimic real-world scenarios, where many users have not previously asked questions, we opted to utilize answers from similar users. In this context, we provided a randomly selected accepted answer to a question in the same area in this scenario.
(5) ***Similar-3-shot.*** Accordingly, we incorporated three randomly selected answers from similar users into this scenario.

Finally, we queried GPT-4o with prompts consisting of three parts: task, context, and instructions (Fig. 2). This prompt design follows best practices, including *explicit task definition*, *structured input and output*, and *contextual guidance* [12]. Notably, the *0-shot* scenario only used the task part.

3 Evaluation

Answers generated by GPT-4o were evaluated using BERTScore [21], LLM evaluation, and human evaluation to assess the quality and personalization alignment with desired accepted answers.

Table 1. BERTScore evaluation results.

	0-shot	Similar-1-shot	Similar-3-shot	Own-1-shot	Own-3-shot
Python	64.99%	68.08%	68.31%	68.20%	**68.50**%
JavaScript	67.34%	69.56%	69.80%	69.89%	**70.20**%
English	52.04%	52.14%	52.32%	52.31%	**52.58**%

BERTScore. This metric [21] is widely used to evaluate text quality using contextual embeddings, making it robust to syntactic differences. It correlates well with human judgment, making it a common choice for assessing text generation models. You can find the BERTScore calculation results in Table 1. It can be seen that the 3-shot scenario using learners' previous desired answers achieved the highest scores across all areas, suggesting improved alignment with user needs. All scenarios outperformed the 0-shot approach, indicating that providing context—such as sample desired answers from the same or similar users—enhances response generation.

LLM Evaluation. With the help of GPT-4o, this method aimed to assess the generated answers by identifying which one most closely resembled and aligned with the user's accepted answer. Therefore, GPT-4o was provided with the (1) original question and the user's accepted (desired) answer, and (2) two generated answers from the five defined scenarios. This resulted in ten pairwise evaluations per question. To mitigate order bias [14], we randomized the ordering of the generated answers in the prompt. To select among the answers, the following criteria were provided to GPT-4o:

(1) **Coverage**: Ensure that all question requirements are met.
(2) **Detail consistency**: Match the detail level of the accepted answer.
(3) **Style consistency**: Match the style of the question and accepted answer.
(4) **Coding consistency** (for programming): Match methods with the accepted answer.
(5) **Correctness**: Ensure correctness of the generated answer.

For each pair of scenarios, we analyzed GPT's selections across all questions. Table 2 shows the most frequently chosen scenario along with its selection frequency for different scenario pairs. Consistent with the BERTScore findings, the own-3-shot and own-1-shot scenarios achieved superior results, reinforcing the notion that providing sample desired answers from the same users can enhance the likelihood of generating preferred responses. Moreover, Table 2 clearly distinguishes the 0-shot scenario from the others, demonstrating a significant improvement when desired answers from the same or similar users are provided.

Human Evaluation. In the final evaluation step, we involved at least two experts per domain to choose between pairs of generated answers based on their similarity to the accepted answer. We used the *own-3-shot* and *similar-3-shot* scenarios, identified as optimal in prior evaluations. This process aimed

Table 2. LLM evaluation results.

Scenario pairs		GPT Selected Scenario (Frequency in %)		
		Python	JavaScript	English
own-3-shot	vs. similar-3-shot	own-3-shot (51.5)	own-3-shot (50.38)	own-3-shot (53)
own-3-shot	vs. similar-1-shot	own-3-shot (53.86)	own-3-shot (56.81)	own-3-shot (57.4)
own-3-shot	vs. own-1-shot	own-3-shot (53.45)	own-3-shot (56.23)	own-1-shot (50.61)
own-3-shot	vs. 0-shot	own-3-shot (80.08)	own-3-shot (83.77)	own-3-shot (72.24)
own-1-shot	vs. similar-1-shot	own-1-shot (51.29)	own-1-shot (50.58)	own-1-shot (54.77)
own-1-shot	vs. similar-3-shot	similar-3-shot (50.16)	similar-3-shot (56.63)	own-1-shot (50.23)
own-1-shot	vs. 0-shot	own-1-shot (79.05)	own-1-shot (79.65)	own-1-shot (70.22)
similar-1-shot	vs. 0-shot	similar-1-shot (79.79)	similar-1-shot (81)	similar-1-shot (70.12)
similar-1-shot	vs. similar-3-shot	similar-3-shot (53.2)	similar-3-shot (56.34)	similar-3-shot (55.47)
similar-3-shot	vs. 0-shot	similar-3-shot (79.59)	similar-3-shot (84.33)	similar-3-shot (72.62)

to gather expert opinions—human judgment being the standard for qualitative assessments—and to verify the validity of results from both BERTScore and LLM evaluations. To ensure fair labeling, we developed an application that allows experts to blindly vote on which generated answer best matches the user's desired answer, with an option to skip if unsure. Due to the time-intensive nature of this process, taking approximately ten minutes per question, we gathered 100 labeled questions from each of the domains. Experts preferred the own-3-shot scenario in 67% of Python, 68% of JavaScript, and 70% of English cases, aligning with earlier findings. We also asked the experts to justify their choice of answers in the form of open-ended text. The most frequently mentioned reasons were: (1) citing references from dictionaries and/or the history of terms in the English language domain, (2) employing a similar coding style (e.g., using comments and functions) and applying the same methods within the programming domain, and (3) offering a comparable level of detail (e.g., providing step-by-step explanations).

4 Conclusion

The emphasis on personalization in online learning has led to solutions that tailor the experience to individual needs, including generating answers specific to each learner. By leveraging LLMs, educators can save time while offering more personalized support. This study highlights the potential of LLMs to generate personalized answers in online learning environments.

By developing a framework and a dataset from StackExchange for evaluating personalized answers, we provided opportunities for future studies in this area. Our investigation of different strategies for generating personalized answers showed that providing GPT-4o with sample desired answers from the same learner or similar learners would increase the chance of generating more preferred answers. Our results were based on leveraging various evaluation strategies, including BERTScore, LLM evaluation, and human evaluation.

Our work-in-progress study showed promising results as an initial step, demonstrating that even without complex processing, LLMs are able to capture certain aspects of personalization from previous users' answers. However, further research is needed to explore the limitations and challenges of using LLMs for personalized question answering. Future work should focus on improving answer quality, addressing biases in LLM training data, and investigating the long-term impact on learning outcomes. Additionally, exploring more datasets and domains is necessary to validate the solution's applicability across various contexts.

References

1. Archambault, L., Leary, H., Rice, K.: Pillars of online pedagogy: a framework for teaching in online learning environments. Educ. Psychol. **57**(3), 178–191 (2022). https://doi.org/10.1080/00461520.2022.2051513
2. Boguslawski, S., Deer, R., Dawson, M.G.: Programming education and learner motivation in the age of generative AI: student and educator perspectives. Inf. Learn. Sci. (2024)
3. Choi, L.J.: English as an important but unfair resource: university students' perception of English and English language education in South Korea. Teach. High. Educ. **29**(1), 144–158 (2024)
4. Gemeda, M.G.C., Yadavalli, P.K., Legesse, M.Y.K., Yadessa, M.A.A., Hirpa, M.D.A., Luta, M.R.B.: Inclusiveness in Education. Laxmi Book Publication (2023)
5. Gros, B., García-Peñalvo, F.J.: Future trends in the design strategies and technological affordances of e-learning. In: Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy, pp. 345–367. Springer (2023)
6. Hasan, M.M., Hasan, M., Nayeem, J.U.: An exploratory analysis of community-based question-answering platforms and GPT-3-driven generative AI: is it the end of online community-based learning? (2023)
7. Kokku, R., Sundararajan, S., Dey, P., Sindhgatta, R., Nitta, S., Sengupta, B.: Augmenting classrooms with AI for personalized education. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6976–6980. IEEE (2018)
8. Liu, D.Y.T., Bartimote-Aufflick, K., Pardo, A., Bridgeman, A.J.: Data-driven personalization of student learning support in higher education. In: Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning, pp. 143–169 (2017)
9. Messer, M., Brown, N.C., Kölling, M., Shi, M.: Automated grading and feedback tools for programming education: a systematic review. ACM Trans. Comput. Educ. **24**(1), 1–43 (2024)
10. Pratama, M.P., Sampelolo, R., Lura, H.: Revolutionizing education: harnessing the power of artificial intelligence for personalized learning. Klasikal J. Educ. Lang. Teach. Sci. **5**(2), 350–357 (2023)
11. Prosus: Stack Exchange (2025). https://en.wikipedia.org/wiki/Prosus
12. Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A.: A systematic survey of prompt engineering in large language models: techniques and applications arXiv:2402.07927 (2024). http://arxiv.org/abs/2402.07927, arXiv:2402.07927

13. Shadiev, R., Yu, J.: Review of research on computer-assisted language learning with a focus on intercultural education. Comput. Assist. Lang. Learn. **37**(4), 841–871 (2024)
14. Shi, L., Ma, W., Vosoughi, S.: Judging the judges: a systematic investigation of position bias in pairwise comparative assessments by LLMs. arXiv preprint arXiv:2406.07791 (2024)
15. Siddiq, M.L., Roney, L., Zhang, J., Santos, J.C.D.S.: Quality assessment of chatgpt generated code and their use by developers. In: Proceedings of the 21st International Conference on Mining Software Repositories, pp. 152–156 (2024)
16. Tan, Y., et al.: Can chatgpt replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family. In: International Semantic Web Conference. Springer (2023)
17. Tavakoli, M., Hakimov, S., Ewerth, R., Kismihók, G.: A recommender system for open educational videos based on skill requirements. In: 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), pp. 1–5. IEEE (2020)
18. Wikipedia: Stack Exchange (2025). https://en.wikipedia.org/wiki/Stack_Exchange
19. Yang, D., et al.: Ya-ta: towards personalized question-answering teaching assistants using instructor-student dual retrieval-augmented knowledge fusion. arXiv preprint arXiv:2409.00355 (2024)
20. Zhang, L., Carter, R.A., Qian, X., Yang, S., Rujimora, J., Wen, S.: Academia's responses to crisis: a bibliometric analysis of literature on online learning in higher education during covid-19. Br. J. Edu. Technol. **53**(3), 620–646 (2022). https://doi.org/10.1111/bjet.13191
21. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: evaluating text generation with bert arXiv:1904.09675 (2020). https://doi.org/10.48550/arXiv.1904.09675. http://arxiv.org/abs/1904.09675. arXiv:1904.09675
22. Zhang, Y.: Design of an intelligent Q&A system for online education platform based on natural language processing technology. J. Electr. Syst. **20**(6s), 2135–2145 (2024)

Predicting User Actions in Algebra Intelligent Tutoring Systems with Markov Models

Pablo Arnau-González[1(✉)], Yuyan Wu[1], Sergi Solera-Monforte[1], David Arnau[2], and Miguel Arevalillo-Herráez[1]

[1] Departament d'Informàtica, Universitat de València, Valencia, Spain
{pablo.arnau,yuyan.wu,sergi.solera,miguel.arevalillo}@uv.es
[2] Departament de Didàctica de la Matemàtica, Universitat de València, Valencia, Spain
david.arnau@uv.es

Abstract. Providing effective and personalized support for students solving Math Word Problems (MWPs) in Algebra Intelligent Tutoring Systems (ITS) remains a significant challenge. Current help systems often rely on rule-based approaches limiting their adaptability to individual student needs and hindering proactive error detection. This work proposes a novel approach leveraging Hidden Markov Models (HMMs) to model user action sequences during MWP solving. We trained HMMs on a dataset of student interaction logs from an algebra ITS, where each action represents an observation and latent states represent the level of understanding. The HMM can effectively capture the dynamics of student's understanding level by probabilistically modelling the relationship between actions. The trained HMM models were evaluated on their ability to predict student actions and identify potential errors compared to the system's existing rule-based help and a Hidden Markov model. Results demonstrate that the HMM-based approach achieves significantly higher accuracy in predicting subsequent actions.

Keywords: Hidden Markov Model · Student Modelling · Problem Solving · Intelligent Tutoring Systems

1 Introduction

Leveraging the potential of ITSs in mathematics education requires addressing specific challenges, particularly in supporting students. MWPs require students to translate real-world scenarios into mathematical expressions, formulate equations, and apply algebraic manipulations to arrive at a solution [17,18]. Which often leads to cognitive overload and a variety of errors [16]. Existing help systems in algebra ITSs typically employ rule-based approaches [2,20], pre-defined assistance sequences [20], or model-tracing techniques where the system compares student's actions to what the ITS would do in the same situation [1]. While

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 195–202, 2025.
https://doi.org/10.1007/978-3-031-98462-4_25

these methods can provide some level of support, they often struggle to adapt to the diverse and evolving problem-solving strategies of individual students.

Reactive feedback in ITSs, while proven beneficial for learning, only addresses errors after they occur [10,14]. This approach can be limiting, as it may not fully account for potential shifts in student reasoning when they navigate challenging solution paths. For example, while the help provided by the HINTS ITS [4] demonstrates effectiveness in providing hints to the student, its inability to anticipate these shifts may lead to a disconnect between the support offered and the student's immediate requirements. We propose a novel approach using Hidden Markov Models (HMMs) to model student action sequences in MWP solving within the HINTS [3] system. HMMs are capable of capturing latent states based on observed actions, allowing us to infer a student's understanding and predict future steps [8,9]. The innovation of our approach lies in modelling of the solution process and accurately predicting the next action the user would take. The main application of the proposed method lies in delivering unsolicited assistance to students engaging with an ITS, by predicting and guiding their next action. Although the approach has been evaluated within the context of HINTS, its applicability extends beyond this scope. The methodology can be seamlessly adapted to various ITS platforms that incorporate predictive modelling, enabling them to generate contextually relevant hints based on the anticipated next steps. The presented approach can also be integrated with other ITS functionalities, such as step-by-step scaffolding, error correction guidance, or even adaptive feedback mechanisms.

2 Related Work

One of the primary trends in ITS research is adaptive and personalized assistance. Modern ITS dynamically adjust learning activities based on the student's progress, cognitive state, and effective factors [6]. Clement et al. [13] introduced a model that personalizes sequences of learning activities to maximize skill acquisition, tailoring recommendations according to the learner's needs.

Another key area of research is data-driven hint generation in intelligent tutoring systems [12]. Barnes et al. proposed using Markov Decision Processes (MDP) to automatically generate hints [7], while Vanlehn developed an adaptive hint system based on probabilistic models such as Item Response Theory (IRT) [19]. Others have introduced a transformer-based approach to develop proactive hints, further advancing the effectiveness of intelligent tutoring systems [21].

Additionally, the shift toward web-based and scalable ITS has been notable. Web-based ITS solutions emphasize scalability and accessibility, particularly in online education contexts [5]. Brusilovsky [11] discussed the increasing role of web-based ITS and their potential for remote learning applications.

3 Modelling

Within the context of MWPs, we first define the fundamental components: quantities and relationships. Let Q represent the set of quantities present in a given

MWP. These quantities are numerical values, either explicitly stated within the problem text or implicitly present as unknown values. Subsequently, let R denote the set of relationships that exist among these quantities in Q. Each relationship in R is defined as a connection involving at least three members of Q, representing dependencies or operations that link these quantities according to the problem's constraints. A Math Word Problem, as conceptualized in this work, is then represented as a series of hypergraphs $H = \{H_0, ..., H_n\}$. Each hypergraph $H_n = \langle Q', R' \rangle$ in this series embodies a potential solution path for the MWP. Constructed from subsets of Q and R, each hypergraph is composed of a set of hypernodes Q', where $Q' \subseteq Q$, and a set of hyperedges R', where $R' \subseteq R$. Q' represents the quantities involved in a particular stage or aspect of a solution path, and R' represents the relationships among these quantities relevant to that specific path. Originally, the help algorithm in HINTS employed a rule-based strategy [3] to provide targeted assistance to students. Whenever a student defined a quantity with a letter name, in terms of two others, or as part of an equation, the system would mark this quantity as known, and the corresponding relations as completed. The original system provides hints under the assumption that the most advanced hypergraph is the one that the student was solving. To do so, the system first computes the completeness of each hypergraph as $\frac{|R'_{known}|}{|R'|}$. Then, the system examines the relations and quantities defined in that graph, looking first for equations, then for expressions, and quantities that could be named. However, this behavior does not model possible changes of reasoning in the student during the solving of a given hypergraph, i.e. getting to a point where it is easier to start over following a different solution path. Or solve the problem following a mechanic approach, first defining quantities, and then solving them in terms of others.

In the studied ITS, the user action space is limited by the number of quantities and relations that exist in each MWP. In total, there are three different possible actions: name a quantity, define a quantity in terms of two others, or establish an equation. As these actions can only be performed once, the current problem state can be represented as a vector, which encodes whether each of these actions has been performed. It should be noted that although the state-space grows exponentially ($2^{2|Q|+|R|}$ possible actions) most of the states cannot be possibly reached due to logical restrictions. Once the problem solution states have been defined, it is possible to establish the sequence of states. To this end, we propose two different featurizations of the action space. First characterization (SYMBOL) is just an encoding of each of the possible actions as a categorical variable. Therefore, actions A, B, C will be mapped to $1, 2, 3$ and possible user sequences will be transformed like $[A, C, B] \rightarrow [1, 3, 2]$. The second characterization (STATE) encodes whether each action in the set of possible actions has been taken. Therefore, actions A, B, C would be mapped to the vectors $[0, 0, 1], [0, 1, 0], [1, 0, 0]$, respectively. The sequence of actions is then modelled by adding these vectors. Finally, the action sequence $[A, C, B]$ is computed as $[A, A + C, A + C + B]$. These action are represented in binary as $[[0, 0, 1], [1, 0, 1] [1, 1, 1]]$ and are mapped to the categories $[1, 5, 7]$.

4 Methodology

The data has been captured from the logs produced by HINTS. The data was gathered from over 100 primary school students when solving arithmetic and algebraic MWPs. During the preprocessing phase, we extracted relevant action information from the raw session logs, focusing on identifying and mapping events to their corresponding problem-solving path nodes in the graph. Since the goal of the study is to predict the next step the student should take, steps involving mistakes were excluded from the analysis. Additionally, exercises that were incomplete or had fewer than two attempts, or had less than two hypergraphs, were also excluded. After filtering out inconsistent records, a total of 24 816 events were processed from 767 students and 89 different problems.

Bias and Fairness. We have made every effort to avoid introducing bias in the captured data. The original data was collected from primary school classes, where all students participated equally in the system. In addition, the data is fully anonymous. There is no reason to believe that the exclusion criteria are correlated with any individual characteristics such as gender, race, or cultural background.

4.1 Experimental Setting

The validity of the proposed approach has been confirmed for each of the factors separately. First, there is the modelling of the sequence of actions whether it is as the whole problem state or a categorization of the particular action. Second, two different predictors: Markov Chains and Hidden Markov Models have been used to model the sequence. To evaluate each contributions individually, we conducted five experiments, testing each contribution and their combinations. The evaluation of the performance has been carried out following a leave-one-out cross-validation strategy tailored to each individual problem within the dataset. Thus, for every unique problem considered, we treated the collection of all available solution attempts as the dataset specific to that problem. Then, we conducted a leave-one-out cross-validation within these attempts. In each iteration of this problem-specific cross-validation, a single solution attempt was designated as the test set, while the remaining attempts for the same problem formed the training set. Our model was then trained on this training set and evaluated on the held-out test attempt, focusing on the per-step accuracy achieved during the solution process. The training of the HMM has been done using the Baum-Welch algorithm, while the transitions for the Markov Chains, have been set, by computing the frequency of transitions from each state to the others in the test set.

Finally, to obtain a representative performance metric for each isolated MWP, we averaged the per-step accuracy scores across all folds of the leave-one-out cross-validation conducted for that specific problem. This resulted in a single average per-step accuracy value for each individual problem and experimental setting.

5 Results

Experimental results are displayed in Table 1 along with the implementation of the original rule-based algorithm (BASELINE). As the table shows, the best combination consists on the mixture of the encoding of the MWP state along with an HMM classifier (HMM-STATE), reporting a median accuracy of 0.98 which although is not significantly better than the symbol featurization counterpart (Wilcoxon $p = 0.069$) shows more consistency in the predictions according to the boxplot in Fig. 1 which displays the distribution of average per-problem accuracy for each explored experimental setting. The Ordinary Least Squares regression model demonstrated a statistically significant relationship between the predictor and feature variables and the score (F-statistic = 43.67, $p < 0.001$), explaining approximately 30% of the variance in the score ($R^2 = 0.304$). Compared to the baseline condition, significant positive effects were observed for alternative predictor and feature categories. Specifically, utilizing "CHAIN" or "HMM" as the predictor, and "STATE" or "SYMBOL" as the feature, independently resulted in a statistically significant increase in the score.

Table 1. Performance Results for Predictors and Featurization Strategies

Experiment	Mean Acc ± std	Median Acc
HMM-STATE	0.9030 ± 0.18	**0.9827**
HMM-SYMBOL	0.9090 ± 0.12	0.9560
CHAIN-STATE	0.9074 ± 0.13	0.9560
CHAIN-SYMBOL	0.8446 ± 0.23	0.9214
BASELINE	0.5747 ± 0.27	0.5238

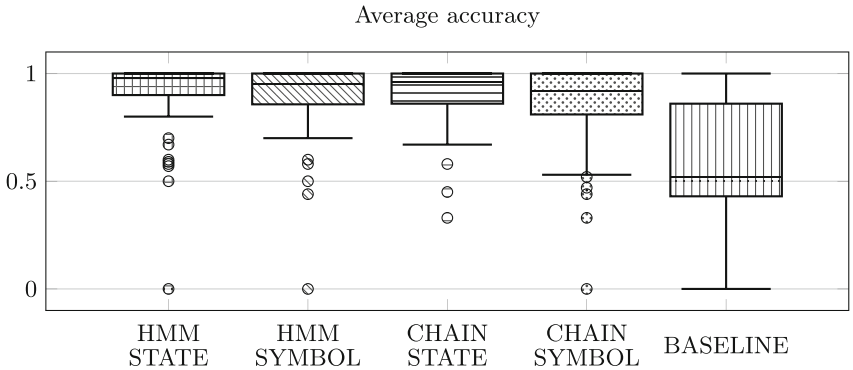


Fig. 1. Difference of accuracy for the proposed (left) and the original method (right)

6 Conclusions

Some ITS can suggest the next action using deterministic algorithms, but this is not always feasible. When such methods are insufficient, an alternative approach is to analyze the actions of other users in similar situations, particularly those that lead to correct outcomes. In this work, we demonstrate that user actions can be modeled as a Markov process, enabling the prediction of the most logical next step based on past user behavior. This model can then be leveraged to propose the next action and generate meaningful hints. In particular, we have presented a novel method for modelling the sequence of actions taken by a student during the solving of a math word problem. The presented method shows significantly better performance for detecting which will be the action taken by the student in the studied cases.

The literature shows that there are different types of learners [15]. Future work should explore this issue and analyze whether students can be grouped based on the strategies they used to solve problems. Although HMMs might be able to capture the different paths that are created, clusterized models would provide tailored experiences for students. The proposed approach is heavily dependent on data availability. HMM models need to be trained, and since there is no direct mapping between the hidden states and the observation they cannot be established following any heuristic. Thus, the proposed approach is not a replacement of the rule-based method when there is not sufficient data to train the HMM. Continuation of our work should delve into finding where is the point in which the trained HMM is likely to improve the rule-based method.

The use of HMMs to predict actions allows students to follow a more natural problem solving scheme compared to a the current approach, as it can adapt quickly to changes in the reasoning of students. Moreover, currently, HINTS needs that the solution paths are encoded in the system. This requires a didactics expert to analyze the problem and manually encode the possible solution paths. HMMs learn from past data, so if initial performance can be sacrificed, there is no need of an expert to perform a detailed analysis of the solution paths, as they would emerge from past experiences.

Acknowledgements. This research has been supported by project TED2021-129485B-C42, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR; and project PID2023-150960NB-I00, funded by the Spanish Ministry of Science, Innovation and Universities and the European Union and by Generalitat Valenciana through grant CIAPOS/2022/163, grant ACIF/2021/439, and project CIAICO/2023/065;

References

1. Aleven, V.: Rule-Based Cognitive Modeling for Intelligent Tutoring Systems, pp. 33–62. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14363-2_3
2. Aleven, V., Mclaren, B., Roll, I., Koedinger, K.: Toward meta-cognitive tutoring: a model of help seeking with a cognitive tutor. Int. J. Artif. Intell. Ed. **16**(2), 101–128 (2006)
3. Arevalillo-Herráez, M., Arnau, D., Marco-Giménez, L.: Domain-specific knowledge representation and inference engine for an intelligent tutoring system. Knowl.-Based Syst. **49**, 97–105 (2013)
4. Arnau, D., Arevalillo-Herráez, M., Puig, L., González-Calero, J.A.: Fundamentals of the design and the operation of an intelligent tutoring system for the learning of the arithmetical and algebraic way of solving word problems. Comput. Educ. **63**, 119–130 (2013)
5. Arnau-González, P., Solera-Monforte, S., Wu, Y., Arevalillo-Herráez, M.: A framework for adapting conversational intelligent tutoring systems to enable collaborative learning. Expert Syst. Appl. 126663 (2025)
6. Arroyo, I., Woolf, B.P., Burelson, W., Muldner, K., Rai, D., Tai, M.: A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. Int. J. Artif. Intell. Educ. **24**, 387–426 (2014)
7. Barnes, T., Stamper, J.C., Lehmann, L., Croy, M.J.: A pilot study on logic proof tutoring using hints generated from historical student data. In: EDM, pp. 197–201 (2008)
8. Beal, C., Mitra, S., Cohen, P.R.: Modeling learning patterns of students with a tutoring system using hidden Markov models. In: Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, pp. 238–245. IOS Press, NLD (2007)
9. Boyer, K., et al.: Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden Markov modeling approach. Int. J. Artif. Intell. Educ. **21**, 65–81 (2011). https://doi.org/10.3233/JAI-2011-018
10. Brown, J.S., Burton, R.R., Bell, A.G.: Sophie: a step toward creating a reactive learning environment. Int. J. Man Mach. Stud. **7**(5), 675–696 (1975)
11. Brusilovsky, P.: Adaptive hypermedia: from intelligent tutoring systems to web-based education. In: Gauthier, G., Frasson, C., VanLehn, K. (eds.) ITS 2000. LNCS, vol. 1839, pp. 1–7. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45108-0_1
12. Cavalcanti, A.P., et al.: Automatic feedback in online learning environments: a systematic literature review. Comput. Educ. Artif. Intell. **2**, 100027 (2021)
13. Clement, B., Roy, D., Oudeyer, P.Y., Lopes, M.: Multi-armed bandits for intelligent tutoring systems. J. Educ. Data Mining **7**(2) (2015). https://doi.org/10.5281/zenodo.3554668
14. D'Mello, S., Olney, A., Williams, C., Hays, P.: Gaze tutor: a gaze-reactive intelligent tutoring system. Int. J. Hum. Comput. Stud. **70**(5), 377–398 (2012)
15. Hariri, D., Mahmudah, H., Wibawa, F., Kania, N.: Unraveling the connection: a systematic review of learning styles and mathematics achievement. Pedagogical Res. **10**(1) (2025)
16. Jaffe, J.B., Bolger, D.J.: Cognitive processes, linguistic factors, and arithmetic word problem success: a review of behavioral studies. Educ. Psychol. Rev. **35**(4), 105 (2023). https://doi.org/10.1007/s10648-023-09821-6

17. Nesher, P., Teubal, E.: Verbal cues as an interfering factor in verbal problem solving. Educ. Stud. Math. **6**(1), 41–51 (1975). https://doi.org/10.1007/BF00590023
18. Riley, M.S., et al.: Development of children's problem-solving ability in arithmetic (1984)
19. Ueno, M., Miyazawa, Y.: IRT-based adaptive hints to scaffold learning in programming. IEEE Trans. Learn. Technol. **11**(4), 415–428 (2018). https://doi.org/10.1109/TLT.2017.2741960
20. Vanlehn, K.: The behavior of tutoring systems. Int. J. Artif. Intell. Educ. **16**, 227–265 (2006)
21. Wu, Y., Arevalillo-Herráez, M., Solera-Monforte, S.: A generative approach for proactive assistance forecasting in intelligent tutoring environments. In: Generative Intelligence and Intelligent Tutoring Systems: 20th International Conference, ITS 2024, Thessaloniki, Greece, 10–13 June 2024, Proceedings, Part I, pp. 14–26. Springer, Heidelberg (2024). https://doi.org/10.1007/978-3-031-63028-6_2



Many-Shot Regurgitation Prompting

Shashank Sonkar[2(✉)], Naiming Liu[1], and Richard Baraniuk[1]

[1] Rice University, Houston, USA
[2] University of Central Florida, Orlando, USA
shashank.sonkar@ucf.edu

Abstract. We introduce Many-Shot Regurgitation (MSR) prompting, a new black-box membership inference attack framework for examining verbatim content reproduction in large language models (LLMs). MSR prompting involves dividing the input text into multiple segments and creating a single prompt that includes a series of faux conversation rounds between a user and a language model to elicit verbatim regurgitation. We apply MSR prompting to diverse text sources, including open educational resources textbooks and Wikipedia articles, which provide high-quality, factual content and are continuously updated over time. For each source, we curate two dataset types: one that LLMs were likely exposed to during training (D_{pre}) and another consisting of documents published after the models' training cutoff dates (D_{post}). To quantify the occurrence of verbatim matches, we employ the Longest Common Substring algorithm and count the frequency of matches at different length thresholds. We then use statistical measures such as Cliff's delta, Kolmogorov-Smirnov (KS) distance, and Kruskal-Wallis H test to determine whether the distribution of verbatim matches differs significantly between D_{pre} and D_{post}. Our findings reveal a striking difference in the distribution of verbatim matches between D_{pre} and D_{post}, with the frequency of verbatim reproduction being significantly higher when LLMs (e.g. GPT models and LLaMAs) are prompted with text from datasets they were likely trained on. Our results provide compelling evidence that LLMs are more prone to reproducing verbatim content when the input text is likely sourced from their training data. Code is available here.

Keywords: Membership Inference Attacks · Large Language Models

1 Introduction

Large Language Models (LLMs) like GPT [1], Gemini [34], and LLaMA [2,35] have revolutionized NLP and significantly impacted the generation of educational content [18,31,32]. As questions about AI alignment [25,36] and training data provenance become increasingly important, we investigate: to what extent do these models reproduce verbatim content from their training data? We develop the Many-Shot Regurgitation (MSR) prompting technique to efficiently investigate this phenomenon. MSR represents a new framework for membership

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 203–211, 2025.
https://doi.org/10.1007/978-3-031-98462-4_26

inference attack (MIA) against language models, which traditionally aims to determine whether inputs were part of training data [19,30]. Unlike prior MIA research focusing on fine-tuning [14,22] and requiring access to model internals [7,11,29], MSR operates in a black-box setting, making it applicable to deployed models like GPT-3.5 [27] and GPT-4 [1,23].

The MSR prompting technique, inspired by the many-shot jailbreaking approach [4,8], involves dividing the input text into multiple segments and creating a single prompt that includes a series of faux conversation rounds between a user and a language model designed to elicit verbatim regurgitation. The input text is segmented into multiple parts, denoted as T_1, T_2, ..., T_n, where n is the total number of segments. A single prompt is then constructed, creating a simulated conversation sequence between a user and a language model. Throughout these rounds of conversation, the user supplies the odd-numbered segments (T_1, T_3, ..., T_{n-1}), while the even-numbered segments (T_2, T_4, ..., T_{n-2}) are presented as if they were generated by the language model. In the final round, the user presents the penultimate segment (T_{n-1}), prompting the language model to generate the concluding segment (T'_n). It is worth highlighting that the segments presented as responses from the language model are not genuine outputs but are integrated into the prompt to mimic a conversation. The sole genuine output from the LLM is the final segment (T'_n).

While quantifying the occurrence of verbatim matches between the generated text (T'_n) and the original segment (T_n) is important, it doesn't provide a complete understanding of verbatim reproduction. Thus, we also explore whether the distribution of these verbatim matches is statistically different from matches in documents that the LLMs are unlikely to have been trained on. To achieve this, we curate two types of dataset sources: D_{pre}, which includes documents that the LLMs were likely exposed to during training, and D_{post}, which consists of documents that were published after the LLMs' training cutoff dates. Specifically, we utilize Wikipedia articles and Open Educational Resource (OER) textbooks as they provide high-quality, factual content and continuously updated over time.

Through experiments employing the MSR prompting technique on both D_{pre} and D_{post} datasets, we aim to compare the distributions of verbatim matches. We employ statistical tests such as Cliff's Delta [10], Kolmogorov-Smirnov (KS) Distance [3], and the Kruskal-Wallis H Test [17] to evaluate the statistical significance of the differences in verbatim reproduction between the two dataset types. Our experiments are performed using GPT-3.5 [27], GPT-4 [1], and LLaMA [2].

Our experiments reveal that verbatim reproduction occurs in text generated by different LLMs, with the extent of reproduction varying across dataset sources and language models. We find that the frequency of verbatim matches is generally higher for D_{pre} compared to D_{post}, suggesting that the availability of the dataset influences the likelihood of verbatim reproduction. For example, using GPT-3.5 on Wikipedia articles, we observe a Cliff's delta of -0.984 and a KS distance of 0.875, indicating a significant difference between D_{pre} and D_{post}.

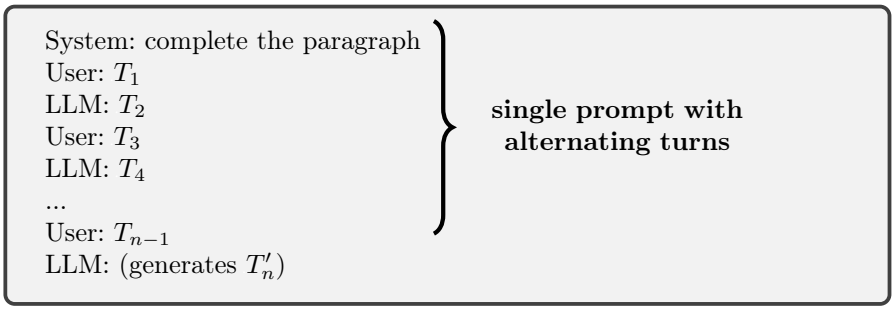
Related Work. MIAs determine if data samples were used in model training [7,29]. Studied extensively in NLP [14,19,22,30], MIAs help quantify pri-

vacy risks [21], measure memorization [6], audit models [33,37], and detect test set contamination [24] or unauthorized copyrighted content use [12,20]. Most MIA research focuses on fine-tuning rather than pretraining [14,19,22]. Previous works have explored automated approaches: studying embedding models' information leakage [30], auditing training data via gradient updates [26], measuring data leakage with exposure metrics [7], and mitigating risks through deduplication [16]. Large pre-trained models present unique challenges due to limited training data access, massive datasets, and proprietary pretraining corpora [11,28]. Most methods assume access to model internals (logits, gradients, token probabilities) [7,19,28,29,33], creating challenges for models where such access is restricted or deprecated [11,28]. Our MSR technique addresses these limitations as a black-box MIA framework requiring no model internal access.

2 Methodology: Many-Shot Regurgitation Prompting

The MSR prompting technique involves dividing the source text into multiple segments (shots) and creating a single prompt with a series of alternating turns between a user and the language model. The key steps of MSR are as follows:

1. **Text Segmentation:** The source text is divided into an even number of segments, denoted as T_1, T_2, ..., T_n, where n is the total number of segments.
2. **Prompt Construction:** A single prompt is created by concatenating the segments in an alternating manner, simulating a conversation between a user and the language model. The user provides the odd-numbered segments (T_1, T_3, ..., T_{n-1}), and the LLM is prompted to generate the even-numbered segments (T_2, T_4, ..., T_n).



   It's important to note that the parts labeled as LLM's responses (T_2, T_4, ..., T_{n-2}) are not actually generated by the model but are included in the prompt to create the illusion that LLM has generated those parts. Only the last part (T'_n) is the genuine output from the LLM.
3. **Text Generation:** The language model is tasked with generating the final segment T_n based on the constructed prompt. The generated segment is denoted as T'_n.

Verbatim Match Analysis. To quantify the extent of verbatim regurgitation, we compare the generated text (T'_n) and the corresponding original segment

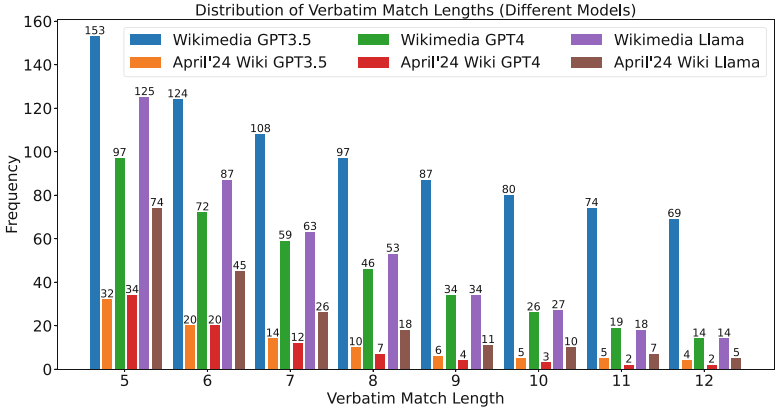


Fig. 1. Verbatim match frequency distribution across GPT and LLAMA models using Wikipedia data. All models show significantly higher verbatim match rates for pre-training articles (Wikimedia) versus post-training articles (April 2024 Wiki), especially at longer match lengths. Wikimedia content maintains higher frequencies even at 10+ word matches, while April Wiki frequencies decline sharply as match length increases.

(T_n) using the Longest Common Substring (LCS) algorithm. The LCS algorithm identifies the longest contiguous substring that is present in both T'_n and T_n.

We define a verbatim match as a substring of a specified minimum length (e.g., $\ell_{\min}$ to $\ell_{\max}$ words) that is identical between the generated and original segments. Formally, let M_k be the set of all substring matches of length k between T'_n and T_n. The frequency of exact verbatim matches of length k be $f_k = |M_k|$.

For each language model and dataset source combination, we obtain an array of frequencies $F = [f_{\ell_{\min}}, f_{\ell_{\min}+1}, ..., f_{\ell_{\max}}]$ corresponding to verbatim match length thresholds from $\ell_{\min}$ to $\ell_{\max}$ words. By counting the frequency of verbatim matches at different length thresholds, we obtain a distribution of verbatim match frequencies for each combination of language model and dataset source.

Statistical Analysis. We use the following complementary measures:

- Cliff's delta [10]: A non-parametric effect size measure (-1 to +1) quantifying distribution differences, chosen for its robustness to outliers and non-normal distributions when comparing frequencies between D_{old} and D_{new}.
- Kolmogorov-Smirnov (KS) distance [3]: Measures the maximum distance between cumulative distribution functions (0 to 1), complementing Cliff's delta in assessing dissimilarity between D_{old} and D_{new} distributions.
- Kruskal-Wallis H test [17]: Determines statistical significance of distribution differences between groups, suitable for our analysis as it requires no assumptions about normality or variance homogeneity.

3 Dataset Curation and Experiments

Datasets and Language Models. In our experiments, we utilized two types of dataset sources: Open Educational Resource (OER) textbooks and Wikipedia articles. For OER textbooks, we selected D_{pre}, comprising textbooks with publication dates before the language models' training cutoff dates, such as "OpenStax Biology 2e" [9], "OpenStax Economics 2e" [15], and "OpenStax Concepts of Biology" [13], and D_{post}, including recently released textbooks, such as two new OpenStax textbooks on nutrition for nurses [5] and pharmacology, which are unlikely to be part of the language models' training data. For Wikipedia, we curated two subsets: D_{pre}, containing approximately 600 articles with a length greater than 1,000 words, selected from a time period significantly before the language models' training cutoff dates, and D_{post}, consisting of around 600 articles with a length greater than 1,000 words, crawled from the last few months, which are likely to be absent from the language models' training data. This controlled setting helps us investigate the influence of dataset availability on the extent of verbatim regurgitation. We evaluated three language models: GPT-3.5, GPT-4, and LLaMA3-70B. For GPT-3.5 and GPT-4, we used a temperature setting of 0.1 to encourage more deterministic and less diverse outputs.

Table 1. Statistical analysis shows verbatim match distribution differences between pre/post-training datasets. Negative Cliff's Delta indicates higher match rates in D_{pre}, with KS test quantifying distribution differences and KW test confirming significantly higher frequencies in D_{pre} compared to D_{post} ($p < 0.05$).

Datasets	LLM	Cliff's δ	KS Distance	KW H-statistic	KW p-value
Wikipedia	GPT-3.5	−1.0	1.0	11.31	0.0008
	GPT-4	−0.828	0.75	7.77	0.0053
	LLAMA	−0.578	0.50	3.78	0.0519
OER	GPT-3.5	−0.984	0.875	11.12	0.0009
	LLAMA	−0.828	0.625	7.80	0.0052

Primary Experiments for Verbatim Regurgitation Analysis. In our base experiments, we applied MSR prompting technique to each combination of language model and dataset source. We used the following many shot configurations: 6 shots (3 human & 3 assistant turns). For each source text, we obtained the generated text (T'_n) and calculated the frequency of exact verbatim matches for substring lengths ranging from 5 to 12 words.

Figure 1 presents the distribution of verbatim match frequencies for different language models on the Wikipedia dataset. The results show that the frequency of verbatim matches is consistently higher for the articles published before the training cutoff dates (Wikimedia) compared to those published after (April 2024 Wiki), across all three language models (GPT-3.5, GPT-4, and LLaMA). The

statistical analysis results, summarized in Table 1, confirm the significance of these differences, with GPT-3.5 showing the most substantial difference (Cliff's Delta = −1.0, KS Distance = 1.0, H-statistic = 11.31, p = 0.0008), followed by GPT-4 (Cliff's Delta = −0.828, KS Distance = 0.75, H-statistic = 7.77, p = 0.0053) and LLAMA (Cliff's Delta = −0.578, KS Distance = 0.5, H-statistic = 3.78, p = 0.0519). Similarly, Fig. 2 displays the distribution of verbatim match frequencies for GPT-3.5 and LLAMA on the OER textbook dataset. The frequency of verbatim matches is notably higher for the textbooks published before the training cutoff dates (Old Bio & Econ) compared to the recently released textbooks (New Pharma & Nursing). Table 1 shows that GPT-3.5 exhibits a highly significant difference (Cliff's Delta = −0.984, KS Distance = 0.875, H-statistic = 11.12, p = 0.0009), while LLAMA also shows a substantial difference (Cliff's Delta = −0.828, KS Distance = 0.625, H-statistic = 7.80, p = 0.0052).

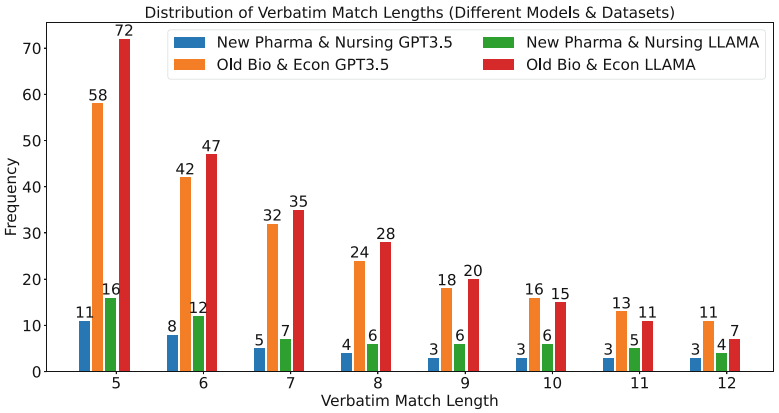


Fig. 2. Distribution of verbatim match frequencies for GPT-3.5 and LLAMA on the OER textbook dataset. The frequency of verbatim matches is consistently higher for the textbooks published before the training cutoff dates (Old Bio & Econ) compared to the recently released textbooks (New Pharma & Nursing).

These results demonstrate a clear difference in the distribution of verbatim matches between the datasets published before and after the training cutoff dates for both dataset sources. The negative Cliff's Delta values, substantial KS distances, and significant Kruskal-Wallis H-statistics consistently indicate that verbatim match frequencies are higher for D_{pre} compared to D_{post}.

Ablation Studies. We conduct experiments examining shot count (2-8) and temperature settings (0.1 vs 0.7). Six shots produced optimal verbatim matching, while lower temperature settings significantly increased verbatim match frequencies ($p < 0.05$), confirmed by moderate effect sizes in statistical tests.

4 Conclusion

We introduced Many-Shot Regurgitation (MSR) prompting, a novel black-box approach for investigating verbatim content reproduction in large language models, particularly for educational documents. Using the Longest Common Substring algorithm, we compared verbatim match distributions between pre-training and post-training datasets from OER textbooks and Wikipedia articles. Our experiments with GPT-3.5, GPT-4, and LLAMA revealed striking differences in verbatim reproduction between content likely in training data versus newer content. These findings have significant implications for educational content creators and institutions using LLMs to process or generate academic materials.

Acknowledgments. Work supported by NSF grant 1842378, ONR grants N0014-20-1-2534 and N00014-18-1-2047, AFOSR grant FA9550-22-1-0060 & Vannevar Bush Faculty Fellowship.

References

1. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. AI@Meta: Llama 3 model card (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
3. An, K.: Sulla determinazione empirica di una legge didistribuzione. Giorn Dell'inst Ital Degli Att **4**, 89–91 (1933)
4. Anil, C., et al.: Many-shot jailbreaking
5. Bilbrew, E.C., Vogelzang, J., Whittington, K.: Nutrition for Nurses. OpenStax, Houston, Texas (2024)
6. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., Zhang, C.: Quantifying memorization across neural language models. arXiv preprint arXiv:2202.07646 (2022)
7. Carlini, N., et al.: Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 2021), pp. 2633–2650 (2021)
8. Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G.J., Wong, E.: Jailbreaking black box large language models in twenty queries (2023)
9. Clark, M.A., Douglas, M., Choi, J.: Biology 2e. OpenStax, Houston, Texas (2018)
10. Cliff, N.: Dominance statistics: ordinal analyses to answer ordinal questions. Psychol. Bull. **114**(3), 494 (1993)
11. Duan, M., et al.: Do membership inference attacks work on large language models? arXiv preprint arXiv:2402.07841 (2024)
12. Duarte, A.V., Zhao, X., Oliveira, A.L., Li, L.: De-cop: detecting copyrighted content in language models training data. arXiv preprint arXiv:2402.09910 (2024)
13. Fowler, S., Roush, R., Wise, J.: Concepts of Biology. OpenStax, Houston, Texas (2013)
14. Fu, W., Wang, H., Gao, C., Liu, G., Li, Y., Jiang, T.: Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. arXiv preprint arXiv:2311.06062 (2023)
15. Greenlaw, S.A., Shapiro, D.: Principles of Economics 2e. OpenStax, Houston, Texas (2017)

16. Kandpal, N., Wallace, E., Raffel, C.: Deduplicating training data mitigates privacy risks in language models. In: International Conference on Machine Learning, pp. 10697–10707. PMLR (2022)
17. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. **47**(260), 583–621 (1952)
18. Liu, N., Sonkar, S., Wang, Z., Woodhead, S., Baraniuk, R.G.: Novice learner and expert tutor: evaluating math reasoning abilities of large language models with misconceptions. arXiv preprint arXiv:2310.02439 (2023)
19. Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., Berg-Kirkpatrick, T.: Membership inference attacks against language models via neighbourhood comparison. arXiv preprint arXiv:2305.18462 (2023)
20. Meeus, M., Jain, S., Rei, M., de Montjoye, Y.A.: Did the neurons read your book? document-level membership inference for llms. arXiv preprint arXiv:2310.15007 (2023)
21. Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., Shokri, R.: Quantifying privacy risks of masked language models using membership inference attacks. arXiv preprint arXiv:2203.03929 (2022)
22. Mireshghallah, F., Uniyal, A., Wang, T., Evans, D.K., Berg-Kirkpatrick, T.: An empirical analysis of memorization in fine-tuned autoregressive language models. In: Proceedings of the 2022 Conference on EMNLP, pp. 1816–1826 (2022)
23. OpenAI: Gpt-4 technical report (2023)
24. Oren, Y., Meister, N., Chatterji, N., Ladhak, F., Hashimoto, T.B.: Proving test set contamination in black box language models. arXiv preprint arXiv:2310.17623 (2023)
25. Ouyang, L., et al.: Training language models to follow instructions with human feedback. Adv. Neural. Inf. Process. Syst. **35**, 27730–27744 (2022)
26. Pan, X., Zhang, M., Ji, S., Yang, M.: Privacy risks of general-purpose language models. In: 2020 IEEE Symposium on Security and Privacy (SP), pp. 1314–1331. IEEE (2020)
27. Schulman, J., et al.: Chatgpt: Optimizing language models for dialogue (2022)
28. Shi, W., et al.: Detecting pretraining data from large language models. arXiv preprint arXiv:2310.16789 (2023)
29. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE (2017)
30. Song, C., Raghunathan, A.: Information leakage in embedding models. In: Proceedings of the 2020 ACM SIGSAC, pp. 377–390 (2020)
31. Sonkar, S., Chen, X., Le, M., Liu, N., Basu Mallick, D., Baraniuk, R.: Code soliloquies for accurate calculations in large language models. In: Proceedings of the 14th Learning Analytics and Knowledge Conference: LAK 2024, pp. 828–835. Association for Computing Machinery, New York (2024)
32. Sonkar, S., Liu, N., Mallick, D., Baraniuk, R.: CLASS: a design framework for building intelligent tutoring systems based on learning science principles. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 1941–1961. Association for Computational Linguistics, Singapore (Dec (2023)
33. Steinke, T., Nasr, M., Jagielski, M.: Privacy auditing with one (1) training run. Advances in Neural Information Processing Systems **36** (2024)
34. Team, G., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

35. Touvron, H., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
36. Wikipedia contributors: Ai alignment — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/AI_alignment (2023), Accessed 13-May-2024
37. Yao, J., et al.: Machine unlearning of pre-trained large language models. arXiv preprint arXiv:2402.15159 (2024)



Distinguishing AI-Generated and Human-Written Text Through Psycholinguistic Analysis

Chidimma Opara(✉)

School of Computing, Engineering and Digital Technologies Teesside University, Middlesbrough, UK
c.opara@tees.ac.uk

Abstract. The increasing sophistication of AI-generated texts highlights the urgent need for accurate and transparent detection tools, especially in educational settings, where verifying authorship is essential. Existing literature has demonstrated that the application of stylometric features with machine learning classifiers can yield excellent results. Building on this foundation, this study proposes a comprehensive framework that integrates stylometric analysis with psycholinguistic theories, offering a clear and interpretable approach to distinguishing between AI-generated and human-written texts. This research specifically maps 31 distinct stylometric features to cognitive processes such as lexical retrieval, discourse planning, cognitive load management, and metacognitive self-monitoring. In doing so, it highlights the unique psycholinguistic patterns found in human writing. Through the intersection of computational linguistics and cognitive science, this framework contributes to the development of reliable tools aimed at preserving academic integrity in the era of generative AI.

Keywords: AI generated text · stylometry · psycholinguistics · plagiarism detection · authorship verification

1 Introduction

AI-generated texts, powered by advanced models such as Openai's GPT series, have significantly reshaped content creation, enabling artificial intelligence to produce text that closely resembles human writing. While these systems enhance efficiency and creativity, such as improving learning experiences, they also introduce new challenges in distinguishing AI-generated content from human-authored work, particularly in education, where written assignments serve as a key measure of critical thinking and subject expertise [2]. In response to these challenges, recent state-of-the-art research has focused on AI-text detection using deep learning techniques, including Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory networks, and Transformer-based models such as RoBERTa and GPT. These models have demonstrated excellent performance in classification tasks [15]. However, they are often criticised for their "black box" nature, offering high accuracy without transparent reasoning.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 212–219, 2025.
https://doi.org/10.1007/978-3-031-98462-4_27

A more interpretable and linguistically grounded approach is **stylometry**, the quantitative analysis of writing style, traditionally used in literary authorship attribution. Stylometric techniques analyse measurable features such as lexical diversity, sentence complexity, and readability to identify patterns indicative of authorship [10,13]. Although stylometric features often yield high classification accuracy when applied on machine learning algorithms, they typically lack explanatory capacity regarding the cognitive processes that give rise to these features in human writing.

Understanding human authorship requires a deeper engagement with the **psycholinguistic** mechanisms that shape written expression. Stylometric features in human-authored texts are not arbitrary; they reflect complex cognitive activities such as self-monitoring, discourse planning, and lexical retrieval [8]. For instance, higher syntactic complexity may indicate intentional structuring of ideas. By aligning stylometric analysis with psycholinguistic theories, it becomes possible to develop more interpretable and cognitively informed approaches to distinguishing human-authored and AI-generated text.

1.1 Research Contribution

This study builds upon prior research [13] and extends the AI-generated text detection model (StyloAI) by grounding stylometric features in psycholinguistic theory. 31 stylometric features are systematically mapped to cognitive and linguistic processes, including Cognitive Load Theory, Metacognition and Self-Monitoring, Lexical Access and Retrieval, and Discourse Planning and Cohesion. This theoretical alignment enhances the understanding of why each of the stylometric features serves as an effective discriminator between human and AI-generated writing.

2 Psycholinguistic Theories in Human and AI-Generated Writing

2.1 Cognitive Load Theory

Cognitive Load Theory (CLT), introduced by Sweller in 1988 [18], posits that human working memory has limited capacity. When writing tasks overwhelm this capacity, through the simultaneous demands of idea generation, sentence construction, and argument organisation, performance declines and error likelihood increases [1]. This strain manifests in natural markers such as pauses, revisions, and stylistic fluctuations, which contribute to the uniqueness of human writing. For instance, under increased cognitive load, writers may simplify sentence structures or omit details [7]. These features are detectable through stylometric analysis and offer insight into the cognitive processes involved in human authorship.

By contrast, AI systems do not experience cognitive load in the human sense. Large Language Models (LLMs) generate coherent, fluent text without pausing, self-monitoring, or strategic planning [17]. While AI is bound by computational

limits, these do not mirror the constraints of human cognition. As a result, AI-generated writing tends to be polished and consistent, yet often lacks the subtle imperfections typical of human expression. Efforts to simulate human-like limitations, such as bounded pragmatic speaker models [11], remain artificial and do not replicate authentic cognitive processes.

2.2 Metacognition and Self-monitoring

Metacognition, as defined by Flavell (1979) [6], refers to one's ability to reflect on and regulate cognitive processes. In writing, this includes revising content to improve clarity, coherence, and alignment with audience expectations. Such self-monitoring is evident in abrupt stylistic changes, iterative corrections, and context-sensitive modifications during drafting.

AI models, by contrast, lack fundamental metacognitive capabilities. Although models like GPT-4 can simulate reasoning through techniques such as chain-of-thought prompting [4,9], they do not engage in conscious reflection or deliberate revision. Instead, their refinements are statistical, not intentional. This limitation is especially evident in phenomena like hallucinations, overconfident but inaccurate outputs [16]. Unlike AI, human authors revise to correct meaning, adjust tone, or refine structure, making self-monitoring a key psycholinguistic marker of human authorship [19].

2.3 Lexical Access and Retrieval

Word selection in human writing involves a complex interplay between semantics, grammar, memory, and personal experience [5]. This cognitive process can result in occasional retrieval errors, such as tip-of-the-tongue moments or unintended repetition, often caught and corrected through self-monitoring. These subtle markers reflect real-time mental effort.

AI models, however, select words based on probabilistic associations derived from training data. While capable of producing text that mimics emotional richness, AI lacks lived experience and contextual depth, limiting the authenticity of its lexical choices [9]. Neuroimaging studies confirm that human writing activates brain areas related to memory and emotion dimensions AI cannot replicate [7,20]. Though AI may demonstrate lexical diversity, it often struggles to maintain tonal consistency and semantic cohesion over longer passages [4], making lexical variation another useful marker for authorship attribution.

2.4 Discourse Planning and Cohesion

Effective writing requires logical progression of ideas, as discussed by Halliday and Hasan (1976) [3]. Human writers use cohesive devices such as pronouns, transitions, and structured argumentation to build narratives hierarchically: from broad ideas to fine-grained sentence-level refinements. Yet, under cognitive load, this planning may falter, leading to abrupt topic changes or weak transitions.

AI-generated texts emulate coherence via statistical predictions rather than conceptual planning [9,20]. While initially appearing structured, these texts often contain contradictions or repetitiveness due to shallow contextual modelling [12]. In contrast, human writing typically reflects more intentional and consistent discourse planning, making cohesion and logical progression key differentiators [4].

3 Framework Mapping Stylometric Features to Psycholinguistic Theories

This section presents a systematic analysis of writing by integrating stylometric features with psycholinguistic theory.

The StyloAI model proposed in the previous work [13] incorporates 31 stylometric features, including 12 novel metrics specifically developed for detecting AI-generated texts. These features are organised into six distinct categories: Lexical Diversity, Syntactic Complexity, Sentiment and Subjectivity, Readability, Named Entities, and Uniqueness and Variety. For an extensive discussion on the theoretical rationale and detailed feature descriptions, readers are referred to this Table

Table 1 summarises the mapping of 18 out of the 31 of these features to psycholinguistic theories.

Table 1. Mapping Stylometric Features to Psycholinguistic Theories.

Stylometric Category	Psycholinguistic Category	Features
Lexical Features	Cognitive Load	word_count
	Metacognition & Self-Monitoring	hapax_legomenon_rate
	Lexical Access & Retrieval	unique_word_count, ttr, avg_word_length, hapax_legomenon_rate
	Discourse Planning & Cohesion	char_count
Syntactic Features	Cognitive Load	avg_sentence_length, complex_sentence_count
	Metacognition & Self-Monitoring	punctuation_count, question_count, contraction_count
	Discourse Planning & Cohesion	complex_sentence_count
Uniqueness Features	Discourse Planning & Cohesion	syntax_variety
Sentiment Features	Cognitive Load	subjectivity
	Metacognition & Self-Monitoring	emotion_word_count, vader_compound
Readability Features	Cognitive Load	gunning_fog
	Metacognition & Self-Monitoring	gunning_fog
	Discourse Planning & Cohesion	flesch_reading_ease
Named Entity Features	Metacognition & Self-Monitoring	first_person_count
	Discourse Planning & Cohesion	direct_address_count

3.1 Detailed Feature Mapping

Lexical Features encompass the fundamental elements of language, words, and their usage patterns. This category includes metrics such as total word count, vocabulary diversity, average word length, and the frequency of unique words. In academic contexts, these features provide insight into cognitive processes such as lexical retrieval, semantic depth, and the ability to structure and integrate knowledge effectively [10,14].

For instance, the following is an excerpt of a human-written text from the dataset used in the StyloAI model, that statistically has a high value of unique word count:

```
"as you suspect a digestive problem in your rabbit, take him
to your veterinarian immediately. If your rabbit has
diarrhea, your veterinarian will test the feces to identify
the specific organism (e.g., Clostridium)"
```

Psycholinguistic Interpretation: Human authors draw on semantic memory, lived experiences, and an awareness of audience expectations when choosing words. This often results in greater vocabulary diversity, precise language use, and contextually refined phrasing. Using stylometry, these behaviours are captured in features such as *Unique Word Count*, *TTR* and *Hapax Legomenon rate*. Psycholinguistically, these behaviours reflect active *lexical retrieval and metacognitive self-monitoring.* In contrast, AI models, although capable of statistically simulating lexical variation, often struggle to produce contextualised or original word choices, especially across longer passages.

Syntactic Features: Syntactic features examine sentence structure and grammatical complexity. Sentence construction places demands on working memory and discourse organisation, making it a key area where human and AI writing differ [18].

Psycholinguistic Interpretation: Human writers demonstrate flexibility in managing syntactic structure, tailoring sentence length, punctuation, and phrasing based on communicative goals such as clarity, emphasis, or engagement. These adjustments are enabled by all psycholinguistic theories discussed in Sect. 2, especially *self-monitoring* and *discourse planning* processes that respond dynamically to context and audience. In contrast, while syntactically fluent, AI-generated texts often exhibit a more rigid, repetitive structure. Without intrinsic self-monitoring, AI cannot vary syntax for rhetorical or communicative effect. As a result, its writing tends to lack the stylistic modulation typical of human authorship.

Sentiment Features: Sentiment analysis evaluates the emotional tone conveyed in a text. For example, the *Emotion Word Count*, which counts the total number of words associated with emotions in the text, is one of the stylometric

features extracted in this category. In academic writing, expressions of enthusiasm, critique, or personal reflection can indicate human cognition [9]. While AI models can mimic emotional language, their expressions typically lack intrinsic emotional depth and intentionality, often resulting in mechanical or contextually misplaced language.

For instance, the following is a generated text from the dataset used in the StyloAI model, demonstrating a mechanical expression of emotion:

> ‘‘My car wont start, it doesnt turn over, it gives a clicking sound, keeps turning over but wont start, HELP.’’

The excerpt demonstrates the AI's attempt to convey emotion through specific linguistic choices. The use of "HELP" adds urgency and signals distress. The short, fragmented phrases create a choppy rhythm that mirrors frustration, while the repetition of "won't start" emphasises the speaker's emotional intensity and sense of helplessness.

Psycholinguistic Interpretation: Human authors naturally embed emotional layers in their writing, drawing from lived experience, empathy, and reflective processes. These expressions evolve organically within the text, drawing on *metacognition* and *cognitive load*, showing contextual variation and dynamic sentiment progression. An example of human-written text from the dataset used in the StyloAI model, that has a high emotion word count, but embeds it in layers:

> ‘‘If you find yourself struggling with patience as you work to change your reputation, remind yourself that your reputation isnt either mature or immature like a light switch is on or off, but is instead on a continuum that varies in maturity...’’

In contrast, AI-generated sentiment is based on statistical patterns rather than experiential grounding. As a result, emotional expressions in AI texts often appear flat, stereotyped, or overly uniform.

Readability Features. Readability metrics assess the complexity and accessibility of a text, often combining indicators of lexical richness and syntactic structure. For example, *Gunning fog*, which estimates the years of formal education needed to understand a text on the first reading, is one of the stylometric features extracted in this category. These features offer insight into a writer's ability to adapt language for clarity and audience comprehension [10].

Psycholinguistic Interpretation: Human authors regulate text complexity through deliberate planning, ongoing self-monitoring, and revision, balancing accessibility with depth. These processes reflect *cognitive loadmetacognitive awareness*, *discourse-level planning*. In contrast, AI lacks true audience modelling capabilities and may produce writing that alternates unpredictably between overly dense and overly simplistic phrasing, without a clear rhetorical strategy.

Named Entity Features. Named entities such as people, organisations, locations, and dates offer important contextual clues about how authors ground their writing in real-world knowledge, personal experience, or domain-specific understanding [10]. Human writers often incorporate actual references, drawing from memory and context. In contrast, AI-generated texts may include fabricated, vague, or inconsistently integrated entities.

Psycholinguistic Interpretation: Human-authored writing reflects deliberate discourse planning, often integrating references to real events, individuals, or timelines. These references emerge from episodic memory, thematic coherence, and personal narrative strategies. These characteristics reference *metacognition and discourse planning cognitive theories*. AI models, while capable of mimicking named entities, lack experiential grounding and may introduce irrelevant or inaccurate references due to limitations in factual recall or context awareness.

Uniqueness Features. Uniqueness refers to how distinct a piece of writing is from commonly observed linguistic patterns. This may involve novel word combinations (e.g., bigrams or trigrams), varied sentence structures, or stylistic originality.

Psycholinguistic Interpretation: Original expression in human writing is underpinned by semantic memory, flexible lexical retrieval, and the ability to adapt tone and structure to audience or purpose. These unique characteristics demonstrate *lexical retrieval and discourse planning cognitive theories*. AI models, even when prompted for variety, generate phrases that are statistically probable but rhetorically shallow, lacking personal voice or adaptive strategy.

4 Conclusion

This study demonstrates that integrating stylometric analysis with psycholinguistic theory offers an interpretable framework for distinguishing AI-generated from human-authored texts. By mapping 31 stylometric features to underlying cognitive processes, such as lexical retrieval, discourse planning, cognitive load management, and metacognitive self-monitoring, this research enhances the understanding of the psycholinguistic signatures embedded in human writing.

References

1. Baddeley, A.: Working memory: looking back and looking forward. Nat. Rev. Neurosci. **4**(10), 829–839 (2003)
2. Bailey, R.A.: Undergraduate students' perceptions the role and utility of written assessment feedback. J. Learn. Developm. Higher Educ. (1) (2009)
3. Carrell, P.L.: Cohesion is not coherence. TESOL Q. **16**(4), 479–488 (1982)

4. Collins, K.M., Wong, C., Feng, J., Wei, M., Tenenbaum, J.: Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 44 (2022)
5. Dell, G.S., O'Seaghdha, P.G.: Stages of lexical access in language production. Cognition **42**(1–3), 287–314 (1992)
6. Flavell, J.: Theories of learning in educational psychology. Am. Psychol. **34**, 906–911 (1979)
7. Goldstein, A., et al.: Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. **25**(3), 369–380 (2022)
8. Huang, B., Chen, C., Shu, K.: Authorship attribution in the era of llms: problems, methodologies, and challenges. ACM SIGKDD Explorations Newsl **26**(2), 21–43 (2025)
9. Mahowald, K., Ivanova, A.A., Blank, I.A., Kanwisher, N., Tenenbaum, J.B., Fedorenko, E.: Dissociating language and thought in large language models. Trends Cognitive Sci. (2024)
10. Mindner, L., Schlippe, T., Schaaff, K.: Classification of human-and ai-generated texts: Investigating features for chatgpt. In: International Conference on Artificial Intelligence in Education Technology. pp. 152–170. Springer (2023). https://doi.org/10.1007/978-981-99-7947-9_12
11. Nguyen, K.X.: Language models are bounded pragmatic speakers. In: First Workshop on Theory of Mind in Communicating Agents (2023)
12. Nolfi, S.: On the unexpected abilities of large language models. Adapt. Behav. **32**(6), 493–502 (2024)
13. Opara, C.: Styloai: distinguishing ai-generated content with stylometric analysis. In: International Conference on Artificial Intelligence in Education, pp. 105–114. Springer (2024). https://doi.org/10.1007/978-3-031-64312-5_13
14. Paas, F., van Merriënboer, J.J.: Cognitive-load theory: methods to manage working memory load in the learning of complex tasks. Curr. Dir. Psychol. Sci. **29**(4), 394–398 (2020)
15. Sarvazyan, A.M., González, J.Á., Franco-Salvador, M., Rangel, F., Chulvi, B., Rosso, P.: Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. arXiv preprint arXiv:2309.11285 (2023)
16. Smith, A.L., Greaves, F., Panch, T.: Hallucination or confabulation? neuroanatomy as metaphor in large language models. PLOS Digital Health **2**(11), e0000388 (2023)
17. Suresh, S., Mukherjee, K., Yu, X., Huang, W.C., Padua, L., Rogers, T.T.: Conceptual structure coheres in human cognition but not in large language models. arXiv preprint arXiv:2304.02754 (2023)
18. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. Learn. Instr. **4**(4), 295–312 (1994)
19. Talboy, A.N., Fuller, E.: Challenging the appearance of machine intelligence: Cognitive bias in llms and best practices for adoption. arXiv e-prints pp. arXiv–2304 (2023)
20. Tuckute, G., Kanwisher, N., Fedorenko, E.: Language in brains, minds, and machines. Annual Rev. Neurosci. **47** (2024)

Improving Open-Response Assessment with LearnLM

Danielle R. Thomas[1(✉)], Conrad Borchers[1], Shambhavi Bhushan[1], Sanjit Kakarla[1], Alex Houk[1], Ralph Abboud[2], Shivang Gupta[1], Erin Gatz[1], and Kenneth R. Koedinger[1]

[1] Carnegie Mellon University, Pittsburgh, USA
{drthomas,cborchers,shivangg,koedinger}@cmu.edu,
{shambhab,sanjitk,ahouk,egatz}@andrew.cmu.edu
[2] Learning Engineering Virtual Institute, Oxford, UK
rabboud@levimath.org

Abstract. Automated grading of learners' open responses remains challenging due to the complexity of language and the subjective nature of human evaluation. Recent advances in generative AI, particularly large language models (LLMs), offer new possibilities to improve assessment. Off-the-shelf LLMs, such as GPT-4, have been applied to this task, as well as dedicated education-oriented models, such as LearnLM. However, little is known about their effectiveness compared to general-purpose models. In this study, we evaluate GPT-4o, GPT-4-turbo, and Gemini-Pro and compare their performance to LearnLM to determine their effectiveness in assessing learning, specifically the professional development of adult tutors. We find that LearnLM outperforms other models on tasks requiring tutor learners to predict the most appropriate response to students. We hypothesize that this is due to the model's fine-tuning on tutor-student interaction data and suggest that LearnLM may be particularly useful in scenario-based tutor training. To further improve automated assessment methods, we challenge the concept of human "ground truth" by proposing alternative validation methods. Specifically, we introduce a predictive validity method by relating open-response scores with corresponding multiple-choice scores that demonstrate statistically significant and moderate correlations, particularly with LearnLM. Our novel method demonstrates predictive validity but should be combined with additional measures to ensure a more comprehensive assessment. This study contributes an open source dataset, human annotation rubrics, and LLM prompts, to improve future assessment applications of LLMs.

Keywords: Generative AI · LLMs · LearnLM · Assessment

1 Introduction and Related Work

Automated grading of learners' textual responses has long been a challenging task due to the complexity of language, context, and the subjective nature of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 220–227, 2025.
https://doi.org/10.1007/978-3-031-98462-4_28

human judgment [6,10,17]. Recent breakthroughs in generative AI, particularly large language models (LLMs), have improved automated assessment. These models, such as GPT-4o [1], GPT-4-turbo, and Gemini-Pro [18], offer scalable solutions to evaluate textual responses. Recently, LearnLM [11], a new family of Gemini models fine-tuned for learning, has been developed to assist with education-related tasks. Fine-tuned for pedagogical instruction from both synthetic and human-written datasets, LearnLM shows promise with results preferred by expert raters over GPT-4o and Gemini-Pro models across several learning scenarios, such as a tutor helping a student solve a math problem [19]. Despite promising preliminary results, the LearnLM's ability to assess pedagogical interaction more broadly, such as equity-related tasks and adult tutor learning, as opposed to student learning and content-related tutoring, is underexplored.

Further Validation of LearnLM. Recent advances in generative AI have significantly influenced the development of specialized models aimed at improving educational outcomes. Introduced by Google in 2024, LearnLM–specifically `learnlm-1.5-pro-experimental`–is a pedagogically fine-tuned extension of the Gemini 1.5 Pro model, designed to emulate behaviors associated with effective human tutors [2]. Its core contribution lies in its ability to follow nuanced pedagogical instructions, enabling targeted responses aligned with specific educational objectives [11,19]. LearnLM has been evaluated through multistage human assessments, where pedagogy experts simulated learner interactions with various AI systems. These evaluations consistently demonstrated that LearnLM outperformed contemporaneous models like GPT-4o, Claude 3.5, and Gemini 1.5 Pro in areas such as adapting to learners' needs, managing cognitive load, and promoting active learning [2,19]. Experts preferred LearnLM across varying scenarios (e.g., a shy student trying to practice addition or a disinterested student wanting a quick response) with mean performance strength 31% higher than GPT-4o and 13% more compared to Gemini 1.5 Pro [19]. Despite promising results, the LearnLM team [19] acknowledges the limitations of their evaluation and invites further extrinsic validation, which this work aims to address.

Revisiting Human Coding as Ground Truth. The reliability of human grading as the "gold standard" for evaluating learners' responses has limitations: inconsistency, subjective interpretation, and bias [3,8,15]. Studies reveal significant variability in human grading practices, often due to poor inter-rater reliability and inconsistent self-grading among human judges [8,15]. Individual biases remain even when rubrics are present. The emerging trend of using LLM-as-a-judge [26] has introduced a novel framework for evaluating performance in open-ended responses without relying on predefined human annotations. However, this approach introduces its own bias problems [9,13,14], sparking the urgency to develop alternative validation frameworks that reduce dependence on subjective human judgment. To address this limitation, we draw from predictive validity, which is the degree to which a test or assessment accurately forecasts or correlates with future performance on a related measure [23]. In this context, we use multiple-choice questions (MCQs) that assess the same learning objectives as open-response questions. Predictive validity evaluates whether performance

on MCQs can reliably predict outcomes on open-response questions or other measures of learner understanding. MCQs are known to be effective and more efficient than open response tasks when instructional time is limited [7,16,21], in addition to being more objective than human scoring of open responses. This approach ensures MCQs serve as effective proxies for more complex assessments.

In this study, we address these gaps by evaluating the performance of several LLM models and comparing their absolute performance to LearnLM to determine LearnLM's ability to generalize to tangential tasks beyond its fine-tuning. We also propose an alternative method to humans as "ground truth." We hypothesize that the fine-tuned model LearnLM will be better at assessing textual responses in a way that is aligned with tutor learning due to its fine-tuning and aim to demonstrate a more objective and efficient method to evaluate its performance. This work also contributes an open source dataset and LLM prompts. We ask: **RQ1:** How does LearnLM perform compared to GPT-4o, GPT-4 turbo, and Gemini Pro on adult tutor learning and equity-related tasks? **RQ2:** How does the alternative validation method of using predictive validity by mapping to the corresponding MCQs (instead of human-scored open response questions as traditional ground truth) compare in evaluating LLM performance?

2 Methods

Participants, Setting, and Lesson Design. The open-source dataset was derived from 81 college students employed as paid tutors for a remote tutoring organization. The lesson was delivered through an online tutoring platform, PLUS.[1] Tutor participants accessed the lesson through the PLUS tutoring app in their own time and submitted their constructed responses through open-text boxes. The focal lesson for this present work, titled *Helping Students Manage Inequities*, was developed by a university research team specializing in learning sciences. The lesson trains human tutors to support students facing various inequities, such as a student not having access to the internet at home to do their homework or a student not being able to hear the teacher during class instruction because they sit in the back of the classroom [20]. All data were anonymously logged, in line with the approved IRB protocol and consisted of responses to a pretest and posttest tutoring scenario where participants were asked to *predict* the best response within an open-response question (i.e., *"What would you say to the student..."*), followed by an MCQ (i.e., *"Which of the following tutor responses..."*). Then, participants were asked to *explain* their rationale in an open-response question (i.e., *"Why did you choose to..."*) and a MCQ (i.e., *"Which of the following explains your rationale..."*). All participants' responses were binary coded as incorrect or correct (0/1) by two human researchers. The interrater reliability for the *predict* open responses was 89% agreement ($\kappa = 0.75$) and 87% agreement ($\kappa = 0.75$) for *explain* open responses.

Comparing LearnLM Performance and Predictive Validity. To evaluate the performance of various models in scoring learners' responses, we

[1] https://www.tutors.plus/.

used a quantitative comparison based on key metrics: accuracy, AUC, and F1 score. The four models used and their specific model strings were GPT-4o (`gpt-4o-2024-11-20`), GPT-4-turbo (`gpt-4-turbo-2024-04-09`), Gemini 1.5 Pro (`gemini-1.5-pro-002`), and the experimental Gemini model LearnLM (`learnlm-1.5-pro-experimental`). Each model assessed both open-response question types: *predicting* the best response and *explaining* the rationale behind that response. Few-shot prompting was applied uniformly across all models, with temperature settings adjusted for select models to observe variability in performance [5]. In this context, temperature is a parameter that controls the randomness of the model's responses—lower values make outputs more deterministic, while higher values increase variability in text generation. Few-shot prompting has been found to outperform zero-shot methods in nuanced lesson scenarios [21,22]. Chain-of-thought prompting was used that required the models to provide rationale for their scores [24]. The dataset, lesson, and LLM system prompts are available via GitHub.[2]

To evaluate the predictive validity of the LLM open response scores on MCQ scores, we calculated the correlation of participants' MCQ scores (0–2 pts) and their LLM scores on open-ended responses (0–2 pts). We also performed correlational analysis of the higher performing LLMs with MCQ scores to determine predictive validity on better performing models.

3 Results and Discussion

RQ1: Model Performance Comparisons with LearnLM. Table 1 summarizes model performance in scoring learners' responses in *predicting* the best response and *explaining* their rationale, with accuracy, AUC, and F1 score as evaluation metrics. Confidence intervals were obtained via bootstrapping. The better performing models are shown in bold. GPT-4o demonstrates the highest overall performance, with accuracies of 0.84 and 0.85 for *predict* and *explain* open responses, respectively. GPT-4-turbo lags behind with significantly lower accuracy for assessing *predict* questions (0.62), but improves substantially for *explain* question (0.78). Gemini-pro-1.5 shows consistent performance, with accuracy and F1 scores improving as temperature increases, peaking at temp=1. LearnLM performs well, second to GPT-4o, achieving accuracy in the low-80s across both tasks, with stable AUC and F1 scores. Temperature adjustments for LearnLM show minimal impact on performance compared to Gemini-pro.

Figure 1 illustrates the pretest and posttest scores, two MCQs and two open-response questions, for a total of four points each at temperature 0. In line with previous work, human scores were generally higher for the pretest and posttest compared to all LLMs, demonstrating leniency among human graders [4]. Although LearnLM performed well, particularly on assessing learners' open-response tasks involving prediction of the best approach, this is not visually apparent due to its comparatively lower performance on evaluating learners' open-response tasks involving explanation of their chosen approach.

[2] https://github.com/conradborchers/learnLM-open-response.

Table 1. Performance against human scores with 95% confidence intervals. All temperatures at 0, unless otherwise specified. Better performing models in bold.

Model	Predicting the best approach			Explaining their rationale		
	Accuracy	AUC	F1	Accuracy	AUC	F1
`GPT-4o`	**0.84** [0.79, 0.89]	**0.87** [0.82, 0.92]	**0.88** [0.83, 0.92]	**0.85** [0.79, 0.90]	**0.86** [0.81, 0.91]	**0.85** [0.79, 0.90]
`GPT-4-turbo`	0.63 [0.57, 0.7]	0.73 [0.69, 0.78]	0.65 [0.57, 0.72]	0.78 [0.72, 0.83]	0.79 [0.74, 0.85]	0.75 [0.66, 0.82]
`Gemini-Pro-1.5`	0.75 [0.68, 0.82]	0.74 [0.67, 0.81]	0.82 [0.76, 0.87]	0.84 [0.78, 0.89]	0.84 [0.79, 0.89]	0.85 [0.79, 0.90]
`Gemini-Pro-1.5`, T=1	0.75 [0.68, 0.81]	0.74 [0.66, 0.81]	0.81 [0.75, 0.86]	**0.87** [0.82, 0.92]	**0.87** [0.82, 0.92]	**0.88** [0.82, 0.93]
`Gemini-Pro-1.5`, T=2	0.73 [0.67, 0.79]	0.72 [0.65, 0.79]	0.80 [0.74, 0.85]	0.86 [0.81, 0.91]	0.87 [0.81, 0.92]	0.87 [0.81, 0.92]
`LearnLM`	**0.82** [0.76, 0.87]	**0.82** [0.76, 0.88]	**0.86** [0.81, 0.91]	0.83 [0.77, 0.88]	0.84 [0.79, 0.89]	0.82 [0.76, 0.88]
`LearnLM`, T=1	0.79 [0.73, 0.85]	0.78 [0.71, 0.85]	0.85 [0.79, 0.89]	0.82 [0.77, 0.88]	0.84 [0.79, 0.88]	0.82 [0.75, 0.87]
`LearnLM`, T=2	0.81 [0.75, 0.87]	0.79 [0.72, 0.86]	0.86 [0.81, 0.91]	0.81 [0.75, 0.86]	0.82 [0.77, 0.87]	0.80 [0.72, 0.86]

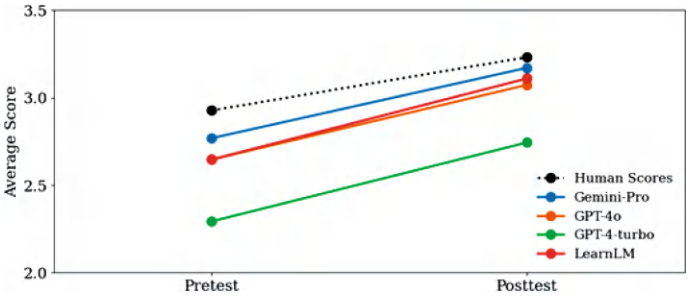


Fig. 1. Comparison of average human and LLM scores at pre/posttest (T=0).

Referring to Table 1, LearnLM performs considerably better than other models, aside from GPT-4o, on assessing responses where tutors *predict the best approach*. For instance, the learner response of what to say to a student who is experiencing an educational inequity is as follows: *"I would discuss how Alexis could present her problem to the teacher"* is an incorrect response as determined by human-expert graders as it does not actively state *what* the tutor will say to the student. Similarly, LearnLM (few-shot and zero-shot prompting at temperatures of 0, 1, and 2) effectively scored the response as incorrect. However, GPT-4o, GPT-4-turbo, and Gemini-Pro-1.5 all ineffectively scored the response as correct. In contrast, the response, *"I am sorry you are going through this. Is there any way we can present this to your teacher in a way she provides an alternative?"* was scored correct by human experts and the LearnLM models, with the remaining models scoring it incorrectly, giving the response a 0. Several more

incidences occurred where human truth and LearnLM agreed in opposition to the remaining models. We hypothesize that LearnLM's strong performance in approach prediction stems from its more curated instruction fine-tuning, which aligns strongly with following pre-set pedagogical practices. In contrast, we conjecture that this same fine-tuning also explains LearnLM's less competitive performance on explaining tutor rationale, as LearnLM, compared to its initial generally fine-tuned version, has explicitly been trained to avoid making explicit pedagogical choices, making it less adept at justifying approaches compared to following them. There were 26 learner open responses where LearnLM (few shot, T=0) and human truth were in agreement and GPT-4o (few shot, T=0) was in disagreement, with the majority related to questions where learners had to *predict* the best response. More specifically, there were seven incidences among the 352 open responses in which human truth and LearnLM were in agreement and the other LLMs (few-shot, T=0) were in opposition. All responses, except one, were prediction tasks and scored as correct by humans and LearnLM.

RQ2: Alternative Validation Approaches to Human as Ground Truth. The analysis revealed a significant, positive correlation between MCQ scores and human-graded open-response scores, $r(86) = 0.42, p < .001$. While this correlation is statistically significant, it is not particularly large. Potential alternative contributors to the moderate size of this correlation are (1) the few test items, just two each, used to produce the MCQ and open-response scores and (2) correlation strength is limited by the imperfect reliability and ambiguity of human grading of open responses. To compare this correlation to the validity of LLM assessment, we computed the correlation between MCQ scores and LLM scores. We correlated MCQ scores with GPT-4o-scored open responses, yielding $r(86) = 0.41, p < .001$; and MCQ scores with LearnLM-scored open responses, yielding $r(86) = 0.48, p < .001$. This descriptive evidence of a difference did not reach statistical significance, z = -1.04, p = .297 but merits further investigation.

We find that LLM scores have significant predictive validity and *this validity can be determined without open-ended grading by humans.* Notably, the LearnLM correlation of 0.48 is 0.06 higher than the human-scored correlation of 0.42. In other words, with this tighter comparison, we find that the predictive validity of the LLM scoring is comparable to that of human scoring. This result supports predictive validity as a complementary method for evaluating model performance, though it should be combined with additional measures to ensure a more comprehensive assessment. Other alternative approaches to using human scores as "ground truth" include: using the average score among several adversarial models (e.g., LearnLM vs. GPT-4o) to establish reliability rather than comparing human judgments; and applying LLM self-consistency measures by comparing evaluations across different prompts to check robustness.

4 Implications, Future Work, and Conclusion

This work presents two main implications for the use of LLMs in education: (1) insight into the boundaries of tasks for which LearnLM is best suited, and

(2) the introduction of predictive validity as a complementary method to using human scoring as ground truth for evaluating open responses. Future research will integrate these implications by using LearnLM to score open responses with much larger sample sizes and across a variety of tutor training lessons. In addition, future work will involve the use of ensemble methods such as bagging and boosting to leverage strengths of different LLMs [12,25]. Our findings reveal that LearnLM can match or even outperform other models on tasks that require tutors to interact with students in situational contexts, especially when they require the generation of specific tutor dialogue moves. Similarly, LearnLM can match or even outperform human scorers using predictive validity methods showing potential for scaling the intricate task of open-response grading.

Acknowledgments. This work was made possible with the support of the Learning Engineering Virtual Institute. The opinions, findings and conclusions expressed in this material are those of the authors.

References

1. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. AI, G.: Learnlm | gemini api | google ai for developers. https://ai.google.dev/gemini-api/docs/learnlm, Accessed 18 Feb 2025
3. Andrade, H.G.: Teaching with rubrics: the good, the bad, and the ugly. Coll. Teach. **53**(1), 27–31 (2005)
4. Awidi, I.T.: Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (ai) tool. Comput. Educ. Artifi. Intell. **6**, 100226 (2024)
5. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
6. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. Int. J. Artif. Intell. Educ. **25**, 60–117 (2015)
7. Butler, A.C.: Multiple-choice testing in education: are the best practices for assessment also good for learning? J. Appl. Res. Mem. Cogn. **7**(3), 323–331 (2018)
8. Chen, G.H., Chen, S., Liu, Z., Jiang, F., Wang, B.: Humans or llms as the judge? a study on judgement biases. arXiv preprint arXiv:2402.10669 (2024)
9. Gallegos, I.O., et al.: Bias and fairness in large language models: a survey. Comput. Linguist., 1–79 (2024)
10. Haller, S., Aldea, A., Seifert, C., Strisciuglio, N.: Survey on automated short answer grading with deep learning: from word embeddings to transformers. arXiv preprint arXiv:2204.03503 (2022)
11. Jurenka, I., et al.: Towards responsible development of generative ai for education: An evaluation-driven approach. arXiv preprint arXiv:2407.12687 (2024)
12. Lee, G.G., Latif, E., Wu, X., Liu, N., Zhai, X.: Applying large language models and chain-of-thought for automatic scoring. Comput. Educ. Artifi. Intell. **6**, 100213 (2024)
13. Lee, J., Hicke, Y., Yu, R., Brooks, C., Kizilcec, R.F.: The life cycle of large language models in education: a framework for understanding sources of bias. Br. J. Edu. Technol. **55**(5), 1982–2002 (2024)

14. Liang, P.P., Wu, C., Morency, L.P., Salakhutdinov, R.: Towards understanding and mitigating social biases in language models. In: International Conference on Machine Learning, pp. 6565–6576. PMLR (2021)
15. Messer, M., Brown, N.C., Kölling, M., Shi, M.: How consistent are humans when grading programming assignments? arXiv preprint arXiv:2409.12967 (2024)
16. Moore, S., Bier, N., Stamper, J.: Assessing educational quality: Comparative analysis of crowdsourced, expert, and ai-driven rubric applications. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 12, pp. 115–125 (2024)
17. Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. Artif. Intell. Rev. **55**(3), 2495–2527 (2022)
18. Team, G., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
19. Team, L., et al.: Learnlm: Improving gemini for learning. arXiv preprint arXiv:2412.16429 (2024)
20. Thomas, D., Yang, X., Gupta, S., Adeniran, A., Mclaughlin, E., Koedinger, K.: When the tutor becomes the student: Design and evaluation of efficient scenario-based lessons for tutors. In: LAK23: 13th International Learning Analytics and Knowledge Conference, pp. 250–261 (2023)
21. Thomas, D.R., et al.: Do tutors learn from equity training and can generative ai assess it? In: Proceedings of the 15th International Learning Analytics and Knowledge Conference, pp. 505–515 (2025)
22. Thomas, D.R., et al.: Does multiple choice have a future in the age of generative ai? a posttest-only rct. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference, pp. 494–504 (2025)
23. Trochim, W.M., Donnelly, J.P., Arora, K.: Research methods: The essential knowledge base (2016)
24. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
25. Zhai, X., He, P., Krajcik, J.: Applying machine learning to automatically assess scientific models. J. Res. Sci. Teach. **59**(10), 1765–1794 (2022)
26. Zheng, L., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Adv. Neural. Inf. Process. Syst. **36**, 46595–46623 (2023)



Multimodal Classroom Diarization with GPT Re-scoring: Teacher or Student?

Matthew Perez(✉), Berk Coker, Kemal Berk Kocabagli, Jessica Vitale, and Alyssa Van Camp

TeachFX, Menlo Park, CA 94025, USA
{matt,berk,berkkocabagli,jessicaboyle,alyssa}@teachfx.com
https://teachfx.com/

Abstract. Speaker diarization, determining "who spoke what", is a critical component of automated, speech-based educational technologies. However, in single-microphone classroom environments, there are many challenges that can make diarization difficult such as background noise, speaker overlap, and interjections. Traditional audio-only diarization methods often fail to capture the nuanced conversational dynamics of a classroom, where the linguistic patterns and discourse styles can provide important context into educator role (i.e. teacher/student). In this paper, we propose a multimodal pipeline that integrates Large Language Model (LLM) re-scoring to enhance classroom speaker diarization. We explore two improvements, the first of which involves Adaptive Centroid Enrollment (ACE), which is able to re-structure the audio embedding space using a cluster taken from the recording itself. The second, involves finetuning generative pretrained transformers to refine clustering predictions using transcriptions associated with each audio segment. Experimental results on our internal classroom dataset reveal relative improvements of 8.2% in word-level accuracy and 16.5% in word-level unweighted F1-score, highlighting the potential of LLM-based approaches for addressing speaker diarization challenges in educational settings.

Keywords: speaker diarization · LLM re-scoring · GPT in education · classroom speaker clustering

1 Introduction

Teachers require ongoing, high-quality feedback on their practice to provide all students the academic, social-emotional, and behavioral support they need to learn and grow. However, providing such feedback is resource-intensive, typically requiring trained staff, classroom observations, data analysis, and personalized coaching. Tools that employ artificial intelligence to analyze classroom recordings have the potential to expand teachers' access to high-quality feedback and improve instruction at scale. For instance, TeachFX is a widely available automated teacher feedback tool, with over 4,000 teachers who make classroom recordings and receive automated feedback on their instruction and student engagement.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 228–235, 2025.
https://doi.org/10.1007/978-3-031-98462-4_29

A critical component of these automated tools is accurate speaker diarization (the task of determining "who spoke when"), which alongside Automatic Speech Recognition (ASR), represent an early bottleneck for downstream analyses. Inaccuracies in these stages can propagate and negatively impact later evaluations such as analyzing teaching practices and student talk [3]. Speaker diarization in Kindergarten to 12th grade (K-12) classrooms is especially challenging due to noisy background acoustics, overlapping speech, and diverse language usage that are often underrepresented in training datasets for existing models [1,16]. As a result, the classroom represents an acoustically challenging environment for unimodal (audio-only) diarization models [15,16]. However, when transcriptions are available, linguistic cues may provide valuable context to improve diarization performance in classroom settings.

Traditionally, speaker diarization is based solely on audio data. However, when transcriptions are available, speaker diarization can be approached as determining "who spoke what". While this framing has been explored in domains such as healthcare [5,14], meetings [6], and telephone conversations [6,18], it still remains largely unexplored in educational settings.

In this work, we investigate multimodal diarization for classroom recordings using cascaded audio and language models. Specifically, we use Generative Pretrained Transformers (GPTs) to re-score initial cluster predictions when given transcript information. We explore both zero-shot prompting and supervised fine-tuning of a GPT model to leverage instructive language cues, thereby enhancing traditional speaker diarization pipelines. Lastly, we investigate ACE as a post-clustering step that improves the audio embedding space and diarization performance. This approach leads us to address two central research questions:

1. Can instructive language serve as a reliable indicator for determining speaker roles in classroom settings?
2. Are LLMs capable of learning these patterns to enhance diarization?

2 Data

We use an internally collected dataset of 291 TeachFX classroom recordings from 247 English-speaking teachers (29 teachers have 2+ recordings), totaling 94 h of audio. In addition, each teacher provided an audio recording of themselves reading a haiku, which we refer to as the *enrollment audio*. The haiku's short, structured format (typically around 6 s long) provides a compact but phonetically rich sample of the speaker's voice, making it ideal for speaker embedding extraction. The classroom recordings span various K-12 grade levels and subject areas, including Mathematics, English Language Arts, Social Studies, Arts and Science. Due to privacy considerations, we do not collect demographic information (e.g. age, ethnicity, gender, race) of teachers, nor do we collect any personally identifiable information about individual students. More information about our privacy policy can be found here[1]. This dataset is considered "in-the-wild", as

[1] https://teachfx.com/privacy-policy.

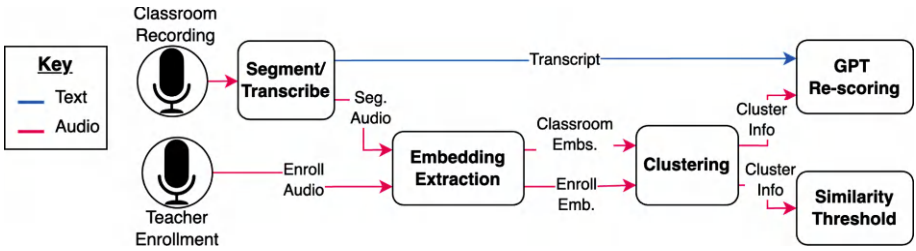


Fig. 1. Diarization Pipeline

the audio is collected in a dynamic classroom environment that is subject to microphone mismatch, recording distance variation, overlapping speech, babble speech, and background noise.

All classroom recordings were transcribed using AssemblyAI[2], yielding transcripts that include spoken words, word time boundaries, and sentence time boundaries. Trained annotators had access to the classroom audio, transcription, and enrollment audio. They followed consistent guidelines when labeling each word as either 'Student' or 'Teacher'. Every recording was reviewed by at least two raters and achieved an interrater agreement defined by a Cohen's Kappa score of $\geq$0.95. For all our experiments, we partitioned the dataset into speaker-independent train and test sets using an 80/20 split.

3 Methods

Figure 1 illustrates our overall diarization pipeline. We begin by segmenting and transcribing the classroom recording using AssemblyAI. Segmented classroom audio and enrollment audio are passed through an embedding extractor. We then cluster these embeddings and apply ACE. Finally, cluster predictions are made using a similarity threshold approach or GPT re-scoring.

3.1 Audio Embedding Extraction

We use ECAPA-TDNN [4], a widely adopted architecture known for its strong performance in speaker identification and verification tasks. The model was trained on the VoxCeleb1 and VoxCeleb2 datasets, which contain speech data from 1,251 and 6,112 speakers, respectively [2,11,12], using noise-robust data augmentation techniques. We use ECAPA-TDNN as an acoustic feature extractor to obtain fixed-length speaker embeddings from segmented audio.

3.2 Audio Clustering

After obtaining ECAPA-TDNN embeddings for each audio segment, we group them using agglomerative clustering [7,13,17]. During clustering, we compute

[2] https://www.assemblyai.com/products/speech-to-text.

the cosine similarity between each cluster's centroid and the enrollment audio. However, one limitation of this approach is that there can be audio mismatch (e.g., microphone proximity, background noise, tonal shift, etc.) between the enrollment audio and the classroom recordings.

One workaround is to sample a teacher segment from the classroom recording as the new enrollment audio, but a single sample may be less robust than using a cluster of recordings. With this in mind, we introduce a post-clustering process called **Adaptive Centroid-based Enrollment (ACE)**. Instead of using the teacher's original enrollment audio, ACE selects the cluster centroid that most closely aligns with the enrollment sample (based on cosine similarity) and redefines the selected centroid as the new "enrollement" reference. We then recompute cosine similarities between this re-enrolled centroid and all other cluster centroids. By effectively re-enrolling with a more representative embedding, ACE mitigates environmental and expressive variations between the original enrollment audio and a specific classroom recording, improving the robustness and accuracy of this clustering step.

3.3 Cosine Similarity Threshold

Our baseline model represents an audio-only approach commonly used in speaker verification tasks where the cluster prediction is based on the cosine similarity between a cluster embedding and the enrollment embedding. Specifically, we use a predetermined threshold to classify clusters as student or teacher, tuning this threshold through a hyperparameter sweep on the training set over values ranging from 0 to 1 in increments of 0.01.

3.4 GPT Re-scoring

To complement the audio-only baseline, we introduce a multimodal approach that leverages GPT-4o-mini to refine diarization predictions using linguistic context. In the system prompt we provide GPT with the transcript, a description of the diarization task, relevant guidelines, and aggregate information. We prompt the model to score each cluster as either "teacher" or "student", based on cosine similarities, the cosine similarity threshold, and transcription content. We explored both zero-shot prompting and fine-tuning. In the zero-shot setting, GPT-4o-mini receives only the system prompt and raw data. For finetuning, we train GPT-4o-mini in a supervised fashion using labeled clusters from the training data, where each cluster is assigned its majority class label (teacher or student). This allows GPT to learn classroom-specific discourse patterns related to the diarization task.

3.5 Evaluation Metrics

To evaluate our diarization performance, we compute word-level accuracy and word-level unweighted F1-score, which have been used in previous speaker

Table 1. Word-level speaker diarization results. † and ‡ are used to represent statistical significance over ACE methods and over all methods respectively.

Prediction Mechanism	Modalities	Accuracy	Unweighted F1-score
Similarity Threshold	audio	88.2 (13.0)	79.0 (19.8)
GPT-zero	audio+text	87.5 (10.1)	81.4 (14.0)
GPT-finetune	audio+text	94.1 (10.1)†	90.5 (12.7)†
Similarity Threshold + ACE	audio	92.6 (10.8)	86.1 (17.8)
GPT-zero + ACE	audio+text	92.0 (8.8)	86.4 (13.8)
GPT-finetune + ACE	audio+text	**95.4 (6.3)**†	**92.0 (9.8)**†‡

diarization research [8–10,19]. For each recording, we calculate these metrics by comparing the predicted label for each word to the ground-truth label. We also perform a paired t-test to assess statistical significance (p-value < 0.05).

4 Results

4.1 Overall Performance

The goal of our experiments are to assess how well different diarization strategies perform in real-world classroom recordings. We compare three main approaches similarity threshold, GPT-zero, and GPT-fine-tune, all with and without ACE. Table 1 summarizes the outcomes of these comparisons by displaying the mean and standard deviation across all test recordings for each metric. We use a † to indicate significance when comparing GPT-finetune, GPT-zero, and Similarity Threshold within its respective grouping (i.e. with or without ACE). We use ‡ to indicate significance when comparing GPT-finetune+ACE against all other methods, including both ACE and non-ACE approaches.

We find that finetuning GPT significantly outperforms similarity threshold and zero-shot prompting achieving higher accuracy and unweighted F1-score suggesting GPT is highly effective at cluster re-scoring given transcript data to learn from. This highlights GPT's ability to integrate both textual and cluster information, making it particularly well-suited for diarization in classroom contexts, where language usage can serve as an important indicator of teacher or student speech. We also observe that ACE consistently enhances diarization performance across all prediction mechanisms, highlighting how re-enrollment within a classroom recording can alleviate issues of enrollment audio mismatch. Ultimately, the combination of ACE and GPT-finetuning achieves the highest average accuracy (95.4%) and average unweighted F1-score (92.0%).

4.2 Role-Specific Language

The strong performance of GPT-based re-scoring motivated a deeper analysis of language usage between teachers and students, specifically by examining trigrams from the transcript. To filter out common phrases used by both groups,

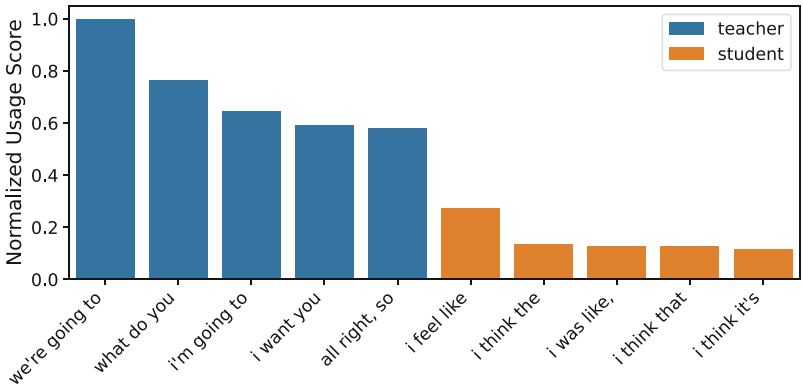


Fig. 2. Top 5 trigrams for teachers and students

we compute the usage difference for teacher (U_T) and student (U_S), shown in Eqs. 1 and 2, where p is a given trigram.

$$U_T(p) = \max(0, \text{count}_T(p) - \text{count}_S(p)) \tag{1}$$

$$U_S(p) = \max(0, \text{count}_S(p) - \text{count}_T(p)) \tag{2}$$

These usage counts are affected by contribution bias, as teacher-student talk time may vary within recordings. To address this, we normalize the usage difference by the total number of trigrams in each speaker group, following Eq. 3, where P_R represents all trigrams used by $R \in \{\text{Teacher}, \text{Student}\}$.

$$\text{Score}_R(p) = \frac{U_R(p)}{|P_R|} \tag{3}$$

Lastly, we perform min-max normalization so that each score lies in the range of [0,1]. Figure 2 shows the top five trigrams for teachers and students. We find that the top teacher phrases include instruction- and direction-related language (i.e., "we're going to"), whereas students more often use response-oriented language (i.e., "I feel like"). These distinct language patterns demonstrate the linguistic cues GPT can leverage to improve speaker diarization.

5 Conclusion

In this work, we demonstrated the importance of language usage for speaker diarization in classroom settings, that ACE can improve speaker embedding robustness, and how GPT re-scoring can improve diarization performance. Our multimodal approach with ACE and finetuned GPT re-scoring ultimately yields significant performance improvements in word-level diarization accuracy and unweighted F1-score (i.e., up to 95.4% and 92.0%, respectively). Additionally, we present some data-driven insight into trigram usage between teachers and

students, to highlight some of the distinct linguistic cues that LLM-based rescoring can take advantage of in classroom settings. These results contribute to advancing speaker diarization for in-the-wild classroom recordings, ultimately enabling automated feedback tools that provide teachers with deeper insights into classroom dynamics.

Acknowledgments. This work was supported, in whole or in part, by the Gates Foundation [INV-060568]. The conclusions and opinions expressed in this work are those of the author(s) alone and shall not be attributed to the Foundation.

References

1. Attia, A.A., Liu, J., Ai, W., Demszky, D., Espy-Wilson, C.: Kid-whisper: towards bridging the performance gap in automatic speech recognition for children vs. adults. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 7, pp. 74–80 (2024)
2. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: deep speaker recognition. In: INTERSPEECH (2018)
3. Demszky, D., Wang, R., Geraghty, S., Yu, C.: Does feedback on talk time increase student engagement? Evidence from a randomized controlled trial on a math tutoring platform. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, pp. 632–644 (2024)
4. Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: Interspeech 2020, pp. 3830–3834 (2020). https://doi.org/10.21437/Interspeech.2020-2650
5. Flemotomos, N., Georgiou, P., Narayanan, S.: Linguistically aided speaker diarization using speaker role information. In: The Speaker and Language Recognition Workshop (Odyssey 2020), pp. 117–124 (2020). https://doi.org/10.21437/Odyssey.2020-17
6. Huang, Y., Wang, W., Zhao, G., Liao, H., Xia, W., Wang, Q.: On the success and limitations of auxiliary network based word-level end-to-end neural speaker diarization. In: Proceedings Interspeech 2024, pp. 32–36 (2024)
7. Kenny, P., Reynolds, D., Castaldo, F.: Diarization of telephone conversations using factor analysis. IEEE J. Sel. Top. Sig. Process. **4**(6), 1059–1070 (2010)
8. Khan, A., et al.: Toward realigning automatic speaker verification in the era of covid-19. Sensors **22**(7), 2638 (2022)
9. Kumar, A.K., Waldekar, S., Sahidullah, M., Saha, G.: Robust acoustic domain identification with its application to speaker diarization. Int. J. Speech Technol. **25**(4), 933–945 (2022)
10. Li, J., Hasegawa-Johnson, M., McElwain, N.L.: Towards robust family-infant audio analysis based on unsupervised pretraining of Wav2vec 2.0 on large-scale unlabeled family audio. In: Proceedings Interspeech 2023, pp. 1035–1039 (2023)
11. Nagrani, A., Chung, J.S., Zisserman, A.: VoxCeleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017)
12. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: VoxCeleb: large-scale speaker verification in the wild. Comput. Sci. Lang. (2019)

13. Sell, G., Garcia-Romero, D.: Speaker diarization with PLDA I-vector scoring and unsupervised calibration. In: 2014 IEEE Spoken Language Technology Workshop (SLT), pp. 413–417. IEEE (2014)
14. Shafey, L.E., Soltau, H., Shafran, I.: Joint speech recognition and speaker diarization via sequence transduction. In: Interspeech 2019, pp. 396–400 (2019). https://doi.org/10.21437/Interspeech.2019-1943
15. Wang, J., Dudy, S., He, X., Wang, Z., Southwell, R., Whitehill, J.: Speaker diarization in the classroom: how much does each student speak in group discussions? In: Proceedings of the 17th International Conference on Educational Data Mining, pp. 360–367 (2024)
16. Wang, J., et al.: Optimizing speaker diarization for the classroom: applications in timing student speech and distinguishing teachers from children. J. Educ. Data Mining **17**(1), 98–125 (2025)
17. Wang, Q., Huang, Y., Lu, H., Zhao, G., Moreno, I.L.: Highly efficient real-time streaming and fully on-device speaker diarization with multi-stage clustering. arXiv preprint arXiv:2210.13690 (2022)
18. Wang, Q., Huang, Y., Zhao, G., Clark, E., Xia, W., Liao, H.: DiarizationLM: speaker diarization post-processing with large language models. In: Interspeech 2024, pp. 3754–3758 (2024). https://doi.org/10.21437/Interspeech.2024-209
19. Yang, Y., Achananuparp, P., Huang, H.Y., Jiang, J., Lim, E.P.: Speaker verification in agent-generated conversations. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5655–5676 (2024)



How Generative AI Impact Student Emotion and Engagement in Programming Tasks?

Boxuan Ma[1(✉)], Liyuan Guo[2], Tianyuan Yang[1], and Jihong Ding[2]

[1] Kyushu University, Fukuoka, Japan
boxuan@artsci.kyushu-u.ac.jp, yang.tianyuan.791@s.kyushu-u.ac.jp
[2] Hainan University, Haikou, Hainan, China
jhding@hainanu.edu.cn

Abstract. The rapid advancement of AI has created transformative opportunities in education, reshaping how students learn. While much of the existing research has focused on AI's impact on student behavior and performance, its effects on emotional experiences and learning engagement remain underexplored. This study aims to address this gap using questionnaires and video analysis to capture students' real-time emotions and engagement levels, investigating how AI influences these two dimensions in programming education. Our preliminary results reveal interesting findings and lay the foundation for further research.

Keywords: Generative AI · Programming · Engagement · Emotion

1 Introduction

Generative AI (GenAI) provides personalized feedback, enables dynamic content generation, and supports adaptive learning environments, thereby enhancing both instructional practices and student learning experiences [1,3,8,9]. Their integration into educational contexts, particularly in programming education, holds the potential to influence not only cognitive and behavioral aspects of learning but also students' emotional experiences and overall engagement [5]. While most research has focused on AI's impact on academic performance and learning behaviors [13], the psychological dimensions, such as emotional experiences [14] and sustained learning engagement [16], remain underexplored.

Some researchers suggest that Generative AIs, such as ChatGPT, can enhance student motivation and foster positive emotional states [6]. By providing immediate feedback and reducing the stress associated with complex tasks, it may foster confidence, boost interest, and facilitate deeper engagement [4,8,12]. However, there are also concerns about potential negative effects. For example, recent studies have indicated that overreliance on AI-generated content could diminish critical thinking skills, reduce meaningful cognitive engagement, and even lead to anxiety or a sense of disconnection due to excessive dependency on AI tools [1,2,10,11,18].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 236–243, 2025.
https://doi.org/10.1007/978-3-031-98462-4_30

In the field of programming education, while the impact of AI on students' engagement and emotional experiences is important, existing research has primarily focused on evaluating AI' performance in solving programming tasks [15,17,20] or assessing students' overall experiences and perceptions of AI-supported programming learning [7,10,11,22]. Only a few studies examined students' emotional experiences and learning engagement when using AI in learning. However, they just included such factors in questionnaires and lacked a comprehensive investigation. Also, they rely predominantly on self-reported measures with only a single question. This may not accurately capture the dynamic nature of emotions and engagement that occur during learning [7,22].

Given these gaps, there is a critical need to investigate how AI influences both students' emotional experiences and their learning engagement when using AI for programming tasks. This study aims to address these gaps by adopting a mixed-method approach that combines video analysis and questionnaire surveys to capture students' real-time emotions and learning engagement. We hope to provide a better understanding of how AI affects both the emotional and engagement dimensions of student learning.

2 Method

2.1 Study Context and Design

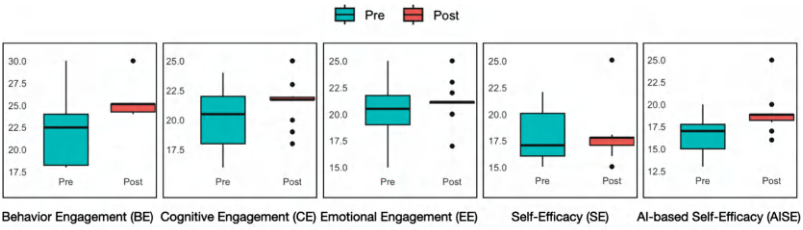
Study Context and Design. The study took place in a 2-h practice lab session at our university. We recruited 14 undergraduate students to participate in a voluntary practice lab session of an introductory Python course. All students (7 male, 7 female, average age is 18.8, SD is 0.86) provided informed consent to participate in the study.

The students first completed a pre-questionnaire designed to assess their baseline engagement and emotional state. Then, they were given a brief presentation introducing the course material and explaining relevant concepts. After this, the students engaged in a 75-min practice session. For the practice session, students were required to solve two Python tasks using the Turtle library. The first task involved drawing a snake, and the second assignment was similar but more complex. These two assignments were designed to assess the students' ability to apply their learning in a practical context. The participants were allowed to solve the assignments with access to AI assistance (we used Ernie bot[1], a widely used AI chatbot in China). During this session, students' emotions, motivation, and engagement were measured twice using a 5-point Likert scale (1 = completely disagree, 5 = completely agree) through three simple questions designed to minimize interference with tasks. Then, they need to fill in a post-questionnaire consisting of the same questions as the pre-questionnaire. This approach helped track changes in students' emotional experiences and engagement over time. The study finished with a voluntary interview about the impact of AI on student learning engagement, emotions, and self-efficacy.

[1] https://yiyan.baidu.com/.

2.2 Data Collection

We employed different methods to collect and analyze data on students' emotional experiences and learning engagement. The questionnaire was designed to assess students' emotional experiences and engagement based on established instruments [21,23]. We also conducted a semi-structured interview, and all interview questions were open-ended. All interviews were transcribed for data analysis. During the experiment, students' computer screens were recorded, and their facial expressions and movements were captured via webcam. We coded programming behaviors, student emotions, and engagement levels from screen recordings and webcam videos, following a recurrent procedure based on an established framework [19]. Four researchers independently coding the data and refining the scheme through discussion. Detailed questionnaire items and coding framework are given in supplementary material[2].



Dimension	Pre(Mean)	Post(Mean)	p-value	Test
Behavior Engagement (BE)	20.214	21.714	0.079	t-test
Cognitive Engagement (CE)	20.143	21.143	0.189	t-test
Emotional Engagement (EE)	17.929	17.714	0.712	t-test
Self-Efficacy (SE)	16.429	18.857	0.007	t-test
AI-supported Self-Efficacy (AISE)	22.286	25.143	0.016	t-test

Fig. 1. Pre- and Post-Questionnaires Scores Across Dimensions.

3 Results

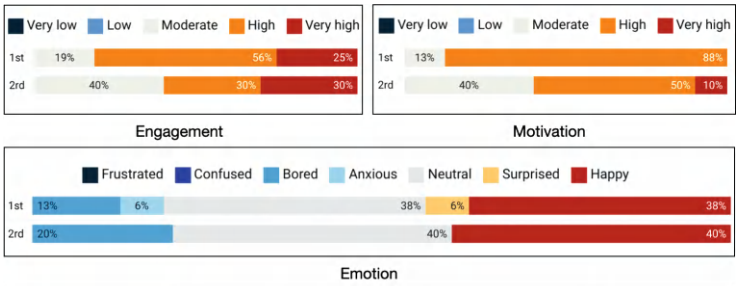
3.1 Analysis of Questionnaires

Figure 1 summarize the results of two questionnaires for five dimensions: Behavior Engagement (BE), Cognitive Engagement (CE), Emotional Engagement (EE), Self-Efficacy (SE), and AI-based Self-Efficacy (AISE). We can observe that SE and AISE demonstrated statistically significant increases, suggesting that using AI in the study had a meaningful positive impact on students' self-efficacy and AI-based self-efficacy, which enhance students' perceived confidence

[2] https://bit.ly/4hHOc0T.

in their abilities. In addition, BE and CE showed slight increases in mean scores, and EE remained relatively stable, but these changes were not statistically significant.

Figure 2 summarizes the results of mid-session questionnaires. The findings indicate a decline in students' learning engagement and motivation in the second session, accompanied by an increase in negative emotions such as confusion and anxiety. Possible explanations include increased task difficulty, reduced novelty, or growing reliance on AI. However, we also observed a higher proportion of students reporting very high engagement and motivation, along with an increase in positive emotional responses. These mixed trends suggest the need for a detailed analysis of student behaviors and video recordings to gain deeper insights.

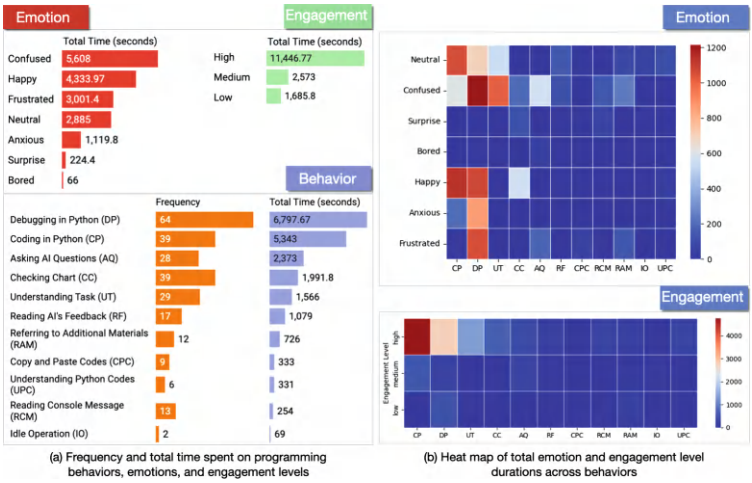


Metric	1st (Mean)	2rd (Mean)	p-value	Test
Engagement	4.072	3.947	0.325	Wilcoxon Signed-Rank Test
Motivation	3.856	3.773	0.448	Wilcoxon Signed-Rank Test
Emotion	-	-	0.009	Chi-square Test

Fig. 2. Comparison of Mid-session Questionnaires.

3.2 Analysis of Video Record

To better understand how students' emotions and engagement levels were affected during AI interactions, we analyzed the video coding data. Figure 3 presents the frequency and total time spent on various programming behaviors, emotional states, and engagement levels during the practice session. Overall, debugging and coding were the most dominant in both time and frequency. Additionally, asking AI questions and reading AI's feedback were also notable behaviors, highlighting students' active engagement in AI-assisted learning. From an emotional perspective, confusion and frustration accounted for a significant portion of the time, reflecting challenges in problem-solving. However, happiness was also notable, suggesting moments of success or effective AI support. Engagement analysis indicates that high engagement was dominant, while low engagement was minimal. We also observe that tasks such as coding (CP) and debugging



(a) Frequency and total time spent on programming behaviors, emotions, and engagement levels

(b) Heat map of total emotion and engagement level durations across behaviors

Fig. 3. The relationships between behaviors, emotions, and engagement levels.

(DP) engage students the most, but also evoke higher levels of negative emotions, such as frustration and confusion, suggesting that these tasks are particularly challenging. However, tasks like reading AI feedback (RF) and asking AI questions (AQ) appear to have little impact on generating positive emotions or enhancing engagement. This is likely due to the uncertainty about the effectiveness of the AI's responses.

To explore this further, we closely examined students' behaviors as they transitioned to debugging their code after interacting with AI. We also coded whether the AI's feedback were helpful, enabling us to assess how the effectiveness of AI feedback influenced students' emotional states and levels of engagement. Overall, when students found the AI's responses helpful, there was a significant increase in positive emotions (happy) compared to other emotional states ($\chi^2 = 7.82, p < 0.05$). Conversely, when the AI's responses were unhelpful or incorrect, students predominantly experienced confusion ($\chi^2 = 12.34, p < 0.01$). These results not only suggest statistical significance but also imply that the effectiveness of AI in enhancing positive emotions directly correlates with its perceived usefulness in the task. For example, during debugging (DP), one student asked AI a question, and the AI provided a useful solution. The student then modified his code based on the feedback, and the expected output confirmed the AI's effectiveness. This positive outcome led to happy emotions, reflected in behaviors such as smiling and leaning forward. However, when the AI's response was inadequate or misaligned with the student's expectations, the student's emotional state shifted to confusion. This was evident in behaviors like repeatedly reading the AI's feedback (RF), staring at the screen, shaking their head, and re-examining error messages (RCM). In another case, a student continuously asked AI questions and modified the code based on the feedback, but after several unsuccessful attempts, the student's emotions visibly shifted to frustration.

These findings underscore the significant role that AI feedback plays in shaping students' emotions and engagement. When students perceive the AI's responses as helpful, their emotional states become more positive, and their engagement deepens. On the other hand, when the AI's responses are ineffective or inadequate, students experience frustration and confusion, which diminishes their learning experience and engagement.

3.3 Analysis of Interviews

Two students participated in the interview, providing valuable insights into the positive and potential negative effects of AI on learning engagement, emotion, and self-efficacy. For learning engagement, S1 noted, "When I struggle to find a solution, my engagement drops significantly. However, if I use AI, my engagement increases." Similarly, S2 emphasized, "Overall, AI makes me more passionate about learning. It's like having an expert by your side, helping to solve any problem." Despite these benefits, both students recognized the potential downsides of AI: "You might become overly dependent, which could reduce your engagement with learning." Additionally, S1 mentioned that relying on AI can hinder active thinking: "When I look at AI's solution, I just passively accept the result and don't think it actively."

S1 explained that AI's impact on emotions is closely linked to the difficulty of the task. "For easier tasks, I mainly use AI to fill in the gaps. My interest is highest in these cases because I'm curious about where I'm lacking and want to get inspiration from AI, and my interest in learning remains high for a long time." However, S1 also noted, "For more challenging tasks, I still find the process interesting when I ask AI for help. But once I know the answer, my interest wanes, and I might not fully master it." S2 expressed: "When I face a problem I can't solve, knowing I can use AI prevents me from feeling frustrated or anxious. But after asking AI several times without getting a correct solution, I start feeling irritated and disappointed."

AI also has varied effects on self-efficacy. S1 believes that using AI correctly can enhance learning abilities, while reliance on it may reduce independent learning skills. On the other hand, S2 expressed, "With AI tools, my sense of self-efficacy, achievement, and motivation to learn all improve. Tools empower human intelligence. They are the key reason humans are more advanced than animals."

4 Conclusion

This study explores the impact of generative AI on students' emotional experiences and engagement in programming education. By analyzing questionnaires, interviews, and video data, we examined how students' emotions and engagement levels were influenced during their interactions with AI. Our preliminary findings highlight both the benefits and challenges of integrating AI into programming education. However, due to the limited sample size, these results may not be generalizable. In future research, we plan to conduct experiments with

a larger group of users and further investigate the relationship between emotion, engagement, and behaviors, as well as how these factors influence students' learning outcomes.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Number 24K20903.

References

1. Al Shloul, T., et al.: Role of activity-based learning and ChatGPT on students' performance in education. Comput. Educ. Artif. Intell., 100219 (2024)
2. Alneyadi, S., Wardat, Y.: ChatGPT: revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in emirates schools. Contemp. Educ. Technol. **15**(4), ep448 (2023)
3. Chiu, T.K.: The impact of generative AI (GENAI) on practices, policies and research direction in education: a case of ChatGPT and midjourney. Interact. Learn. Environ. **32**(10), 6187–6203 (2024)
4. Escalante, J., Pack, A., Barrett, A.: AI-generated feedback on writing: insights into efficacy and ENL student preference. Int. J. Educ. Technol. High. Educ. **20**(1), 57 (2023)
5. Heung, Y.M.E., Chiu, T.K.: How ChatGPT impacts student engagement from a systematic review and meta-analysis study. Comput. Educ. Artif. Intell., 100361 (2025)
6. Huesca, G., et al.: Effectiveness of using ChatGPT as a tool to strengthen benefits of the flipped learning strategy. Educ. Sci. **14**(6), 660 (2024)
7. Humble, N., Boustedt, J., Holmgren, H., Milutinovic, G., Seipel, S., Östberg, A.S.: Cheaters or AI-enhanced learners: consequences of ChatGPT for programming education. Electron. J. e-Learn. (2023)
8. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
9. Li, H., Ma, B.: Design of AI-powered tool for self-regulation support in programming education. arXiv preprint arXiv:2504.03068 (2025)
10. Ma, B., Chen, L., Konomi, S.: Enhancing programming education with ChatGPT: a case study on student perceptions and interactions in a Python course. In: International Conference on Artificial Intelligence in Education, pp. 113–126. Springer (2024)
11. Ma, B., Chen, L., Konomi, S.: Exploring student perception and interaction using ChatGPT in programming education. In: 21st International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA 2024, pp. 35–42. IADIS Press (2024)
12. Pellas, N.: The effects of generative AI platforms on undergraduates' narrative intelligence and writing self-efficacy. Educ. Sci. **13**(11), 1155 (2023)
13. Rahman, M.M., Watanobe, Y.: ChatGPT for education and research: opportunities, threats, and strategies. Appl. Sci. **13**(9), 5783 (2023)
14. Reeve, J.: How students create motivationally supportive learning environments for themselves: the concept of agentic engagement. J. Educ. Psychol. **105**(3), 579 (2013)

15. Shoufan, A.: Exploring students' perceptions of ChatGPT: thematic analysis and follow-up survey. IEEE Access (2023)
16. Sinatra, G.M., Heddy, B.C., Lombardi, D.: The challenges of defining and measuring student engagement in science (2015)
17. Skjuve, M., Følstad, A., Brandtzaeg, P.B.: The user experience of ChatGPT: findings from a questionnaire study of early users. In: Proceedings of the 5th International Conference on Conversational User Interfaces, pp. 1–10 (2023)
18. Stojanov, A., Liu, Q., Koh, J.: University students' self-reported reliance on ChatGPT for learning: a latent profile analysis. Comput. Educ. Artif. Intell. **6**, 100243 (2024)
19. Sun, D., Boudouaia, A., Zhu, C., Li, Y.: Would ChatGPT-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? An empirical study. Int. J. Educ. Technol. High. Educ. **21**(1), 14 (2024)
20. Tlili, A., et al.: What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learn. Environ. **10**(1), 15 (2023)
21. Wang, Q., Gao, Y., Wang, X.: Exploring engagement, self-efficacy, and anxiety in large language model EFL learning: a latent profile analysis of Chinese university students. Int. J. Hum.–Comput. Interact., 1–10 (2024)
22. Yilmaz, R., Yilmaz, F.: Augmented intelligence in programming learning: examining student views on the use of ChatGPT for programming learning. Comput. Hum. Behav. Artif. Hum. **1**(2), 100005 (2023)
23. Yilmaz, R., Yilmaz, F.: The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. Comput. Educ. Artif. Intell. **4**, 100147 (2023)



Student Perceptions of Adaptive Goal Setting Recommendations: A Design Prototyping Study

Conrad Borchers(✉), Cindy Peng, Qianru Lyu, Paulo F. Carvalho, Kenneth R. Koedinger, and Vincent Aleven

Carnegie Mellon University, Pittsburgh, PA, USA
{cborcher,cindypen,qlyu,pcarvalh,krk,aleven}@cs.cmu.edu

Abstract. Many AIED systems support self-regulated learning, yet, support for setting and achieving practice goals has received little attention. We examine how middle school students respond to system-recommended practice goals, building on the success of similar data-driven recommendations in other domains. We introduce an adaptive dashboard in an intelligent tutoring system. The dashboard tracks progress (time, mastery), suggests dynamic goal adjustments, and visualizes performance—aiming to increase learning opportunities. In a prototyping study with five middle schoolers (grades 6–7), we used scenario-based tasks to examine reactions to goal recommendations during under- or overachievement. Thematic analysis revealed: (1) students accepted system suggestions as guidance but wanted final control; (2) they valued adaptive recommendations tied to past performance to increase their chance of meeting their goal and pace themselves; and (3) they strongly preferred motivational features (e.g., streaks, celebrations) for engagement. Our study contributes design implications for practice goal-setting in tutoring systems, highlighting personalized, explainable learning analytics and motivational design to support goal-directed behavior.

Keywords: human-centered design · goal setting · self-regulated learning · prototyping study · intelligent tutoring systems · K-12

1 Introduction and Related Work

Goal setting is a crucial component of self-regulated learning (SRL) [23]. Effective goals feature clear timeframes and performance standards [20]. AIED systems often provide such features [5,17]. However, in addition to setting and monitoring goals, learners must also set *intentions* to work toward goals. Gollwitzer's model of action phases [13] suggests that goal pursuit involves a pre-actional phase of planning goal-directed behavior and a post-actional phase dedicated to reflecting on performance and goals. Without support at these stages, defining and acting on goals remains challenging for students using technology [7,22].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 244–251, 2025.
https://doi.org/10.1007/978-3-031-98462-4_31

How might AIED systems better support students in setting intentions and executing goal-directed behavior? Learners must be supported in *identifying and evaluating* opportunities to work toward goals [14] and seek regular performance feedback [18]. However, past research has focused on supporting student goal setting *during* practice, after practice has already been initiated [3,5].

We designed a goal-setting recommendation dashboard building open learner models (OLMs) displaying mastery progress, which can support performance reflection [7,8], but, alone, are limited in providing goals and rewards, which are known to be effective [4,15]. Conversely, non-digital goal contracts may not effectively integrate feedback and data capabilities (e.g., log data) [22].

We build on research showing that adaptive, performance-based goal-setting improves outcomes in health domains [1,18], and extend our prior design work where students valued difficulty feedback in learning analytics for digital homework [22]. The dashboard supports goal intentions and planning *outside of practice* by (a) showing current goals, including those set externally (e.g., goal contracts [4]), and (b) enabling goal reflection and adjustment via data-driven recommendations. We explore student perceptions of these recommendations to inform future learning environment design. We ask:

– RQ: What are students' perceptions of adaptive practice goal recommendations and constraints for tutoring system practice?

2 Goal-Setting Dashboard System

The Goal-Setting Dashboard is embedded in TutorShop and the Lynnette tutoring system [2]. The system tracks practice time and skill mastery (e.g., subtracting constants) through log data [19]. Students can work on multiple assignments, and progress is tracked and aggregated across all of them. The dashboard is compatible with any tutoring system and subject domain that uses an underlying skill model and complies with tutoring system log data standards [16]. Currently, the dashboard supports tutoring systems that are authored with CTAT [2].

2.1 Dashboard and Goal Recommendations

The homepage displays the current focus goal and its associated rewards based on deadlines. Goals and rewards—set collaboratively by students and parents or teachers through contracts [12]—are manually uploaded after paper-based goal setting [4]. The dashboard supports time-based (e.g., 30 min/week) and skill mastery goals (e.g., 2 skills/day), tracking progress toward both (Fig. 1).

After practice, for each goal, the dashboard provides students with adaptive recommendations that foster growth. In the first eight days, if their goal achievement has been above or below 25% of their goal, the recommendation is $goal + (performance - goal) * 0.5$. Later, the algorithm recommends a goal according to the 60% percentile of students' most recent, non-missing nine days of goal achievement [1]. Recommendations are accompanied by encouragement

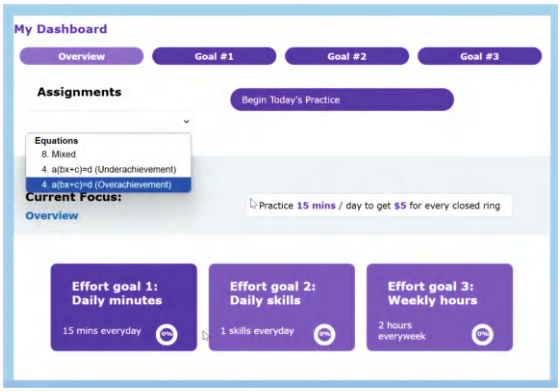


Fig. 1. Overview page where achievement conditions were simulated for the user study (default; underachievement; overachievement) with current goal focus.

(Fig. 2). Students can accept, decline, or overwrite goal recommendations, but overwritten goals must be within 25% of the suggested value to prevent unrealistic or trivial targets. Since recommendations are based on past performance, they remain suitable across achievement levels.

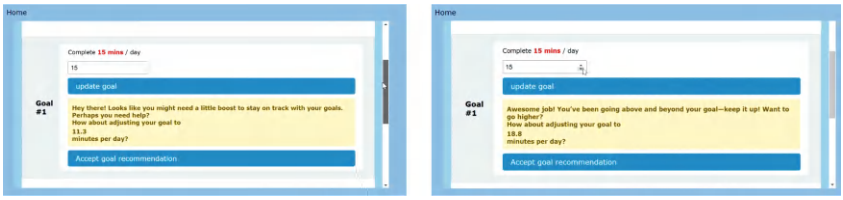


Fig. 2. (Left) underachievement and (right) overachievement scenario showing messages that encourage effort, celebrate progress, and suggest realistic goals.

2.2 Learning Analytics

The interface's goal tabs incorporate three visualization elements: a ring progress visualization showing time-based or skill-based goals with encouraging messages, a horizontal bar graph displaying skill mastery progress, and a line graph comparing daily practice minutes against weekly goal targets using a red dash-dotted line to represent the goal benchmark (Fig. 3). The red benchmark line represents the expected daily practice effort or achievement needed to meet the weekly goals to encourage spaced practice and repetition [11,28].

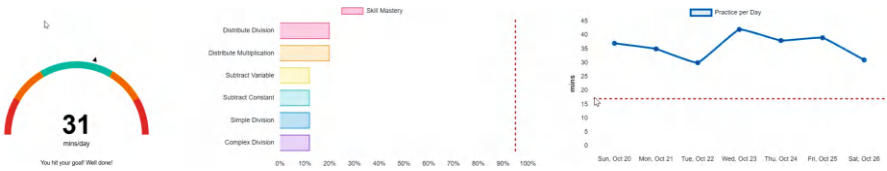


Fig. 3. Learning analytics: (1) a ring progress tracker, (2) a horizontal bar graph for skill mastery, and (3) a line graph comparing daily practice minutes against goals. (Color figure online)

3 Methods

Five middle school students participated in the study. The participants, predominantly from urban areas in the Mid-Atlantic U.S. region, included two 6th and three 7th graders, aged 11 to 13. Students were 60% male, with three being Caucasian, one East Asian, and one mixed race. Recruitment was conducted through university outreach centers. Each participating student received $50 in Amazon gift cards. The study was approved by the relevant institutional review board and included informed consent from participants and their guardians.

Participants interacted with the prototype on laptops, with all activities screen- and audio-recorded. All students, sitting in groups of 1–2 in a large room, were guided through UX tasks by a researcher each, prompting them to think aloud [21]. Tasks included navigating tabs and reviewing goal progress.

Afterward, participants engaged in scenario-based simulations. Using Carroll's scenario-based design framework, all scenarios elicited participants' interaction patterns and decision-making, providing contexts for reflecting on system features [9,10]. They worked through three scenarios in a fixed order: (1) *average achievement*, where participants' performance was just about enough to meet their goal; (2) *underachievement*, which prompted participants to adjust their goals; (3) *overachievement*, where participants were encouraged to set more ambitious goals. Students were prompted to adjust goals within system constraints.

Finally, a group feedback activity using sticky notes encouraged active participation and visual organization of thoughts [6]. Participants wrote reflections and categorized them under themes: *Ease of Use*, *Goal Motivation*, *Support & Feedback*, *Suggestions for Improvement*, and *Future Use*, placed them on the board, and voted on the reflections that resonated most with their experiences.

Transcripts, completed worksheets, and sticky notes were analyzed through an independent thematic analysis by two researchers [25]. Themes were cross-referenced and discussed. The analysis focused on understanding the dashboard system's usability, functionality, and areas for improvement.

4 Results

Theme 1: All five students appreciated the autonomy to set goals with dashboard guidance and wanted to retain final control. When asked to adjust goals in the dashboard, recommendations supported students shifting from broad aspirations (e.g., [Green] *"to get it done"*) to more concrete targets (e.g., [Green] *"10 min a day, 1 skill a week"*; [Black] *"20 min per week, 3 skills per week"*). Students emphasized wanting to retain final control over their goal-setting: [Black] noted, *"The system doesn't know my situation,"* while [Green] stated, *"I prefer to set my goal myself."*

Theme 2: Students found system recommendations and guardrails helpful and broadly acceptable. ***Subtheme 2.1: Students responded positively to achievement feedback and adjustment reminders.*** [Pearl] found suggestions helpful, stating, *"It's helpful, it's like, if it's too little it can say like hey you can bring it up if you want, so that helps, I guess."* [Red] appreciated the system recommendations, stating they *"would use the system recommendations of adjusting goals to make them harder under different conditions."* Students' underlying reasons for liking underachievement feedback aligned with our design assumption that such feedback would make future goal achievement more attainable, boosting motivation. When asked about responding to unmet goals, [Green] stated *"I would change my goals so I can meet them"*, and [Gold] noted *"I would want to try harder and make my goals a little bit easier so I can meet them."* ***Subtheme 2.2: Students had varied responses to overachieving and appreciated flexibility.*** Some students were motivated to increase their goals after exceeding expectations, as [Black] shared, *"Probably up it seems?"* Conversely, others were more cautious when overachieving, emphasizing the importance of not overworking. [Pearl] appreciated that the system set limits on goal increases, noting, *"It was said at first you cannot go up that high. I like that because then you don't like overwork yourself or anything like that,"* while [Green] similarly advised, *"Don't practice more than you should."* ***Subtheme 2.3:*** We observed that at least two students ([Gold], [Red]) had challenges comprehending goal recommendations expressed as percentages (e.g., 75% of your current goal) and decimals (e.g., 11.3 min).

Theme 3: Students strongly preferred motivational features and visual elements related to performance achievement to reward effort. ***Subtheme 3.1: Students requested gamified elements like streaks or milestones.*** [Pearl] shared, *"I think the goals would probably help you because you want to get your goals and finish what you set."* [Red] echoed: *"And like having this streak in any program is like having a reason to keep someone coming back."* [Black] further supports this claim by stating *"that's like a super big thing for other apps is just like setting a goal to make a streak for seven days and then it just moves on. For this, it's like it's right here so you can keep track of your goals and they're important."* ***Subtheme 3.2: Students desired cohesive aesthetics and visualizations to recognize goal progress.*** Several students suggested ways to make the interface more engaging. [Red] highlighted the need for

celebratory elements, suggesting, *"the way the like completion is set up is cool, but nothing really happens when you complete your goals. Having even something simple like some confetti would be cool."* [Gold] suggested incorporating seasonal themes to make the interface more dynamic, stating, *"I feel like you have some very seasonal ... like in fall you can make for like a farm ... and then summer and winter, it's something that could be like getting snow, maybe Christmas tree."* These findings indicate students increasingly expect AIED dashboards to integrate gamification, recognizing its role in sustaining engagement and motivation.

5 Discussion

Students appreciated setting their own goals and viewed system recommendations as helpful refinements—consistent with prior research showing students often use recommendations as starting anchors [22]. Students still wanted control over final practice goals, echoing research on open learner models [7]. Results suggest this control can be balanced with recommendations: students found over- and underachievement feedback helpful for reflection and goal adjustment [14].

Students generally did not question goal-setting guardrails (e.g., preventing setting too easy goals). In contrast, they agreed that preventing overwork, one pitfall of goal setting [18], is important. Some students welcomed opportunities to increase their goals after overachieving, once again highlighting the need for flexibility and user control. Gentle prompts and visual cues in the dashboard could motivate students to reflect on their progress and preferences to adjust their goals. This result is encouraging as this active reflection on achievement may boost self-efficacy when intermediate, and progressively challenging goals are achieved [24,26,27]. Motivational streaks, celebratory visuals, and projections of potential future achievement [7] could further enhance these reflection processes and persistence in future design revisions.

5.1 Limitations and Future Work

First, our sample of five students limits generalizability to broader or more diverse populations. We plan to expand our sample upon revisions of our prototype. Second, we relied on scenario-based simulations rather than long-term, in situ usage. While these scenarios allowed us to observe students' immediate reactions to underachievement or overachievement, they may not capture the full complexity of how goals evolve over weeks of practice. Third, this study does not fully explore constraints imposed by deadlines, milestones, and in-class work time, which may vary by subject domain and content. Maintaining student autonomy may be challenging in these contexts, and we plan to interview teachers to study design alignment with classroom expectations.

6 Conclusion

Goal setting is a core component of SRL, but past AIED solutions are limited in helping students set realistic goals through data and spur reflection on goal progress. This study introduced an adaptive goal-setting dashboard for middle school learners and validated its design in a prototyping study. Three main findings emerged: (1) students appreciated system recommendations when they retained final control; (2) students were responsive to adaptive adjustments balancing under- and overachievement and goal guardrails, though at times preferred keeping current goal levels; and (3) students described motivational elements (e.g., streak counters, confetti) as key to sustained engagement. This work advances design knowledge for K–12 AIED dashboards, illustrating how adaptive feedback, explainable data, and gamified elements could support students in pre- and post-actional phases of goal setting, which past solutions paid relatively little attention to. We anticipate that shared student-system control while maintaining student autonomy and regular feedback will increase student learning opportunities to benefit from tutoring systems and increase the chance of goal achievement and its motivational benefits.

Acknowledgments. This research was funded by the Institute of Education Sciences (IES) of the U.S. Department of Education (Award #R305A220386). We thank Dorian Pan, Kennion Gubler, and Octav Popescu for their contributions to the development of the dashboard.

References

1. Adams, M.A., et al.: Adaptive goal setting and financial incentives: a 2 × 2 factorial randomized controlled trial to increase adults' physical activity. BMC Public Health **17**, 1–16 (2017)
2. Aleven, V., et al.: An integrated platform for studying learning with intelligent tutoring systems: CTAT+TutorShop. arXiv preprint arXiv:2502.10395 (2025)
3. Aleven, V., Mclaren, B., Roll, I., Koedinger, K.: Toward meta-cognitive tutoring: a model of help seeking with a cognitive tutor. Int. J. Artif. Intell. Educ. **16**(2), 101–128 (2006)
4. Alwahbi, A.: The use of contingency contracting in educational settings: a review of the literature. Educ. Res. Rev. **15**(6), 327–335 (2020)
5. Azevedo, R., Witherspoon, A., Chauncey, A., Burkett, C., Fike, A.: MetaTutor: a metacognitive tool for enhancing self-regulated learning. In: 2009 AAAI Fall Symposium Series (2009)
6. Ball, L.J., Christensen, B.T., Halskov, K.: Sticky notes as a kind of design material: how sticky notes support design cognition and design collaboration. Des. Stud. **76**, 101034 (2021)
7. Borchers, C., Ooge, J., Peng, C., Aleven, V.: How learner control and explainable learning analytics about skill mastery shape student desires to finish and avoid loss in tutored practice. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference, pp. 810–816 (2025)
8. Bull, S.: There are open learner models about! IEEE Trans. Learn. Technol. **13**(2), 425–448 (2020)

9. Carrol, J.M.: Five reasons for scenario-based design. In: Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. In: HICSS-32. Abstracts and CD-ROM of Full Papers, p. 11. IEEE (1999)
10. Carroll, J.M.: Making Use: Scenario-Based Design of Human-Computer Interactions. MIT Press (2003)
11. Dempster, F.N.: Spacing effects and their implications for theory and practice. Educ. Psychol. Rev. **1**, 309–330 (1989)
12. Fraser, E., Mallek, C., Sigourney, D., Watland, M.: Strategies to improve student motivation to complete homework assignments (1999)
13. Gollwitzer, P.M.: Mindset theory of action phases. In: Handbook of Theories of Social Psychology, vol. 1, pp. 526–545 (2012)
14. Gollwitzer, P.M., Sheeran, P.: Implementation intentions and goal achievement: a meta-analysis of effects and processes. Adv. Exp. Soc. Psychol. **38**, 69–119 (2006)
15. Kahle, A.L., Kelley, M.L.: Children's homework problems: a comparison of goal setting and parent training. Behav. Ther. **25**(2), 275–290 (1994)
16. Koedinger, K.R., Baker, R.S., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the EDM community: the PSLC DataShop. In: Handbook of Educational Data Mining, vol. 43, pp. 43–56 (2010)
17. Li, X., et al.: FLoRA engine: using analytics to measure and facilitate learners' own regulation activities. J. Learn. Anal., 1–24 (2025)
18. Locke, E.A., Latham, G.P.: The development of goal setting theory: a half century retrospective. Motiv. Sci. **5**(2), 93 (2019)
19. Long, Y., Holstein, K., Aleven, V.: What exactly do students learn when they practice equation solving? Refining knowledge components with the additive factors model. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge, pp. 399–408 (2018)
20. McCardle, L., Webster, E.A., Haffey, A., Hadwin, A.F.: Examining students' self-set goals for self-regulated learning: goal properties and patterns. Stud. High. Educ. **42**(11), 2153–2169 (2017)
21. O'Brien, L., Wilson, S.: Talking about thinking aloud: perspectives from interactive think-aloud practitioners. J. User Experience **18**(3), 113–132 (2023)
22. Peng, C., Borchers, C., Aleven, V.: Designing homework support tools for middle school mathematics using intelligent tutoring systems. In: Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024, pp. 1730–1733. International Society of the Learning Sciences (2024)
23. Schunk, D.H.: Goal setting and self-efficacy during self-regulated learning. Educ. Psychol. **25**(1), 71–86 (1990)
24. Srivastava, S.: To study the effect of academic achievement on the level of self confidence. J. Psychosoc. Res. **8**(1) (2013)
25. Terry, G., Hayfield, N., Clarke, V., Braun, V., et al.: Thematic analysis. In: The SAGE Handbook of Qualitative Research in Psychology, vol. 2, no. 17–37, p. 25 (2017)
26. Wäschle, K., Allgaier, A., Lachner, A., Fink, S., Nückles, M.: Procrastination and self-efficacy: tracing vicious and virtuous circles in self-regulated learning. Learn. Instr. **29**, 103–114 (2014)
27. Wesson, C.J., Derrer-Rendall, N.M.: Self-beliefs and student goal achievement. Psychol. Teach. Rev. **17**(1), 3–12 (2011)
28. Yan, V.X., Sana, F., Carvalho, P.F.: No simple solutions to complex problems: cognitive science principles can guide but not prescribe educational decisions. Policy Insights Behav. Brain Sci. **11**(1), 59–66 (2024)



Ordinal Classification for Transformer-Based Automated Essay Scoring Models

Sungjin Nam(✉)

ACT Education Corp., Iowa City, IA 52243, USA
sungjin.nam@act.org

Abstract. This study examines the application of ordinal classification methods in transformer-based automated essay scoring (AES) models and compares them with conventional regression-based approaches. We evaluate our models in various settings, including the types of pre-trained encoders, essay prompts, and scoring scales. Our results show that the ordinal classification models perform comparably to the regression models regarding the overall QWK score and across different score ranges, as evidenced by centrality measures and qualitative analysis.

Keywords: Automated Essay Scoring · Natural Language Processing · Neural Network · Ordinal Classification

1 Introduction

Automated essay scoring (AES) systems are widely used to predict human-annotated scores with efficiency and consistency for evaluating a large volume of open-ended essay responses [1,17]. Solving the AES task as a regression problem is a popular approach [4,11,21,22]. It can conveniently handle multiple writing tasks with different scoring ranges by applying simple scaling, such as min-max scaling [4,11]. However, many AES tasks are fundamentally *ordinal classification* tasks, where human raters score the essays in numeric categories using a scoring rubric. Solving ordinal classification tasks using regression or general classification methods misinterprets that the score categories are equally spaced or independent and can lead to suboptimal performance [12]. Using ordinal classification methods can also enhance interpretability by providing probability scores for each class and helping prevent the use of incorrect discretization cutoffs [3,15].

Previous studies have investigated ordinal classification methods for AES tasks but used less modern machine learning model architectures [2,6,19,20] or limited score ranges [2,6]. In this study, we focus on applying ordinal classification methods to transformer-based AES models across different types of essay tasks, rather than striving for the best available results. Our contributions are:

1. We show that ordinal classification methods can provide comparable performance to the regression-based models for transformer-based AES models.
2. We demonstrate how the ordinal classification methods perform on various essay tasks and score ranges using centrality metrics and qualitative analysis.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 252–259, 2025.
https://doi.org/10.1007/978-3-031-98462-4_32

2 Related Studies

2.1 Ordinal Classification for Neural Networks

Previous studies have investigated how to preserve the natural order of class labels for the classification models [7,9]. Recent studies showed that using cross-entropy loss with softmax normalization can effectively estimate people's age [12]. Studies like Consistent Rank Logits (CORAL) [3] and Conditional Ordinal Regression for Neural Networks (CORN) [15] presented advanced methods to address the rank consistency problem, such as the model's binary prediction results are not being ordinarily sorted, by applying shared weights [3] or the chain rule for conditional probabilities [15].

However, these studies were based on computer-vision tasks [3,12,15] or numeric and categorical datasets [15]. On the other hand, AES datasets are mainly composed of open-ended text responses with unbalanced score label distributions, which might pose different challenges. Pre-trained NLP models that are used by AES studies, such as BERT (base model with 110M parameters) [6,13,17,22], also have significantly different parameter sizes than the computer-vision models, like the ResNet-34 model used in [3,12,15] (21M). In this paper, we examine how the ordinal classification methods for neural networks perform in comparison to regression-based models on the AES task.

2.2 AES Using Ordinal Classification

Neural network-based AES models have become popular as they can capture complex linguistic patterns from essays and achieve the state-of-the-art performance [13,17]. However, many studies tend to treat AES as a regression problem. It incorrectly assumes that the distance between scores is both equal and continuous. Using simple cutoff points, such as rounding, to discretize continuous scores into classes may not yield the best results. A classification approach that disregards the natural order between score categories would also struggle [7,9].

Some AES studies have looked into how well ordinal logistic regression models work when using either manually created features [19,20] or convolutional (CNN) [6] or recurrent neural network (RNN) [2] representations. However, these studies were limited to using less modern NLP model architectures [2,6,19,20] or did not fully cover a wide variety of scoring ranges [6,20]. In this study, we investigate how transformer-based ordinal classification AES models can perform as well as regression-based models. We compare our models' performance across different pre-trained encoders and essay tasks with various score ranges. We also examine whether the score predictions from the ordinal models are more biased toward the majority scores through both quantitative and qualitative analyses.

3 Methods

3.1 Ordinal Loss Functions

We compared three ordinal classification methods for neural network AES models. All methods converted numeric scores into binary vector representations from [7].

For example, an essay score from the 0–3 range was represented with three ordinal categories (e.g., [0, 0, 0] for a score of 0, and [1, 1, 1] for a score of 3).

OR-NN. [12] proposed using a binary cross-entropy to learn $(K-1)\cdot 2$ sized matrices, where softmax normalized rows represent categories, and columns represent the negative and positive probabilities. N and K represent the number of examples and label categories. $p(o_i^k|x_i,\theta)$ is the predicted probability for a class (k) given by the model parameters (θ) and an essay (x_i). $\mathbb{1}\{o_i = y_i\}$ indicates if the prediction (o_i) is equal to the true class (y_i).

$$L_O = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K-1}\mathbb{1}\{o_i = y_i\}w_i \log(p(o_i^k|x_i,\theta)) \tag{1}$$

CORAL. One drawback of `OR-NN` is that it does not guarantee the rank consistency. To address this, `CORAL` [3] suggested using an output layer with shared weights (W) for all categories and bias values (b_k) for each label category. It uses the cross-entropy of $K-1$ binary classifiers with a single sigmoid logit value $(\hat{P}(y_i = 1) = \sigma(\theta(x_i, W) + b_k))$ to represent each class label's probability.

$$L_R = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K-1}[y_i^k \cdot \log(\hat{P}(y_i = 1)) + (1 - y_i^k)\cdot \log(1 - \hat{P}(y_i = 1))] \tag{2}$$

CORN. Based on `CORAL`, `CORN` [15] improved the method further by using trainable weights for *each* score category (w_k). It uses a cross-entropy of the summed subset probabilities (S_j : j^{th} class' probability; $\hat{P}(y_i = 1) = \sigma(\theta(x_i, w_k) + b_k)$) and estimates a series of conditional probabilities by applying the chain rule $(\hat{P}(y_i > r^k) = \prod_{k=1}^{K}\sigma(\theta(x_i, w_k) + b_k))$, maintaining the rank consistency among $K-1$ probabilities.

$$L_N = -\frac{1}{\sum_{j=1}^{K-1}|S_j|}\sum_{i=1}^{S_j}\sum_{k=1}^{K-1}[\log(\hat{P}(y_i = 1))\cdot\mathbb{1}\{y_s > r^k\} + \log(1-\hat{P}(y_i = 1))\cdot\mathbb{1}\{y_s \le r^k\}] \tag{3}$$

3.2 AES Model Design

We tested the ordinal classification methods with `BERT` [8,18,22], `RoBERTa`, and `Bigbird-RoBERTa` (all `base-uncased` versions) for generalizability. Based on a pre-trained model, each essay representation was weighted with a four-headed self-attention block, then the `[CLS]` token's vector was passed to the scoring block. The scoring block consisted of a fully connected layer, a dropout (p = 0.3), ReLU activation, and the last linear layer for each individual learning method. All models were trained for 20 epochs, with a batch size of 32, a learning

Table 1. Essays included in the ASAP dataset.

	P1	P2	P3	P4	P5	P6	P7	P8
Number of Essays	1,783	1,800	1,726	1,772	1,805	1,800	1,569	723
Avg. Length in Words	350	350	150	150	150	150	250	650
Score Range	2–12	1–6	0–3	0–3	0–4	0–4	0–30	0–60

rate of 2.5e−05, and linear learning rate schedulers. We trained separate models per prompt and selected the best model based on the development set loss. The ordinal classification models' outputs counted the number of score labels that had probabilities greater than 0.5. We also tested a combination of ordinal and regression losses using pre-determined weights (e.g., $L = L_{RMSE} + 0.1 \cdot L_{OR-NN}$).

3.3 Baseline Models

We compared our ordinal classification models with the RMSE baselines. They provided similar results to the MSE baselines [4,21,22], but provided better loss-value scaling when combined with ordinal classification losses. The output from the regression models was rounded to the integer scores.

`R`^{2} `BERT` [22] is a strong baseline AES model that employs a regression objective, which can be integrated with a rank-based loss objective. `NPCR` [21] is a state-of-the-art AES model that combines scoring and contrastive ranking tasks. Both models use the `BERT-base-uncased` pre-trained encoder. The ordinal logistic regression model (`OL Reg.`) [20] is included for informal comparisons because they did not use the same data splits [16] and missed the results for Prompt 8 due to their model's poor performance.

3.4 Dataset

We used the popular ASAP dataset[1] (Table 1). Eight prompts include argumentative (P1-P2), source-dependent (P3-P6), and narrative (P7-P8) writing tasks in various response lengths and score ranges. We brought the 5-fold cross-validation splits [16] for better comparison. We used the Quadratic Weighted Kappa (QWK) to measure the models' overall prediction performance [16,20,22], and the Coefficient of Variation (CoV), the ratio of the standard deviation to the mean ($\frac{s}{\mu}$), to quantify the centrality [10].

4 Results

4.1 Overall Prediction Performance

The upper rows of Table 2 show that our baseline models (e.g., `BERT-RMSE`) perform similarly to the R^2 BERT [22] (MSE) models, suggesting that our simple AES model design was reasonably effective. Results from other `RMSE` baselines suggest that different pre-trained models may offer additional improvements.

[1] https://www.kaggle.com/c/asap-aes.

Table 2. QWK results per prompt show that the ordinal classification methods like `OR-NN` can provide comparable performance to `(baseline)` models. *: the performance is significantly lower ($^*_-$) or higher ($^*_+$) than the corresponding RMSE model (one-tailed wilcoxon test, $p < 0.05$). **Bold**: The best QWK score per prompt. Underline: The best QWK score from the ordinal classification models.

Method	Model	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
(LR)	OL Reg. [20]	0.770	0.710	0.700	0.810	0.820	0.830	0.630	N/A	0.753
(MSE)	NPCR [21]	0.856	0.750	0.756	0.851	0.847	0.858	0.838	0.779	0.817
	R^2 BERT [22]	0.800	0.679	0.679	0.822	0.803	0.797	0.837	0.725	0.768
	(+Rank)	0.817	0.719	0.698	0.845	0.841	0.847	0.839	0.744	0.794
(RMSE)	BERT	0.815	0.645	0.680	0.802	0.801	0.820	0.825	0.722	0.764
	RoBERTa	0.812	0.659	0.667	0.819	0.805	0.816	**0.832**	0.707	0.765
	Bigbird	**0.830**	0.678	0.687	0.806	0.803	0.806	**0.832**	0.750	**0.774**
CORN [15]	BERT	0.782$^*_-$	0.648	0.690	0.796	0.799	0.806	0.805	0.566$^*_-$	0.737$^*_-$
	RoBERTa	0.782$^*_-$	0.665	0.700	0.811	0.801	0.809	0.810$^*_-$	0.622$^*_-$	0.750$^*_-$
	Bigbird	0.794$^*_-$	0.657	0.705	0.815	0.806	0.809	0.823$^*_-$	0.666$^*_-$	0.759$^*_-$
OR-NN [12]	BERT	0.788$^*_-$	0.636	0.679	0.808	0.806	0.808	0.809	0.654$^*_-$	0.749$^*_-$
	RoBERTa	0.785$^*_-$	0.663	0.683	0.818	0.801	0.806	0.827	0.681	0.758$^*_-$
	Bigbird	0.810	**0.683**	0.694	**0.822**	0.798	0.792	0.815	0.716	0.766
RMSE + OR-NN	BERT	0.796	0.650	0.680	0.814	0.815	**0.823**	0.824	0.704	0.763
	RoBERTa	0.817	0.648	0.659	0.813	0.796	0.815	0.822$^*_-$	0.714	0.760
	Bigbird	**0.830**	0.638$^*_-$	0.683	0.816	0.806	0.821$^*_+$	0.831	**0.755**	0.773

The middle rows show that our ordinal classification models can provide comparable QWK results to the regression baselines across all essay prompts. The `CORAL` method did not provide meaningful performance overall, with the overall QWK scores around 0.54, especially the prompts with wider score ranges (P1 and P2 (0.414–0.490), 7 (0.550–0.563), and 8 (0.200–0.219)). Models using `CORN` and `OR-NN` methods showed that the ordinal classification methods could provide comparable QWK results to other regression baselines, like the RMSE or R^2 BERT [22] models. Our results also surpassed the `OL Reg.` [20] model, even when we included the longest essay set with the largest score range (P8). Ordinal models performed better than regression models in some prompts, but tend to perform worse when the score range is wider (P1, P7, and P8).

The `Bigbird`-based and the `OR-NN` models performed better than other models. However, these ordinal classification models and the combined models (e.g., `RMSE+OR-NN` models) did not demonstrate significant improvements over the `RMSE` baselines or other advanced models [21,22].

4.2 Prediction Centrality and Score-Wise Accuracy

A higher CoV indicates more deviation from the mean and lower centrality in the predicted scores [10]. All AES models had lower CoV scores than the human-annotated scores. The ordinal classification models had marginally higher CoV scores than the RMSE baselines, suggesting that these models can produce similarly diverse predictions as the regression baselines across all essay prompts (Table 3).

Table 3. CoV results per prompt from the `Bigbird`-based models. The ordinal classification models could provide similarly diverse predictions as did the regression baselines. *: Statistically significantly better CoV score than the RMSE model ($p < 0.05$).

Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
(Human)	0.236	0.320	0.441	0.656	0.403	0.356	0.285	0.155	0.356
(`RMSE`)	0.209	0.270	0.400	0.594	0.361	0.324	0.250	0.122	0.316
`OR-NN`	0.203	0.262	0.408	0.640*	0.361	0.316	0.247	0.109	0.318
`RMSE + OR-NN`	0.213	0.259	0.406	0.624	0.366	0.333*	0.250	0.123	0.322

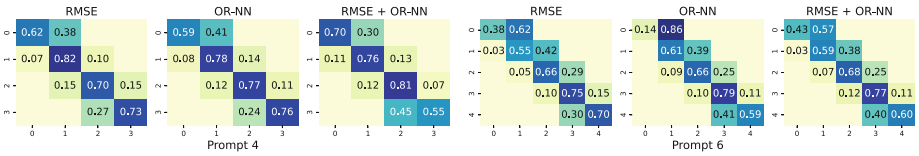


Fig. 1. Confusion matrix from the `Bigbird`-based models for Prompts 4 and 6.

Figure 1 shows that the ordinal classification models could provide comparably accurate results to the regression models across different score ranges, and the results might be improved by combining the learning objectives. In Prompt 4, the `OR-NN` model performed better than the `RMSE` model (QWK: 0.822 vs. 0.806). The combined model (QWK: 0.816) improved the predictions for the lowest (0) and a mid-score (2). In Prompt 6, the `OR-NN` model performed worse than the `RMSE` model (QWK: 0.806 vs. 0.792), especially for the lowest score point (0). However, the combined model (QWK: 0.821) could improve the performance for the lowest score point while sacrificing the highest score point (4) results.

5 Conclusion and Discussion

We demonstrated that the ordinal classification AES models can match the performance of regression-based models, while providing more accurate and interpretable representations of discrete scores. Combining the ordinal and regression loss functions could improve the predictions for the lower- or higher-ends.

Using a more rigorous method did not necessarily lead to more accurate predictions. `CORAL` and `CORN` achieved 100% rank consistency in predicted score *probabilities* (e.g., the probabilities are sorted in descending order), whereas `OR-NN` marked 0% for P7 and P8. All methods marked nearly perfect results for predicted *labels* (99.8%–100%; e.g., binary score classes are ordinarily sorted). More advanced training strategies, like contrastive or reinforcement learning, could help the rank-consistent methods to learn better to handle scores with large ranges or unbalanced distributions while meeting the requirements.

Additional efforts on hyperparameter sweeps or refining the methods for combining the loss functions [21,22] could also improve predictions. Future works on

exploring cross-prompt [5], multi-trait [14] AES, or non-English datasets would improve the generalizability of ordinal classification methods.

References

1. Attali, Y., Burstein, J.: Automated essay scoring with e-rater® v. 2. J. Technol. Learn. Assess. **4**(3) (2006)
2. Berggren, S.J., Rama, T., Øvrelid, L.: Regression or classification? Automated essay scoring for Norwegian. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 92–102 (2019)
3. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recogn. Lett. **140**, 325–331 (2020)
4. Cao, Y., Jin, H., Wan, X., Yu, Z.: Domain-adaptive neural automated essay scoring. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1011–1020 (2020)
5. Chen, Y., Li, X.: PMAES: prompt-mapping contrastive learning for cross-prompt automated essay scoring. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1489–1503. Association for Computational Linguistics, Toronto, Canada, July 2023. https://doi.org/10.18653/v1/2023.acl-long.83. https://aclanthology.org/2023.acl-long.83/
6. Chen, Z., Zhou, Y.: Research on automatic essay scoring of composition based on CNN and OR. In: 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 13–18. IEEE (2019)
7. Cheng, J., Wang, Z., Pollastri, G.: A neural network approach to ordinal regression. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1279–1284. IEEE (2008)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)
9. Frank, E., Hall, M.: A simple approach to ordinal classification. In: De Raedt, L., Flach, P. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 145–156. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44795-4_13
10. Kernbach, J.M., Staartjes, V.E.: Foundations of machine learning-based clinical prediction modeling: Part I–introduction and general principles. In: Staartjes, V.E., Regli, L., Serra, C. (eds.) Machine Learning in Clinical Neuroscience, pp. 7–13. Springer, Cham (2022)
11. Lee, Y., Jeong, S., Kim, H., Kim, T.I., Choi, S.W., Kim, H.: NC2T: novel curriculum learning approaches for cross-prompt trait scoring. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2204–2208 (2023)
12. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output CNN for age estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4920–4928 (2016)
13. Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. Artif. Intell. Rev. **55**(3), 2495–2527 (2022)

14. Ridley, R., He, L., Dai, X.Y., Huang, S., Chen, J.: Automated cross-prompt scoring of essay traits. Proc. AAAI Conf. Artif. Intell. **35**(15), 13745–13753 (2021). https://doi.org/10.1609/aaai.v35i15.17620. https://ojs.aaai.org/index.php/AAAI/article/view/17620
15. Shi, X., Cao, W., Raschka, S.: Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. Pattern Anal. Appl. **26**(3), 941–955 (2023)
16. Taghipour, K., Ng, H.T.: A neural approach to automated essay scoring. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1882–1891 (2016)
17. Uto, M.: A review of deep-neural automated essay scoring models. Behaviormetrika **48**(2), 459–484 (2021). https://doi.org/10.1007/s41237-021-00142-y
18. Wang, Y., Wang, C., Li, R., Lin, H.: On the use of BERT for automated essay scoring: joint learning of multi-scale essay representation. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3416–3425. Association for Computational Linguistics, Seattle, United States, July 2022. https://doi.org/10.18653/v1/2022.naacl-main.249. https://aclanthology.org/2022.naacl-main.249
19. Wang, Y., Hu, R.: A prompt-independent and interpretable automated essay scoring method for Chinese second language writing. In: China National Conference on Chinese Computational Linguistics, pp. 450–470. Springer (2021)
20. Woods, B., Adamson, D., Miel, S., Mayfield, E.: Formative essay feedback using predictive scoring models. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Ciscovery and Data Mining, pp. 2071–2080 (2017)
21. Xie, J., Cai, K., Kong, L., Zhou, J., Qu, W.: Automated essay scoring via pairwise contrastive regression. In: Calzolari, N., et al. (eds.) Proceedings of the 29th International Conference on Computational Linguistics, pp. 2724–2733. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, October 2022. https://aclanthology.org/2022.coling-1.240/
22. Yang, R., Cao, J., Wen, Z., Wu, Y., He, X.: Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1560–1569 (2020)



Modeling Intrinsic Load: Can Item Difficulty Be a Reasonable Proxy?

Guher Gorgun[1,2,3(✉)], Minghao Cai[1], and Carrie Demmans Epp[1]

[1] EdTeKLA Group, University of Alberta, Edmonton, AB, Canada
{gorgun,minghaocai,cdemmansepp}@ualberta.ca
[2] Europa-Universität Flensburg, Flensburg, Germany
[3] Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Abstract. In addition to adapting to learner knowledge, the materials given to learners could be adapted to ensure they have the cognitive resources available for performing the learning task. In order to perform such adaptation, we need to be able to model the different domains of cognitive load (i.e., intrinsic, extraneous, and germane), which continues to be a logistical challenge. In this study, we examined whether item difficulty estimated using item response theory—a long-established and widely-used psychometrics method—could be used to model intrinsic load in adaptive learning environments. We found that item difficulty could be a viable option for modeling the intrinsic load of a task at the learning session level. We also found that learner experiences of intrinsic load varied according to their abilities. This finding underscores the need for modeling intrinsic load for creating adaptive systems that match task complexity with learner needs.

Keywords: Cognitive load · IRT · domain modeling

1 Introduction

Adaptive learning environments, such as intelligent tutoring systems (ITS), aim to adjust the learning environment or content to meet individual learner needs [32]. This type of adaptive behavior has supported fairly large gains in learner knowledge when the adaptation has focused on domain knowledge [33]. However, there are other cognitive elements to consider when adapting the learning environment to meet learner needs. One such element is the cognitive load that is imposed by the recommended learning materials or activities [5,7]. It has been well-established that the design of learning materials can impact the learner's cognitive load both for better and for worse [8,9,15,30,35]. While there are design guidelines and practices for how to minimize unnecessary load [27], we have no way of estimating or representing the inherent load of a learning task within the domain model of adaptive learning systems. We, therefore, investigate an approach for automatically modeling this aspect of cognitive load.

Cognitive load theory postulates three interrelated domains: intrinsic load (IL), extraneous load (EL), and germane load (GL) [31]. Intrinsic load (IL)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 260–267, 2025.
https://doi.org/10.1007/978-3-031-98462-4_33

[28,29] stems from the inherent complexity of the learning task. The sources of IL [17,19] include element interactivity [10] and prior knowledge of the learner [14]. Extraneous load (EL) is influenced by the design of the learning material [28]. The final domain is germane load (GL) and refers to activities that a learner should undertake to contribute to their performance and learning processes [22]. While cognitive load theory has emerged as a crucial constituent for optimizing instructional design, its domains need to be measured and modeled separately to enable system adaptation.

1.1 Challenges with Cognitive Load Measurement

There are two mainstream methods used for measuring cognitive load: self-report and physiological measures [17].

Self-report measures ask the learner to report their experience (e.g., item difficulty or mental effort) using a rating scale [4]. This approach is subjective and easy to administer [21]. When there are existing measures, this approach could be easily integrated into an ITS, but taking these measures would interrupt learning processes. Additionally, the idiosyncratic interpretation of self-report scales may threaten the validity and reliability of the cognitive load instrument.

Physiological measures rely on using cameras or sensors to automatically collect data from learners to understand the cognitive load exerted [17]. Using physiological sensors poses the challenge of determining the optimal time window for predicting cognitive load from various sensor signals [6].

Many studies highlight a strong relationship between perceived difficulty and the IL of a task [12,13,16,23,24,26]. Thus, we propose to use established measurement models of item difficulty and learner ability from item response theory (IRT) to approximate the IL of a learning task. Several advantages of the proposed approach include being more objective, economic, and scalable compared to physiological sensors and self-report measures. Learner ability—a proxy for pre-knowledge and expertise—can be estimated with IRT to model a more nuanced IL. In addition, the proposed method should be less intrusive during the learner's engagement with the task [25] and should not interfere with the task flow. Using two different large-scale datasets coming from adaptive learning environments, we investigated the feasibility of using IRT-based difficulty and ability estimates to approximate the reported IL levels for tasks in adaptive learning environments. We ask "To what extent can IRT parameters be used as a proxy for the IL of a task in adaptive learning environments?".

2 Methods

This study uses data from two adaptive learning environments: a game-based learning environment targeting reading comprehension and English language arts (RC hereafter) and a learning environment targeting Python programming skills (PP hereafter). The collection and use of these datasets underwent institutional research ethics reviews, and consent was collected from participants.

In the RC platform, the game continually inferred learners' reading abilities using their performance on literacy tasks and covertly adapted the difficulty of reading materials and questions. The difficulty level was initially set to grade 8; it was then automatically adjusted based on learner performance. Participants were between 17 and 33 years old ($M = 24.1$). Twenty-three of them were women and twelve were men. The languages first spoken at home varied: 24 spoke Chinese, 4 spoke Japanese, and 3 spoke Persian (Farsi). Additionally, Korean, Gujarati, French, and Spanish were each reported as the home language for one participant.

In the PP platform, the system employed various types of interactive activities to support computer-science education. All participants had some prior experience with Python. They were between 18 and 34 years old ($M = 22.6$). Of the 39 participants, 1 was non-binary, 19 were women, and 19 were men. Many ($n = 12$) had previously used PP in one of their courses.

2.1 Modeling Procedure and Analysis

We administered a cognitive load scale to participants who used the RC ($n = 35$) or PP ($n = 39$) platforms. This was a timed study, and participants' self-reported cognitive load was recorded at approximately 5-min intervals (see [7]). There was an average of ten data collection intervals; the number varied across participants.

To estimate item difficulty, all historic response data from RC and PP have been used. Learner performance data (responses) from all participants on the RC platform ($N = 6{,}464$) and the PP platform ($N = 164$) were used to estimate the difficulty of the questions. In RC, there were a total of 11,142 items. In PP, there were a total of 164 items. Because we used adaptive systems, the historical response data were sparse, meaning that the number of responses per item varied. To obtain stable difficulty parameters, we followed recommendations regarding sample size requirements per item (e.g., [20]) and selected items administered at least 100 times, resulting in the removal of items ($n = 1{,}915$) from the difficulty analysis.

The simplest IRT model is Rasch, where only item difficulty is estimated [1]. IRT models allow linking learner ability with item difficulty. The analyses were conducted in R with the *mirt* package.

To answer our research question, we identified the questions that learners answered between administrations of the cognitive load questionnaire. We mapped the completed questions to those in the question difficulty database and calculated the average difficulty for each data collection interval. We also estimated learner ability using all questions that learners answered as well as the probability of a correct response associated with each item that a learner answered. The final dataset included average item difficulty per data collection interval, average learner ability, average probability of a correct answer, and the reported IL, EL, and EL +IL (i.e., the sum of IL and EL) values for each learning session (i.e., collection interval). As the measures were collected on different scales, we standardized all measures.

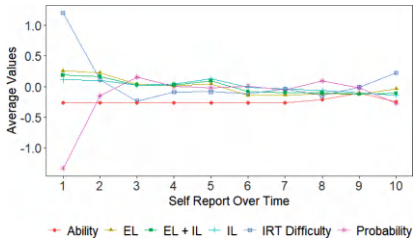


Fig. 1. Average cognitive load and IRT values per session for RC

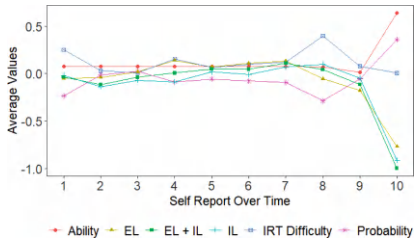


Fig. 2. Average cognitive load and IRT values per session for PP

3 Results

To answer our research question, we compared the averaged and standardized self-reported IL and EL values with the averaged and standardized values of item difficulty, learner ability, and probability of answering an item correctly across 10 learning sessions. For both the RC and PP platforms, we observed that the relationship between IL, EL, and IRT parameters was similar, with a few exceptions (see Fig. 1 and Fig. 2).

The difficulty parameters and IL, EL, and EL + IL values were parallel to one another across learning sessions, suggesting that difficulty could be a viable proxy for IL and EL for each learning session. Since item difficulty was parallel to the reported IL and EL levels per session, instructors, instructional designers, and adaptive systems could leverage averaged item difficulty values as a proxy for IL and EL.

Consistent with our expectation, when item difficulty was higher than learner ability across the learning session, the reported IL and EL values were higher as well (Fig. 1). However, when item difficulty and learner ability were close to one another, the reported IL and EL values were lower (Fig. 2). Thus, aligning item difficulty with ability might be important to create learning environments with appropriate levels of IL and EL. Additionally, for RC, those with lower ability levels, and for PP, those with high ability levels, reached the last session during the study. This is partly due to item and passage length, as some individuals had data collection sessions with longer than 5-min intervals, as expected with the employed event-contingent sampling procedure.

We also analyzed how IL and EL values changed for each individual when difficulty and ability levels were (mis)matched. When learners' ability was consistently higher than item difficulty, learners reported lower levels of IL and EL, and when their ability was lower than item difficulty, learners reported higher levels of IL and EL for both RC and PP (see Fig. 3 for example items).

We expected that the mismatch between ability and difficulty at the beginning of the session might prime learners regarding how they perceive the IL and EL of questions throughout the sessions. For both RC and PP, when learner ability was lower than the item difficulty in Session 1, learners persistently reported higher levels of EL and IL. In contrast, learners reported lower levels of EL and

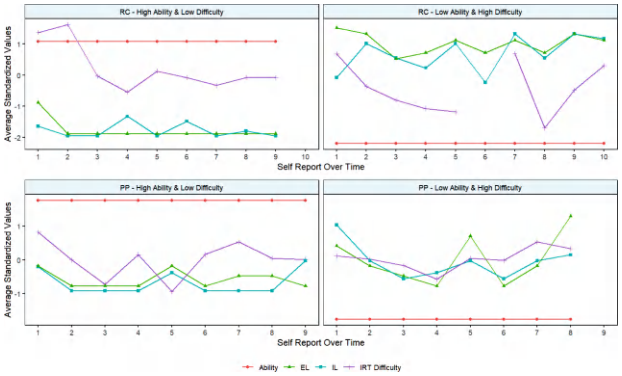


Fig. 3. Example items showing the mismatch between item difficulty and ability for RC and PP

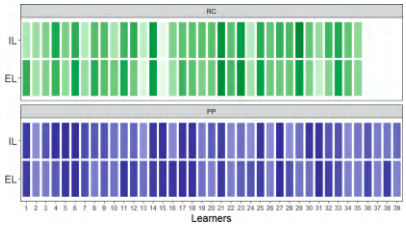


Fig. 4. Heatmap showing the relationship between IL and EL for RC and PP

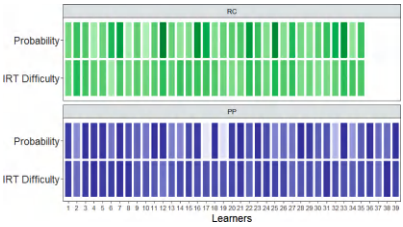


Fig. 5. Heatmap showing the relationship between difficulty and probability for RC and PP

IL when their ability was higher than item difficulty in Session 1, especially for PP (Fig. 3).

We also analyzed the heatmaps showing the relationship between IRT parameters and reported EL and IL levels per learner averaged across the data collection intervals (Fig. 4 and Fig. 5). We found that learners typically did not differentiate IL from EL, which is expected given that these two constructs are often correlated in theory. While average difficulty values were similar for learners, the probability of answering an item correctly showed more variations, highlighting the importance of considering learner ability.

4 Discussion and Implications for Practice

Our findings indicated that difficulty could be used for adjusting the IL and EL of a session. When considered at the individual level, our findings highlighted the importance of jointly representing ability with difficulty to adjust the IL of a task. Specifically, when items were more difficult than the learner's ability level, the learner reported higher levels of IL and EL. This is an important

finding for designing effective adaptive systems: system designers can use IRT to estimate item difficulty and learner ability to adjust for the IL of the task. An item bank with calibrated difficulty parameters could facilitate task selection tailored to learners' ability levels in such learning environments [2,3,11,18].

Our results also indicated another persistent issue in learner modeling and adaptive systems: the cold-start problem [34]. Selecting items adaptively requires predicting how learners will answer questions (i.e., learner ability); however, before a learner engages in any interaction, such knowledge about the learner cannot be known, creating problems for task selection and thus the IL level adjustment. This was a pronounced issue for the RC platform, where difficulty was initially higher than learner ability, and learners reported higher levels of IL and EL.

Unfortunately, using item difficulty at the individual level is unlikely to resolve this issue. This is highlighted in our finding that the relationship between difficulty, IL, and EL cannot be well distinguished by learners. This might be due to two reasons. First, in our analysis, we averaged item difficulty across sessions and found that learners received items at around the same difficulty levels. Thus, this finding might be due to the underlying task selection functions of the systems. Second, item difficulty is an objective measure of task complexity, whereas a self-report measure is the learner's experience of task complexity (i.e., perceived task complexity). This measurement nuance may help explain the discrepancy between item difficulty and IL at the individual level.

5 Conclusion

This study examined an alternative, easy, and non-intrusive approach for modeling IL in adaptive learning environments through IRT modeling. Using two different platforms, we showed that at the learning session level, item difficulty is a reasonable proxy for estimating the average IL self-reported by the learners per session. A remarkable finding for both platforms was the importance of considering ability and difficulty jointly for IL, as IL values changed based on the mismatch between ability and difficulty. While difficulty was a reasonable proxy for these two adaptive learning environments, the generalizability of difficulty being a proxy for IL should be investigated across different learning environments. We will continue to refine and explore the integration of this proxy measure to gain a more nuanced understanding of learner experience in complex educational settings. We will also investigate how adaptations based on estimated cognitive load may better support learning experiences and outcomes.

Acknowledgements. This work draws on research supported by the Social Sciences and Humanities Research Council (SSHRC) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Andrich, D.: The Rasch model explained. In: Applied Rasch Measurement: A Book of Exemplars (2005)
2. Bjorner, J.B., Kosinski, M., Ware, J.: Computerized adaptive testing and item banking. In: Assessing Quality of Life in Clinical Trials, pp. 95–112 (2005)
3. Bjorner, J.B., Chang, C.H., Thissen, D., Reeve, B.B.: Developing tailored instruments: item banking and computerized adaptive assessment. Qual. Life Res. **16**, 95–108 (2007)
4. Bratfisch, O., et al.: Perceived item-difficulty in three tests of intellectual performance capacity. Stockholm University (Sweden). Institute of Applied Psychology (1972)
5. Cai, M., Demmans Epp, C.: Modeling cognitive load and affect in interactive game-based learning based in physiological features. In: Adjunct Proceedings of the ACM International Conference on Interactive Media Experiences (IMX), pp. 9–14. ACM, Aveiro, Portugal (2022)
6. Cai, M., Demmans Epp, C.: Exploring the optimal time window for predicting cognitive load using physiological sensor data (2024). https://arxiv.org/abs/2406.13793
7. Cai, M., Rebolledo Mendez, G., Arevalo, G., Tang, S.S., Abdullah, Y.A., Demmans Epp, C.: Toward supporting adaptation: exploring affect's role in cognitive load when using a literacy game. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–17. ACM, Honolulu, HI, USA, May 2024. https://doi.org/10.1145/3613904.3642150
8. Caskurlu, S., et al.: Cognitive load and online course quality: insights from instructional designers in a higher education context. Br. J. Edu. Technol. **52**(2), 584–605 (2021)
9. Castro-Alonso, J.C., de Koning, B.B., Fiorella, L., Paas, F.: Five strategies for optimizing instructional materials: instructor-and learner-managed cognitive load. Educ. Psychol. Rev. **33**(4), 1379–1407 (2021)
10. Chandler, P., Sweller, J.: Cognitive load while learning to use a computer program. Appl. Cogn. Psychol. **10**(2), 151–170 (1996)
11. Dai, J., Gu, X., Zhu, J.: Personalized recommendation in the adaptive learning system: the role of adaptive testing technology. J. Educ. Comput. Res. **61**(3), 523–545 (2023)
12. Dindar, M., Kabakçı Yurdakul, I., Dönmez, F.İ: Measuring cognitive load in test items: static graphics versus animated graphics. J. Comput. Assist. Learn. **31**(2), 148–161 (2015)
13. Ehrich, J.F., Howard, S.J., Bokosmaty, S., Woodcock, S.: An item response modeling approach to cognitive load measurement. Front. Educ. **6**, 648324 (2021)
14. Gerjets, P., Scheiter, K., Catrambone, R.: Designing instructional examples to reduce intrinsic cognitive load: molar versus modular presentation of solution procedures. Instr. Sci. **32**, 33–58 (2004)
15. Huang, Y.H.: Influence of instructional design to manage intrinsic cognitive load on learning effectiveness. Eurasia J. Math. Sci. Technol. Educ. **14**(6), 2653–2668 (2018)
16. Jung, J.: Effects of task complexity and working memory capacity on l2 reading comprehension. System **74**, 21–37 (2018)
17. Klepsch, M., Schmitz, F., Seufert, T.: Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Front. Psychol. **8**, 1997 (2017)

18. Luo, H., Yang, X.: Efficiency of computerized adaptive testing with a cognitively designed item bank. Front. Psychol. **15**, 1353419 (2024)
19. Moreno, R.E., Park, B.: Cognitive load theory: historical development and relation to other theories. In: Cognitive Load Theory. Cambridge University Press (2010)
20. Osterlind, S.J., Wang, Z.: Item response theory in measurement, assessment, and evaluation for higher education. In: Handbook on Measurement, Assessment, and Evaluation in Higher Education, pp. 191–200. Routledge (2017)
21. Ouwehand, K., Kroef, A.V.D., Wong, J., Paas, F.: Measuring cognitive load: are there more valid alternatives to Likert rating scales? Front. Educ. **6**, 702616 (2021)
22. Paas, F., Renkl, A., Sweller, J.: Cognitive load theory and instructional design: recent developments. Educ. Psychol. **38**(1), 1–4 (2003)
23. Paas, F., Tuovinen, J.E., Tabbers, H., Van Gerven, P.W.: Cognitive load measurement as a means to advance cognitive load theory. Cogn. Load Theor., 63–71 (2016)
24. Révész, A., Michel, M., Gilabert, R.: Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: a validation study. Stud. Second. Lang. Acquis. **38**(4), 703–737 (2016)
25. Rikers, R.M., Van Gerven, P.W., Schmidt, H.G.: Cognitive load theory as a tool for expertise development. Instr. Sci. **32**, 173–182 (2004)
26. Sasayama, S.: Is a 'complex' task really complex? Validating the assumption of cognitive task complexity. Mod. Lang. J. **100**(1), 231–254 (2016)
27. Sentz, J., Stefaniak, J., Baaki, J., Eckhoff, A.: How do instructional designers manage learners' cognitive load? An examination of awareness and application of strategies. Educ. Tech. Res. Dev. **67**(1), 199–245 (2019). https://doi.org/10.1007/s11423-018-09640-5
28. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. Learn. Instr. **4**(4), 295–312 (1994)
29. Sweller, J., Chandler, P.: Why some material is difficult to learn. Cogn. Instr. **12**(3), 185–233 (1994)
30. Sweller, J., Chandler, P., Tierney, P., Cooper, M.: Cognitive load as a factor in the structuring of technical material. J. Exp. Psychol. Gen. **119**(2), 176 (1990)
31. Sweller, J., Van Merriënboer, J.J., Paas, F.G.: Cognitive architecture and instructional design. Educ. Psychol. Rev. **10**, 251–296 (1998)
32. VanLehn, K.: The behavior of tutoring systems. Int. J. Artif. Intell. Ed. **16**(3), 227–265 (2006). http://dl.acm.org/citation.cfm?id=1435351.1435353
33. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. **46**(4), 197–221 (2011)
34. van der Velde, M., Sense, F., Borst, J., van Rijn, H.: Alleviating the cold start problem in adaptive learning using data-driven difficulty estimates. Comput. Brain Behav. **4**(2), 231–249 (2021)
35. Zumbach, J., Mohraz, M.: Cognitive load in hypermedia reading comprehension: influence of text type and linearity. Comput. Hum. Behav. **24**(3), 875–887 (2008)



Artificial Intelligence Bias on English Language Learners in Automatic Scoring

Shuchen Guo[1,2(✉)], Yun Wang[3], Jichao Yu[3], Xuansheng Wu[3], Bilgehan Ayik[4], Field M. Watts[5], Ehsan Latif[2], Ninghao Liu[3], Lei Liu[5], and Xiaoming Zhai[2(✉)]

[1] Nanjing Normal University, Nanjing 210023, Jiangsu, China
69010@njnu.edu.cn
[2] AI4STEM Education Center, University of Georgia, Athens, GA 30602, USA
xiaoming.zhai@uga.edu
[3] School of Computing, University of Georgia, Athens, GA 30602, USA
[4] College of Education and Human Development, George Mason University, Fairfax, VA 22030, USA
[5] Educational Testing Service, Princeton, NJ, USA

Abstract. This study investigated potential scoring biases and disparities toward English Language Learners (ELLs) when using automatic scoring systems for middle school students' written responses to science assessments. We specifically focus on examining how unbalanced training data with ELLs contributes to scoring bias and disparities. We fine-tuned BERT with four datasets: responses from (1) ELLs, (2) non-ELLs, (3) a mixed dataset reflecting the real-world proportion of ELLs and non-ELLs (unbalanced), and (4) a balanced mixed dataset with equal representation of both groups. The study analyzed 21 assessment items: 10 items with about 30,000 ELL responses, five items with about 1,000 ELL responses, and six items with about 200 ELL responses. Scoring accuracy (Acc) was calculated and compared to identify bias using Friedman tests. We measured the Mean Score Gaps (MSGs) between ELLs and non-ELLs and then calculated the differences in MSGs generated through both the human and AI models to identify the scoring disparities. We found that no AI bias and distorted disparities between ELLs and non-ELLs were found when the training dataset was large enough (ELL $\approx 30,000$ and ELL $\approx 1,000$), but concerns could exist if the sample size is limited (ELL ≈ 200).

Keywords: Artificial Intelligence · Automatic scoring · English language learners · AI bias · AI disparities · Science assessment

1 Introduction

AI-based automatic assessment offers transformative potential for assessment practices [19] but raises concerns about scoring bias [5,9], particularly for English

S. Guo, Y. Wang, J. Yu and X. Wu—These authors contributed equally to this work.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 268–275, 2025.
https://doi.org/10.1007/978-3-031-98462-4_34

Language Learners (ELLs) [15]. Several empirical studies have found that AI scoring models, typically trained using natural language materials, possess the capability to score most student writings but vary in their ability to score ELLs' responses [10]. The potential bias can arise from the underrepresented ELL training samples and differences in writing styles [4,6,15]. While much research has been conducted on AI bias in language education, little is known about this issue in science assessments. Unlike essay writing tasks in language learning, student responses in science assessments are typically much shorter and prioritize science thinking rather than linguistic fluency [16]. Thus, this study aimed to answer the question: How do training samples impact the AI scoring bias and disparities for ELLs?

2 Related Work

2.1 Automatic Scoring in Science Assessment

In response to the call for promoting three-dimensional (3D) science learning, researchers advocate for developing assessment items that move beyond rote memorization tasks and require the application of knowledge and higher-order thinking skills. Thus, students are asked to write constructed responses to express their ideas [20]. Compared to written tasks in language learning, these constructed responses are typically short, which provides limited information for AI scoring. Students are often asked to briefly explain phenomena, make an argument, or provide their reasoning. Based on students' responses, AI scoring aims to assess student's performance in these practices (e.g., explanation, argumentation, scientific reasoning).

However, one challenge of using constructed response assessments in science teaching is the significant time commitment required for teachers to score student-constructed responses. Studies have demonstrated the potential of AI-based automatic scoring to assess student written responses in science assessment using various AI technologies, including machine learning [8] and large language models [7] for both text-based [7,8] and image-based responses [17]. However, there is still ample opportunity for improvement [20].

2.2 AI Bias and Disparities in Automatic Scoring

AI scoring bias is one of the major concerns researchers have when developing and using automatic scoring techniques [9,15]. AI bias refers to systematic rather than random error. To detect AI bias, studies use the scoring accuracy (Acc) of the AI model as an indicator. Meanwhile, AI disparities can be defined as the difference in Mean Score Gaps (MSGs) between human and AI scores for responses written by different groups of students [5]. When human MSGs and AI model MSGs are the same, it indicates no disparity. However, if the AI model MSG is significantly different from the human MSG, it means that using the AI scoring model can distort the scoring pattern for different groups of students [5].

2.3 AI Bias and Disparities for ELLs

Bias and uneven treatment of ELLs and non-ELLs have always been a concern in traditional educational environments. With the introduction of AI in science education, it is not surprising that educators and researchers have expressed concerns about whether the use of AI may exacerbate the existing bias in education that mostly impacts minoritized ELL groups [1]. Theoretically, the AI scoring model could have a statistical and computational bias if the training dataset lacks representation of certain groups since the AI scoring model would reflect the pattern of the training dataset [13]. When it comes to ELLs, although the proportion of ELL students is increasing rapidly in some multicultural and multilingual countries, the percentage of ELLs compared to non-ELLs is often significantly much lower (e.g., 1:9 in the US), which can result in unbalanced datasets for AI model training. Meanwhile, the characteristics of linguistic features of ELLs and non-ELLs are obvious, which makes it a potential pattern to influence AI model scoring on ELLs' responses. AI may struggle to interpret nonstandard or unexpected responses of ELLs, resulting in misinterpretation or underestimation of valid yet diverse expressions of knowledge from students with different English proficiency levels [1]. This situation can perpetuate existing inequities in educational assessment.

3 Methods

3.1 Dataset

The dataset used in this study is from the California Science Test (CAST), a state-administered assessment designed to measure students' knowledge and abilities in science. For this study, we utilized 21 constructed response items from the Grade 8 test. In response to each item, students are required to write some words or sentences. Each item is scored on a scale of 0, 1, or 2, with scores typically based on the correctness of the response. According to students' English learning proficiency, they are categorized into "English Language Learner (ELL)" and "Non-English Language Learner (Non-ELL)". There are approximately 15% ELLs in the entire student group.

3.2 Experimental Design

We conducted quantitative analysis by training and testing models under various settings to identify any potential bias. For the training phase, four training variants are applied for each item to train a BERT [2] model: only using responses from ELL students (ELL Model), only using responses from non-ELL students (Non-ELL Model), using mixed data with both ELL and non-ELL student responses which reflect the real-world distribution (Unbalanced Mixed Model), and a balanced mixed dataset with equal representation of both groups (Balanced Mixed Model). Note that, for a fair comparison, we down-sample datasets to the same size as that of ELL students so that the total number of

training samples will not impact the results across these four variants. During testing, each model was tested using unbalanced mixed data, which represents the authentic composition of ELL and non-ELL students.

The number of respondents varied across test items. Based on the number of responses, we categorized the items into three groups: Group 1 has over 200,000 total responses (around 30,000 ELLs and 180,000 non-ELLs). Group 2 has about 10,000 total responses (more than 1,000 ELLs and around 10,000 non-ELLs). Group 3 has about 1,800 total responses (around 200 ELLs and 1,500 non-ELLs). Each group includes 6, 5, and 10 items, respectively. These groups were set up to compare the performance of BERT-base [2] under varying data scale conditions. Bias and disparities were calculated within each group based on the test results for all items. Finally, comparing results across the groups helps assess the impact of the data scale on AI scoring bias and disparities related to English proficiency.

3.3 Measure

Building on Latif et al.'s analytical approach [5], we adopted model accuracy and Mean Score Gap (MSG) as key indicators for measuring potential bias and disparities. Model accuracy refers to the percentage of correctly scored students' responses out of the total testing responses. In this research, we applied the Friedman test [3] to analyze the difference in various models' performance on mixed training data. A significant difference ($p < 0.05$) indicates a potential bias. MSG refers to the gap between the average scores of two groups of students [5]. In this research, the MSG in human-coded scores between the ELL and non-ELL students were taken as a baseline for evaluating the machine-generated scores (Fig. 1). If machine-generated MSG is significantly larger or lower than the human-scored MSG, it indicates the disparities between these ELL and non-ELL students have changed. Due to the limited items used in the study, the Wilcoxon signed-rank test was used to detect whether there are significant differences among MSG among different AI scoring models and human scorers.

4 Results

4.1 AI Bias: Accuracy Comparison

Model accuracy and Friedman test results are presented in Table 1. Scoring performance declined with smaller datasets. No significant differences emerged among models with larger ELL samples (ELL $\approx 30,000$: $p = 0.066$; ELL $\approx 1,000$: $p = 0.908$). However, a significant difference occurred with about 200 ELL responses ($p = 0.006$). Results suggest the accuracy of different models can differ significantly only in small datasets.

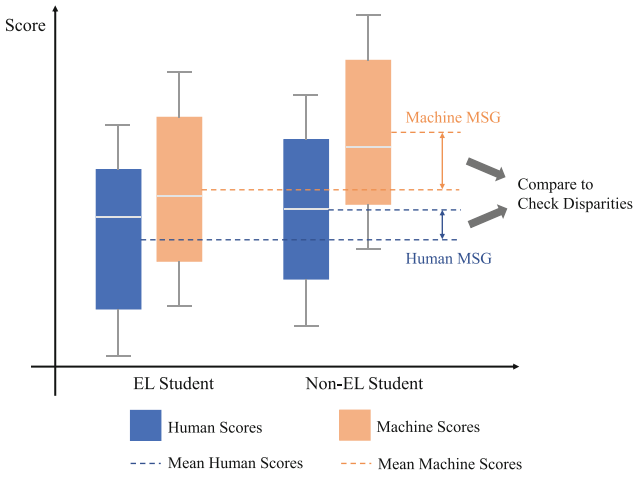


Fig. 1. Analyzing disparity by comparing MSGs between human and machine scores.

Table 1. Accuracy among different models across different group sizes.

Group	Model Type	Accuracy (*Mdn (Range)*)	*p*-value
Group 1(ELL $\approx 30,000$)	Unbalanced mixed model	0.9313 (0.8829–0.9565)	0.066
	Balanced mixed model	0.9294 (0.8812–0.9538)	
	ELL model	0.9200 (0.8560–0.9564)	
	Non-ELL model	0.9300 (0.8890–0.9592)	
Group 2(ELL $\approx 1,000$)	Unbalanced mixed model	0.8189 (0.7747–0.9715)	0.908
	Balanced mixed model	0.8657 (0.7568–0.9573)	
	ELL model	0.8018 (0.6674–0.9614)	
	Non-ELL model	0.8189 (0.7779–0.9606)	
Group 3(ELL ≈ 200)	Unbalanced mixed model	0.7610 (0.6500–0.8722)	0.006**
	Balanced mixed model	0.7639 (0.6243–0.8833)	
	ELL model	0.5389 (0.4167–0.8944)	
	Non-ELL model	0.7425 (0.6944–0.9000)	

Note. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.2 AI Disparities: Mean Score Gap Difference

Table 2 shows MSG comparisons between AI models and human scoring across data scales. For large ELL samples (ELL $\approx 30,000$ and ELL $\approx 1,000$), no significant MSG differences emerged between AI and human scoring. With about 200 ELL responses, no MSG difference was found between the human scoring and the machine scoring for the unbalanced mixed and non-ELL models. Significant differences were observed between the human scoring and the machine scoring for the balanced mixed model ($p = 0.004$) and the ELL model ($p = 0.002$). Also, adding more ELLs to the training data tends to decrease the MSG. These

findings indicate that sufficient data eliminates scoring disparities between ELL and non-ELL groups. Small datasets require training data composition matching natural distributions for optimal human-AI MSG alignment.

Table 2. MSG scores among different models across different group sizes and the Wilcoxon test results for the comparisons between the human scoring and each model.

Group	Model Type	MSG (*Mdn (Range)*)	p-value
Group 1(ELL $\approx 30,000$)	Human	0.5389 (0.2883–0.7116)	—
	Unbalanced mixed model	0.5428 (0.2750–0.7266)	0.438
	Balanced mixed model	0.5398 (0.2887–0.7221)	0.219
	ELL model	0.5213 (0.2427–0.7192)	0.563
	Non-ELL model	0.5487 (0.2903–0.7202)	0.313
Group 2(ELL $\approx 1,000$)	Human	0.3287 (0.2883–0.5786)	—
	Unbalanced mixed model	0.3447 (0.3175–0.6137)	0.188
	Balanced mixed model	0.3112 (0.2786–0.6438)	0.813
	ELL model	0.2831 (0.2574–0.5693)	0.188
	Non-ELL model	0.3372 (0.3016–0.5856)	0.313
Group 3(ELL ≈ 200)	Human	0.4899 (0.2919–0.7615)	—
	Unbalanced mixed model	0.4663 (0.1874–0.7611)	0.922
	Balanced mixed model	0.3732 (0.0305–0.5955)	0.004**
	ELL model	0.1427 (0.0000–0.5976)	0.002**
	Non-ELL model	0.4314 (0.1394–0.6967)	0.131

Note. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5 Discussion and Conclusion

While existing work on AI scoring bias primarily focuses on essay writing tasks, where AI often favors non-ELLs [10,12,15], this study contributes to the field by analyzing the AI bias in science constructed-response items. The study investigates AI scoring bias for ELLs in science assessments by comparing model accuracy and MSG across different training datasets with varying proportions of ELL and non-ELL responses. The results show that when training data is large enough, AI scoring models do not result in bias or distort the disparities between ELLs and non-ELLs. However, smaller datasets may introduce bias. For disparity measures, training models with naturally representative ELL/non-ELL ratios yield undistorted results.

Our findings are unsurprising and contribute to the ease of unnecessary fears of AI bias, which are critical for all populations to benefit from AI [18]. Unlike essay writing assessments, science constructed-response items are typically shorter and less dependent on language proficiency, reducing the impact

of language proficiency on AI scoring. Our findings therefore suggest that AI scoring neither amplifies existing bias nor distorts true disparities for ELLs. Similarly, Wilson and Huang (2024) found that their AI scoring system did not exhibit unique bias against ELLs but rather replicated biases present in human scoring [15]. Thus, the results could be the potential consequence of the fact that machines imitate humans, as Nesbitt and Quesen also argued [11]. Until more empirical evidence emerges, we recommend using adequately sized datasets that reflect real-world ELL proportions as training data to minimize additional scoring bias in science assessment.

6 Limitation and Future Study

Although the results of this study show that AI scoring generally does not lead to biased results for ELLs and non-ELLs in science constructed response assessments, it is important to note that AI bias does not solely stem from the factor investigated (i.e., the training dataset in this study). Schwartz et al. (2022) [13] classify AI biases into three groups: systemic biases (e.g., racism, sexism) are embedded in datasets due to historical inequities; statistical and computational biases are caused by the lack of representation or misrepresentation of certain groups; human biases (e.g., prompt design, output interpretation), emerge during user and AI interaction. Therefore, future studies should continue to address the potential AI bias for ELLs, considering all these possible dimensions. Moreover, future research could examine bias across finer-grained English proficiency levels (e.g., WIDA tiers) [14] instead of binary codes used in this study. Finally, while it is crucial to explore the potential bias in automated scoring, it is even more important to reconsider the possible embedded biases in the entire assessment process, from creating items to rubric design and the underlying human assumptions about race, ethnicity, and language of those who participate in the assessment [11]. Future research is needed to provide a more comprehensive perspective on more inclusive science assessments that accurately capture all students' understanding of science concepts.

Acknowledgment. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305C240010 (PI Zhai) and R305A240356 (PI Liu). The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. Additionally, we acknowledge the use of OpenAI's ChatGPT for polishing the language and enhancing the clarity of this manuscript.

References

1. Cheuk, T.: Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. Sci. Educ. **105**(5), 825–836 (2021)
2. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018). http://arxiv.org/abs/1810.04805

3. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. **32**(200), 675–701 (1937). https://doi.org/10.1080/01621459.1937.10503522
4. Jourdain, L., Sharma, S.V.: Language challenges in mathematics education for English language learners: a literature review. Waikato J. Educ. **21**(2) (2016)
5. Latif, E., Zhai, X., Liu, L.: AI gender bias, disparities, and fairness: Does training data matter? arXiv preprint arXiv:2312.10833 (2023). https://arxiv.org/abs/2312.10833
6. LeClair, C., Doll, B., Osborn, A., Jones, K.: English language learners' and non-English language learners' perceptions of the classroom environment. Psychol. Sch. **46**(6), 568–577 (2009)
7. Lee, G.G., Latif, E., Wu, X., Liu, N., Zhai, X.: Applying large language models and chain-of-thought for automatic scoring. Comput. Educ. Artif. Intell. **6**, 100213 (2024)
8. Lee, H.S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., Liu, O.L.: Automated text scoring and real-time adjustable feedback: supporting revision of scientific arguments involving uncertainty. Sci. Educ. **103**(3), 590–622 (2019)
9. Li, C., Xing, W., Leite, W.: Using fair AI to predict students' math learning outcomes in an online platform. Interact. Learn. Environ. **32**(3), 1117–1136 (2024)
10. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., Zou, J.: GPT detectors are biased against non-native English writers. Patterns **4**(7) (2023)
11. Nesbitt, J., Quesen, S.: Machines imitating humans: scoring systems and the risk of bias (2023). https://www.wested.org/blog/machines-imitating-humans-scoring-systems-and-the-risk-of-bias/. Accessed 19 Feb 2025
12. Reed, D.K., Mercer, S.H.: Potential scoring and predictive bias in interim and summative writing assessments. School Psychol. **38**(4), 215 (2023)
13. Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P.: Towards a standard for identifying and managing bias in artificial intelligence, vol. 3. US Department of Commerce, National Institute of Standards and Technology (2022)
14. WIDA Consortium: WIDA English Language Development Standards Framework, 2020 edn. Technical report, Board of Regents of the University of Wisconsin-Madison (2020). https://wida.wisc.edu/resources/wida-english-language-development-standards-framework-2020-edition
15. Wilson, J., Huang, Y.: Validity of automated essay scores for elementary-age English language learners: evidence of bias? Assess. Writ. **60**, 100815 (2024)
16. Zhai, X.: AI and machine learning for next generation science assessments. In: Machine Learning, Natural Language Processing, and Psychometrics, p. 201. IAP (2024)
17. Zhai, X., He, P., Krajcik, J.: Applying machine learning to automatically assess scientific models. J. Res. Sci. Teach. **59**(10), 1765–1794 (2022)
18. Zhai, X., Krajcik, J.: Pseudo artificial intelligence bias, pp. 568–578 (2024). https://doi.org/10.1093/oso/9780198882077.003.0025
19. Zhai, X., Nehm, R.H.: AI and formative assessment: the train has left the station. J. Res. Sci. Teach. **60**(6), 1390–1398 (2023)
20. Zhai, X., Yin, Y., Pellegrino, J.W., Haudek, K.C., Shi, L.: Applying machine learning in science assessment: a systematic review. Stud. Sci. Educ. **56**(1), 111–151 (2020)



Using Self-regulated Learning Theory to Inform the Design of Educational Recommender Systems for Introductory Programming

Jordan Barria-Pineda(✉), Deniz Sonmez Unal, Kamil Akhuseyinoglu, Peter Brusilovsky, and Erin Walker

University of Pittsburgh, Pittsburgh, PA 15260, USA
{jab464,des204,kaa108,peterb,eawalker}@pitt.edu

Abstract. Many educational recommender systems (EdRecSys) rely on commercial recommendation strategies that emphasize content relevance while neglecting learners' views on recommendation effectiveness. To address this, we conducted a co-design study with computer science students in an introductory programming course to explore their vision of an ideal EdRecSys. The subjects shared preferences and concerns related to three areas: recommendation approaches, transparency, and control. We used Zimmerman's model of self-regulated learning to contextualize their expectations within a broader educational framework. Findings offer actionable insights for designing learner-centered AIED systems that foster engagement, agency, and self-regulation.

Keywords: Educational Recommender Systems · Self-Regulated Learning · User-Centered Design

1 Introduction and Background

Educational Recommender Systems (EdRecSys) are AI-based tools that personalize learning by suggesting resources or connecting students with peers [25]. Early EdRecSys adapted general recommender techniques with some success, but applying them within learning contexts poses unique design challenges. Unlike commercial recommender systems that prioritize user preferences, EdRecSys must account for educational factors such as learners' competencies, goals, and motivation [20]. Although other AI-based learning tools have successfully incorporated educational theories into their design [13,22], EdRecSys lack a systematic, theory-driven framework to guide their design. Existing efforts are limited and typically focus on isolated aspects such as goal recommendation [10]. One promising framework is self-regulated learning (SRL) [33]. SRL consists of three phases: (1) *forethought* (planning and learning goal (LG) setting), (2) *performance control* (strategy application), and (3) *self-reflection* (evaluation and adjustment). It improves learning experience, especially in tech-driven environments with limited tutor support [19]. Although SRL has guided many educational technologies [4,6], it remains underused in EdRecSys.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 276–284, 2025.
https://doi.org/10.1007/978-3-031-98462-4_35

Inspired by commercial systems, EdRecSys research has begun exploring transparency and control, which can support learners' comprehension, persistence, and agency [1,2,5,26]. While such benefits rely on user-centered design in other domains [21,32], few EdRecSys studies take this approach [26,28], despite growing evidence that learner input can enhance these features [7,18]. In this paper, we study how to design transparency and control, grounded in both student input and learning theory. With a user-centered approach, we engaged learners in co-designing an interactive EdRecSys for programming. We explored **(RQ1)** What strategies do students think an effective EdRecSys should follow? **(RQ2)** How do they perceive features that enhance transparency? **(RQ3)** What are their views on controlling recommendation generation? To contextualize students' input, we mapped their responses to Zimmerman's SRL model [33]. Our work advances AI-based EdRecSys by grounding their design in learning theory.

2 Co-design Study

Participants and Protocol: We recruited eight students (2 female, 6 male) taking an introductory Java programming class at a U.S. public university. Recruitment was carried out through classroom visits. Each subject provided informed consent and completed a one-hour, one-on-one co-design session for $20 compensation. The group varied in academic standing (1st–3rd year) and included two non-CS majors. All had prior experience with educational platforms and half had formal high school programming instruction. After collecting demographic information, we introduced the *Mastery Grids* system [23] used in the class, explaining its purpose (adaptive programming practice) and showing a mock screenshot (Fig. 1) since the use of the system was optional and only 3 of 8 participants had previously accessed it. The screenshot illustrated the need for personalized recommendations among 70 practice activities. The students were asked what features a good EdRecSys should have, which aspects of their *learning context* it should consider and whether adaptation to changing *learning goals* (LG) was important. The students ranked a list of LGs extracted from the literature [15,29] according to their study priorities. Finally, we explored their views on transparency and control, asking what information they desired about the recommendation process, whether they would like to modify it, and inviting them to sketch low-fidelity prototypes while thinking aloud.

Coding and Thematic Analysis. After revising the automatically transcribed recordings, the first and second authors independently coded two transcripts to develop initial codebooks. We used a hybrid approach combining inductive (interview protocol and research questions) and deductive coding (Zimmerman's SRL theory) [31] in order to identify students' awareness of their SRL processes, and ways to align design suggestions with appropriate SRL stages within an EdRecSys. After coding two transcripts separately, we merged our findings into a unified codebook, ensuring a manageable number of codes (fewer than 15) and an appropriate abstraction level. The final codebook comprised seven

main themes: **SRL phase**, **need for adaptation**, **suggestions for EdRecSys design**, **transparency**, **learner control**, **proficiency assessment**, and **GUI suggestions**. We focus on three major themes aligned with our research questions: **suggestions for EdRecSys design** (RQ1), **transparency** (RQ2—value of interpretable recommendations), and **learner control** over the EdRecSys (RQ3). To contextualize findings, we analyzed intersections between these themes and the **SRL phase** theme, identifying the specific learning stages where certain student needs arise. Table 1 shows a summary of our findings.

3 Findings

Recommendation Approach. From students' responses to the short LGs survey (final rank from more to less important: 1-Fix misconceptions, 2-Fill knowledge gaps, 3-Prepare for a test), we confirmed that they were aware of different study goals and used them to guide their study sessions (i.e., **forethought** phase). They would like an EdRecSys platform to suggest content that matches those specific goals. For example, P2 mentioned: *"If there's no test coming up next. Let's say you've completed the homework assignments, you're doing well, I think you go to acquiring new knowledge becomes number one... because that will help you get a deeper understanding of what you're doing"*. For choosing learning content (i.e., **performance control** phase), most mentioned that the system should recommend *a path* that reflects the natural learning progression, which goes from foundations and worked examples to more complex problems. P1 mentioned: *"it would recommend a couple of examples, how to go through the*

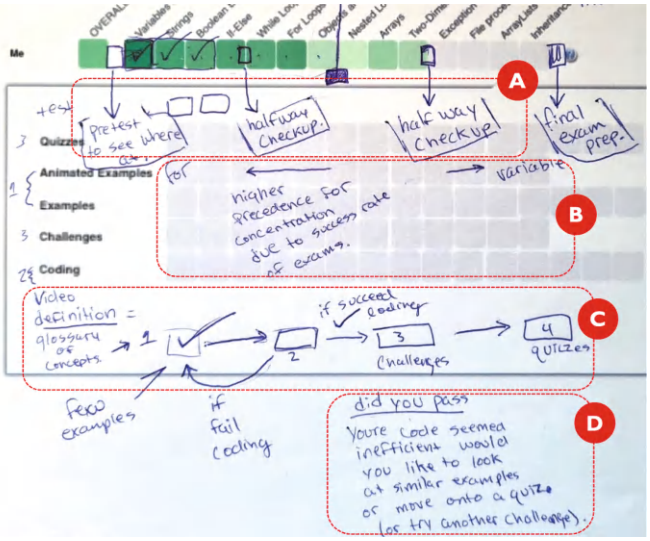


Fig. 1. P1's sketch showing features for a good EdRecSys A) Testing for better student model B) Content allocation C) Learning path recommend. D) Student-AI interaction

Table 1. Summary of findings contextualized by Zimmerman's SRL framework

SRL phase	Findings
Forethought	**Rec. approaches:** Learning context needs to be known (current LG, learning strategy, upcoming evaluations, current performance and struggles). **Transparency:** Up-to-date learning context needs to be visible by students. **Control:** Learners need to state their learning goal for the session, decide on the level of challenge they would like to experience, and be prompted about their confidence.
Performance Control	**Rec. approaches:** Recommending a progressive learning path (i.e., sequence of materials to follow based on their current goal and proficiency). **Transparency:** Explain why a) recommended content should be attempted and b) non-recommended content should not be the learner's priority. Keep them updated about the current level of LG completion after each attempt. **Control:** AI support when struggling with the materials (hints, remedial recommendations). Allow learners to modify system beliefs about their knowledge when wrong.
Self-reflection	**Rec. approaches**: Formal instances for testing student knowledge (and to feed the student model). **Transparency**: Summary about learning goal completion. **Control**: Allowing students to modify the student model when estimations are wrong.

examples... and then after, say, you're getting used to it, then it would quiz you on it, and if you don't do well on a quiz, keep kind of almost doing the cycle" (see C in Fig. 1). Finally, subjects described moments of **self-reflection**, especially when deciding next steps. They suggested that system feedback—such as prompting a topic switch after reaching proficiency could support these decisions. A natural point for encouraging reflection is after completing a problem, particularly if recent performance reveals struggles. P3 stated *"once you finish, like a section or whatever, they can use that to fill the knowledge gaps"*, suggesting that another good moment for self-reflection is after completing a practice section. Subjects stated that they expected to be "tested" through short quizzes and using their performance as input to generate recommendations, e.g., P4 said *"you took a quiz and then it'll just be, 1-2-3...maybe if you've got these two right, but this one wrong, it was like highlight, and then say, like that topic, you need help on, and then just I guess list links to learn about it"*. P1 suggested adding 'checkpoints' across topics to assess their progress (see A in Fig. 1).

Transparency. Subjects emphasized that they wanted transparent access to data on recent grades, detected misconceptions, and upcoming tests, to better define their study goals and guide the EdRecSys (**forethought** phase). Once recommendations were made and they began practicing, students valued explanations that clarified why specific materials were suggested (**performance control** phase). They highlighted two useful indicators: proficiency (e.g. accuracy in previous attempts) and effort (e.g., time spent). P1 mentioned: *"it would show you the one [test problem] that you have the lowest grade on or like the lowest average and it could recommend those to you, rather than the ones that you did well on"*. In the same context, P4 suggested that displaying areas that need improvement within the student model could help connect recommendations to specific weaknesses, making it clearer why certain recommendations were generated. 3 of 8 subjects mentioned *effort*-based indicators, such as time spent or problems completed, which were seen as signs of ongoing struggles

regardless of performance. P1 stated: *"You're doing bad in this one? You've spent more time on this one. That's why we're gonna help you do like you. OK, you spent less time on this one, but you did well on it. So we'll push that to the end"*. P2 sketched a prototype where the system alerted the student about spending too much time on a topic. P3 mentioned that the system should also communicate **why the content has not been suggested to them**. When envisioning recommendations' explanations, students prefer references to identifiable learning episodes, such as coding problem performance (see D in Fig. 1) or external measures like exam scores, e.g., P2 stated: *"So if I was unsure why am I doing poorly in this? Oh, well, when you took this exam, this practice exam, you did this wrong. That's why you should reinforce it"*. Finally, to encourage **self-reflection**, students wanted to know their progress toward the LG set during the forethought phase. P1 stated *"like, if you're doing well, and you're above 80% in that one subject, and there's now a lower one, then it would go to that next"*. This highlights the importance of allowing students to track their progress towards their goal (e.g., 80% proficiency). This reflection then informs the next SRL cycle, where students can adjust their goal or move to a new topic.

Learner Control. Students wanted recommendations aligned with their LGs and the ability to adjust them over time, reflecting the **forethought** phase. They also emphasized adjusting factors like *content difficulty*, e.g., P5 said *"and then if they [materials] seem too simple, I might skip like a little bit ahead, but I probably just go in order from simple to complex"*. Regarding the **performance control** phase, students stated that they would like to "talk" with the system when struggling with the learning content. P1 stated that the system should prompt *"Would you like to try again? Or like with a different set of code, or would you like to see previous examples of how this code works?"* after failing to solve a problem repeatedly. Subjects stated that they would like to be prompted by the system whenever they are struggling so that they can be offered "remedial" options (see D in Fig. 1), but in the end **they should have the final say**.

Addressing system inaccuracies was considered crucial, aligning with the **self-reflection** phase. We identified two ways students felt at ease providing explicit feedback to the system: **student-initiated** and **agent-initiated**. P5 mentioned *"maybe if I wanted to change it, I could just go in and solve a problem, and then it [the system] would understand like, you really do understand this"*. This portrays a learner-initiated action without directly interacting with the agent, intending to update the (inaccurate) student model. On the other hand, some students said that the system should be more proactive by asking the students if they are sure that they know a certain concept, P7 stated *"so if you just go to inheritance...'How comfortable are you?' And then if they're comfortable with it then so they say: 'I'm 5 comfortable with, yes, really comfortable' then you just throw coding at them. But if they say like 'something that indicates they need to review' then going through the examples would be great"*. Subjects said the system should let learners reflect on their mastery and correct the student model, especially when errors result from "slips" rather than misconceptions.

4 Discussion and Conclusion

We gained several insights that could inform *EdRecSys* design in support of SRL:

1. LG setting and tracking are crucial: Many students struggle to set meaningful goals on their own. Adaptive learning platforms effectively scaffold these processes by recognizing student goals/plans based on their interactions [14]. Hence, EdRecSys should offer features that allow students to set and modify goals starting a study session (**forethought** phase). This can also act as explicit feedback to improve the recommendation process. Complementing this with LG progress visualization [12] is expected for SRL support. This *goal awareness component* should provide learners with feedback on their performance, aligning with cognitive and motivational monitoring principles [30].

2. Recommending learning paths: Students reported learning sequentially, moving from foundational to advanced content (e.g., *"from worked examples to quizzes and coding"* (aligning with prior research [3])). This also aligns with SRL's **performance control** phase and suggests EdRecSys should recommend meaningful content sequences that support goals set in the **forethought phase**. Incorporating interactive and practical components to support these trajectories was shown to enhance student engagement [11].

3. More "tangible" explainability: Learners preferred explanations linked to **concrete experiences** like struggles on specific problems or low scores on a certain test, rather than abstract metrics like mastery values (commonly used in OLMs). To support reflection during **performance control** phase, EdRecSys should highlight clear indicators (e.g., exam scores, time spent) when explaining recommendations. EdRecSys should also explain why certain materials **are not recommended**, so learners can judge recommendations' quality.

4. Enabling Student-AI collaboration to improve EdRecSys: The students were open to correct the system's misjudgments and improve recommendation relevance, similar to scrutable and negotiable OLMs [9,17]. This highlights the value of learner control in EdRecSys, an area where broader AI has advanced beyond traditional AIED [8]. To support this, systems should offer editable learner models and prompt students to refine them, especially when students' activity in the system is scarce [16].

5. Optional control to support student agency: Subjects preferred varying degrees of control. Although they appreciated guidance through recommendations, they wanted to remain the ultimate decision makers, especially when they disagreed with the system's suggestions. This aligns with Brod et al. [7], suggesting that effective learning technologies should adapt the level of control based on learners' metacognitive skills and needs. Similarly, Jin et al. [16] highlight supporting learner activeness as a key element of promoting SRL. While prior research has shown benefits of control in automated learning systems [24], some students noted potential downsides, such as inaccurate self-assessment.

We found that SRL theory can offer a framework for designing effective EdRecSys features. It hints at the support learners need on each learning stage,

and its link to the EdRecSys operation (e.g., LG setting to guide the recommendation algorithm in forethought phase). Despite this, balancing student input with theoretical evidence remains an open challenge [27]. While subjects valued transparency and control, more research is needed to define how much of each is helpful across various SRL phases without compromising agency/trust. The small sample size is a limitation of our study, although the co-design approach provided rich in-depth insights. In conclusion, our findings lay the foundation and motivate the design of EdRecSys by reconciling learning research principles with actual student interests/needs.

Acknowledgments. This work was partially funded by CONICYT PFCHA/ Doctorado Becas Chile/2018-72190680

References

1. Abdi, S., Khosravi, H., Sadiq, S., Gasevic, D.: Complementing educational recommender systems with open learner models. In: Proceedings of LAK 2020, pp. 360–365 (2020)
2. Ain, Q.U., Chatti, M.A., Tsoplefack, W.K., Alatrash, R., Joarder, S.: Designing and evaluating an educational recommender system with different levels of user control. In: Proceedings of the 11th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (2024)
3. Atkinson, R.K., Renkl, A., Merrill, M.M.: Transitioning from studying examples to solving problems: effects of self-explanation prompts and fading worked-out steps. J. Educ. Psychol. **95**(4), 774 (2003)
4. Azevedo, R., Witherspoon, A., Chauncey, A., Burkett, C., Fike, A.: Metatutor: a metacognitive tool for enhancing self-regulated learning. In: 2009 AAAI Fall Symposium Series (2009)
5. Barria-Pineda, J., Akhuseyinoglu, K., Želem Ćelap, S., Brusilovsky, P., Milicevic, A.K., Ivanovic, M.: Explainable recommendations in a personalized programming practice system. In: Artificial Intelligence in Education, pp. 64–76 (2021)
6. Biswas, G., Jeong, H., Kinnebrew, J.S., Sulcer, B., Roscoe, R.: Measuring self-regulated learning skills through social interactions in a teachable agent environment. Res. Pract. Technol. Enhanc. Learn. **5**(02), 123–152 (2010)
7. Brod, G., Kucirkova, N., Shepherd, J., Jolles, D., Molenaar, I.: Agency in educational technology: interdisciplinary perspectives and implications for learning design. Educ. Psychol. Rev. **35**(1), 25 (2023)
8. Brusilovsky, P.: AI in education, learner control, and human-AI collaboration. Int. J. Artif. Intell. Educ. **34**(1), 122–135 (2024)
9. Bull, S.: Negotiated learner modelling to maintain today's learner models. Res. Pract. Technol. Enhanc. Learn. **11**(1), 10 (2016)
10. Capuano, N., Gaeta, M., Ritrovato, P., Salerno, S.: Elicitation of latent learning needs through learning goals recommendation. Comput. Hum. Behav. **30**, 663–673 (2014)
11. Erickson, M., Marks, D., Karcher, E., et al.: Characterizing student engagement with hands-on, problem-based, and lecture activities in an introductory college course. Teach. Learn. Inq. **8**(1), 138–153 (2020)

12. Fessl, A., Maitz, K., Dennerlein, S., Pammer-Schindler, V.: The impact of explicating learning goals on teaching and learning in higher education: evaluating a learning goal visualization. In: Technology-Enhanced Learning for a Free, Safe, and Sustainable World, pp. 1–15. Springer (2021)
13. Gibson, D., Kovanovic, V., Ifenthaler, D., Dexter, S., Feng, S.: Learning theories for artificial intelligence promoting learning processes. Br. J. Edu. Technol. **54**(5), 1125–1146 (2023)
14. Goslen, A., Carpenter, D., Rowe, J.P., Henderson, N., Azevedo, R., Lester, J.: Leveraging student goal setting for real-time plan recognition in game-based learning. In: International Conference on Artificial Intelligence in Education, pp. 78–89. Springer (2022)
15. Janssen, O., Prins, J.: Goal orientations and the seeking of different types of feedback information. J. Occup. Organ. Psychol. **80**(2), 235–249 (2007)
16. Jin, S.H., Im, K., Yoo, M., Roll, I., Seo, K.: Supporting students' self-regulated learning in online learning using artificial intelligence applications. Int. J. Educ. Technol. High. Educ. **20**(1), 37 (2023)
17. Kay, J.: Scrutable adaptation: because we can and must. In: Wade, V., Ashman, H., Smyth, B. (eds.) 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2006), pp. 11–19. Springer (2006)
18. Khosravi, H., et al.: Explainable artificial intelligence in education. Comput. Educ. Artif. Intell. **3**, 100074 (2022)
19. Kizilcec, R.F., Saltarelli, A.J., Reich, J., Cohen, G.L.: Closing global achievement gaps in MOOCs. Science **355**(6322), 251–252 (2017)
20. Klašnja-Milićević, A., Ivanović, M., Nanopoulos, A.: Recommender systems in e-learning environments: a survey of the state-of-the-art and possible extensions. Artif. Intell. Rev. **44**(4), 571–604 (2015)
21. Knijnenburg, B.P., Willemsen, M.C.: Evaluating recommender systems with user experiments. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 309–352. Springer, Boston (2015)
22. Lajoie, S.P., Li, S.: Theory-driven design of AIED systems for enhanced interaction and problem-solving. In: Handbook of Artificial Intelligence in Education, pp. 229–249. Edward Elgar Publishing (2023)
23. Loboda, T.D., Guerra, J., Hosseini, R., Brusilovsky, P.: Mastery grids: an open-source social educational progress visualization. In: Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education, p. 357 (2014)
24. Long, Y., Aleven, V.: Enhancing learning outcomes through self-regulated learning support with an Open Learner Model. User Model. User-Adap. Inter. **27**(1), 55–88 (2017)
25. Manouselis, N., Drachsler, H., Verbert, K., Duval, E.: Recommender Systems for Learning. Springer Briefs in Electrical and Computer Engineering. Springer (2013)
26. Ooge, J., Dereu, L., Verbert, K.: Steering recommendations and visualising its impact: effects on adolescents' trust in e-learning platforms. In: Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 156–170 (2023)
27. Rau, M.A., Aleven, V., Rummel, N., Rohrbach, S.: Why interactive learning environments can have it all: resolving design conflicts between competing goals. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 109–118 (2013)
28. Santos, O.C., Boticario, J.G.: Practical guidelines for designing and evaluating educationally oriented recommendations. Comput. Educ. **81**, 354–374 (2015)
29. Schunk, D.H.: Goal setting and self-efficacy during self-regulated learning. Educ. Psychol. **25**(1), 71–86 (1990)

30. Schunk, D.H.: Self-regulated learning: the educational legacy of Paul R. Pintrich. Educ. Psychol. **40**(2), 85–94 (2005)
31. Thomas, D.R.: A general inductive approach for analyzing qualitative evaluation data. Am. J. Eval. **27**, 237–246 (2006)
32. Tintarev, N., Masthoff, J.: Effective explanations of recommendations: user-centered design. In: Proceedings of the 2007 ACM Conference on Recommender Systems, pp. 153–156. ACM (2007)
33. Zimmerman, B.J.: Attaining self-regulation: a social cognitive perspective. In: Handbook of Self-regulation, pp. 13–39 (2000)



StoryLab: Empowering Personalized Learning for Children Through Teacher-Guided Multimodal Story Generation

Zhaohui Li[1(✉)], Feiwen Xiao[2], Jiaju Lin[2], Xiaohan Zou[2], Qingxiao Zheng[1], and Jinjun Xiong[1]

[1] University at Buffalo, Buffalo, NY 14068, USA
{zli253,qingxiao,jinjun}@buffalo.edu
[2] The Pennsylvania State University, University Park, PA 16802, USA
{ffx5014,jjl7137,xfz5266}@psu.edu

Abstract. Personalized story reading enhances child literacy by aligning content with individual interests, backgrounds, and developmental needs. However, implementing such systems presents challenges, including data privacy concerns, the need for culturally diverse materials, limited resources, and balancing personalization with standardized benchmark objectives. To address these challenges, we introduce **StoryLab**, a multimodal system designed for K-2 teachers and students. The system leverages advanced generative AI to integrate students' personal interests with teacher-defined learning objectives to generate comprehensive learning materials, including story text, illustrated figures, vocabulary support, and a consistent narrative voice. A teacher-in-the-loop design ensures pedagogical alignment and trust. Evaluations demonstrate StoryLab's effectiveness and usability, positioning it as a promising and scalable tool for personalized literacy instruction.

Keywords: Early Literacy · Story Generation · Personalized Learning · Large Language Models · Multimodal Learning Environment

1 Introduction

Personalized reading experiences have been shown to enhance engagement, comprehension, and literacy skill acquisition in young learners [15]. By tailoring content to children's unique interests, cultural backgrounds, and developmental needs, educators can foster intrinsic motivation and bridge gaps in foundational skills development such as phonics, fluency, and vocabulary [7]. However, scalable implementation of personalized literacy tools remains a challenge. Current classroom practices demonstrate that reading decodable text is a common approach to teaching phonics [1,6]. Many existing literacy platforms rely on standardized

Z. Li and F. Xiao—Equal contribution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 285–292, 2025.
https://doi.org/10.1007/978-3-031-98462-4_36

texts, which may not adequately represent the classrooms' linguistic and cultural diversity in recent years [10]. Children are more likely to engage in reading when they see their own cultures and identities represented in the content [4]. To support diverse learners, reading texts should be personalized to reflect different cultural backgrounds. Educators must navigate competing priorities, such as balancing individualized instruction with standardized learning objectives, and sourcing culturally responsive materials in digital environments [5,18].

To address these challenges, we designed StoryLab, a multimodal generative AI system supporting K-2 teachers and students in personalized story reading and literacy instruction. StoryLab utilizes SOTA generative models GPT-4o [9] and DALLE-3 [12] from OpenAI to produce adaptive learning materials, including customized texts, illustrations, and vocabulary support. Educators can define learning objectives such as phonics patterns, vocabulary, and grammar, while integrating student profiles (interests, cultural backgrounds) to generate engaging, coherent narratives. Unlike traditional platforms, StoryLab employs a teacher-in-the-loop approach, allowing educators to review and refine generated content, thus maintaining pedagogical rigor and data privacy. We evaluated the story's alignment with literacy objectives using a group TF-IDF metric based on an expanded benchmark corpus incorporating words from the DIBELS 8 [17]. Our generated stories outperformed reference corpora in TF-IDF scores, indicating stronger alignment with targeted vocabulary. A user study involving four teachers confirmed StoryLab's pedagogical effectiveness and usability, highlighting its potential as a scalable solution for personalized literacy instruction.

Overall, StoryLab is an AI-powered literacy platform that delivers personalized, story-based reading experiences for early learners and provides teachers with tools for targeted instruction, combining interactive features, customizable content, and a teacher-in-the-loop design to support culturally and linguistically diverse students.

2 Related Works

Recent advancements in AI techniques have driven the development of adaptive reading tools designed to personalize instruction for students. Commercial systems like iReady and Lexia Core5 use machine learning to adjust content difficulty in real-time, emphasizing foundational skills like phonics and fluency [8]. Similarly, DreamBox customizes activities to individual progress, while tools like Natural Reader enhance accessibility via text-to-speech. Amplify Reading effectively boosts literacy outcomes but provides limited teacher customization, focusing primarily on standardized metrics rather than cultural relevance or student interests. The balance between personalization and standardization remains problematic in tools like Amira Learning and Microsoft Reading Coach. Amira Learning offers real-time fluency and comprehension feedback but lacks sufficient teacher customization [5]. Likewise, Ello integrates student interests into phonics learning but restricts content adaptation through closed subscription

models. Consequently, current AI-powered systems often overlook teacher input, cultural responsiveness, and curricular integration [18]. These gaps include: (1) limited teacher-involved mechanisms for instructional alignment, (2) inadequate culturally responsive content with personal interest, and (3) reliance on proprietary models restricting adaptability in resource-limited classrooms. StoryLab addresses these challenges by integrating generative AI with educator-guided customization, enabling culturally responsive, dynamic story creation aligned with student interests and instructional goals.

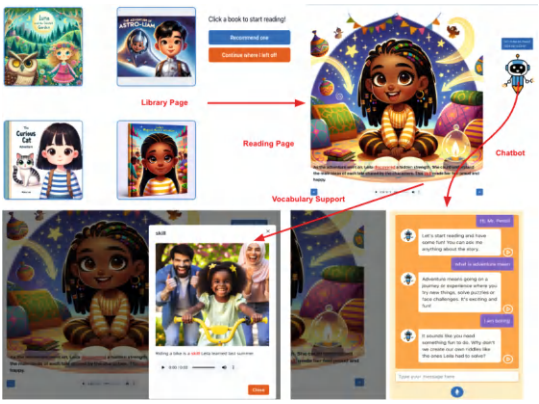


Fig. 1. Overview of the Student Interface in StoryLab.

3 The StoryLab Prototype

The system is designed with two primary user interfaces: Student View and Teacher Dashboard, both accessible through a user-friendly web application.

Student View. The student view (Fig. 1) offers an engaging, adaptive reading experience through a *Library Page* for browsing teacher-assigned or recommended books, and a *Reading Page* that combines narrations and illustrations to enhance comprehension. Literacy support features include clickable, highlighted words for pictorial definitions and pronunciations, TTS-based *Narration* supports for accessibility, and an interactive *Chatbot* that answers story-related questions via text or voice, using child-friendly prompts to ensure age-appropriate engagement.

Teacher Dashboard. As shown in Fig. 2, the teacher dashboard enables educators to monitor student progress and generate AI-assisted learning materials. The *Student Profile* section provides key demographic information, such as grade

level, ethnicity, and interests. Teachers can also access the *Student Library*, which displays each student's reading history and current books in progress. To further personalize instruction, the *Story Generation Canvas* allows educators to specify learning objectives, including vocabulary, phonics, grammar, and comprehension goals. Based on these inputs, the system produces customized stories aligned with individual student needs. Once a story is generated, teachers can use the *Story Review Page*, a pop-up editor that presents text and illustrations in a book-like format to review, edit, or regenerate specific pages. This feature ensures the content meets pedagogical goals while offering flexibility and control over the final output. By integrating AI-assisted storytelling with teacher supervision and student interactivity, StoryLab provides a dynamic and personalized learning experience. The system enhances literacy development by offering narratives, adaptive reading support, and interactive feedback mechanisms, ensuring a more effective and enjoyable reading journey for students.

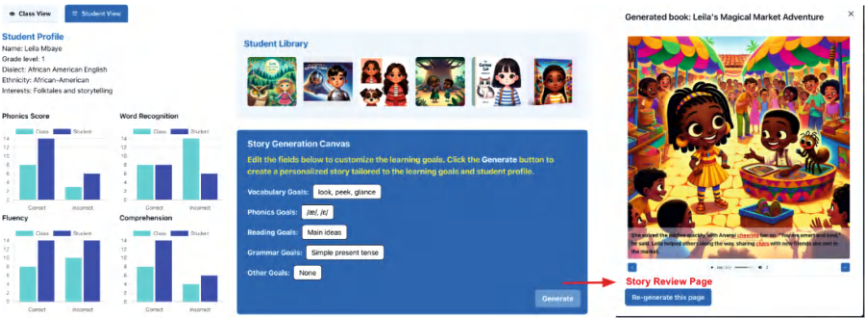


Fig. 2. Overview of the Teacher Dashboard in StoryLab.

Story Generation Canvas. The core of the teacher's dashboard is an editable story generation canvas (Fig. 2), where teachers create stories aligned with student profiles. Content consistency is ensured via detailed multimodal prompts describing character appearances, available in supplementary materials. The generation pipeline involves three steps: (1) outline drafting, (2) story filling, and (3) multimodal content generation. **Outline drafting**: This initial phase employs LLM planning capabilities to produce structured storylines, a method empirically validated for enhancing narrative quality and consistency [3,11,20]. The outline, generated by a story planner agent, includes the story title, plot points for logical progression, and target vocabulary tailored to learners' reading levels and objectives, thus boosting engagement [13]. **Story filling**: Following the outline, a writer agent develops the narrative by integrating learners' interests, vocabulary, and reading levels, ensuring alignment with educational objectives [19]. The writer agent employs: Dynamic adaptation that personalizes

content based on learners' profiles to maintain engagement; Pedagogical alignment that embeds learning objectives within narratives for natural skill development; Complexity gradation that adjusts text difficulty to learners' proficiency to facilitate manageable learning without cognitive overload. **Multimodal content generation**: Multimodal content is created using OpenAI APIs, specifically DALL-E 3 for images and TTS for audio, including: 1. Annotations for target vocabulary: Providing illustrations, sample sentences, and audio pronunciations for new vocabulary, aiding multimodal comprehension. 2. Illustrations for story content: Ensuring visual consistency through detailed character descriptions generated by the story planner agent. This maintains recognizable characters throughout the narrative (Fig. 3). These multimodal integrations enhance learner engagement and comprehension by addressing visual consistency challenges.



Fig. 3. Images (a, b) generated with detailed character prompts show consistent visuals, while images (c, d), lacking detailed prompts, show greater variation.

To ensure the reliability and pedagogical effectiveness of the AI-generated stories, we propose a straightforward teacher-in-the-loop framework. Once the system generates the story and corresponding illustrations, a story review page is presented, displaying the full set of illustrated pages alongside the story text and relevant dictionary entries. Teachers can then review, edit, or refine the story text to ensure that each page aligns with the intended learning objectives and teaching standards.

4 Evaluation

Quantitative Evaluation. We employed a novel group TF-IDF metric, extending traditional TF-IDF [14] by grouping synonyms and plural word forms to evaluate vocabulary alignment with learning objectives from the DIBELS 8 benchmark dataset [17]. For grades K–2, we selected vocabulary words—such as *job*, *team.*, and *time*, and generated ten stories per set, forming our sample set. We compared these with: *Generated Sample Set*: 10 generated stories per DIBELS 8 sample. *StoryLab Corpus*: 30 total generated stories aligned with DIBELS 8. *StoryDB*: English subset of a multilingual narrative dataset [16]. *Children Stories Text Corpus*: 98 public domain children's books from Project

Gutenberg [2]. As shown in Fig. 4, our stories consistently achieved higher scores across grades K–2 compared to other corpora, indicating strong alignment with target vocabulary. However, further evaluation needs to be conducted for fluency or pedagogical effectiveness.

Teacher Evaluation. To better evaluate the pedagogical value of our prototype, we conducted a user study with four educators focusing on: (1) teachers' perceptions of generated stories (language quality, age appropriateness, engagement, pedagogical value, story structure), and (2) potential design improvements. Participants' information, study procedure, and data analysis methods can be found in supplementary materials. Teachers responded positively to the StoryLab prototype, highlighting its intuitive dashboard and multimodal tools like read-aloud narration and multimedia dictionaries, which supported word recognition and comprehension. They also praised the story content for its linguistic quality, coherence, and demographic personalization, noting its potential for differentiated instruction (P1–P3). Despite StoryLab's potential, teachers identified key areas for improvement. P4 noted inconsistent character illustrations that disrupted narrative coherence, while others flagged mismatches between story complexity and intended reading levels, citing excessive sight words and advanced vocabulary. P3 recommended a "one sentence per picture" format for Pre-K, and both P3 and P4 emphasized that complex vocabulary without visual support may hinder recognition. Concerns about cultural representation also emerged: while teachers appreciated demographic personalization, P2 and P4 warned against superficial, ethnicity-based depictions and advocated for more authentic, diverse cultural narratives [4]. Additionally, although teachers valued the system's flexibility, P1–P3 found the interface cognitively demanding and called for intelligent scaffolding, such as system-suggested objectives or curated menus to ease use. These findings highlight the need to improve story alignment, cultural sensitivity, and interface usability in future iterations.

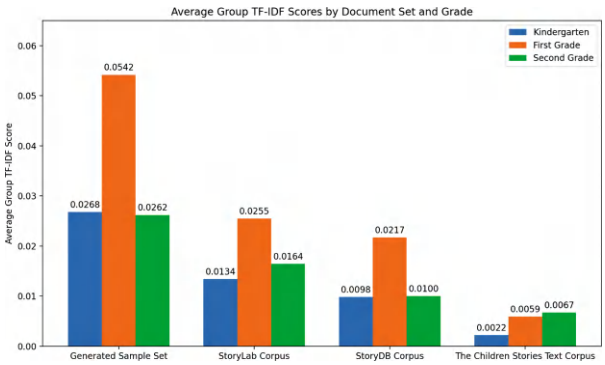


Fig. 4. Comparison of Average Group TF-IDF Scores Across Four Corpora.

5 Discussion and Conclusion

In conclusion, we introduced StoryLab, a multimodal platform supporting early childhood language and literacy development by aligning personalized student experiences with classroom learning objectives. StoryLab integrates student interests, cultural backgrounds, and assessment data, empowering teachers through customizable tools and a teacher-in-the-loop mechanism for reviewing AI-generated content. Our evaluations demonstrate that the generated stories effectively align with educational objectives and student interests, with strong adoption potential among educators. Despite promising outcomes, several limitations must be addressed in future work. Technically, high costs associated with OpenAI's API suggest exploring proprietary LLMs to lower operational expenses. Consistency in character portrayal also requires improvement, achievable through additional training data and model fine-tuning. Furthermore, our user study highlighted inconsistencies in aligning language complexity with grade levels, underscoring the need for detailed classroom benchmark data. Lastly, co-designing interface features with teachers and students, such as recommendation systems and progress tracking, could further enhance usability and instructional effectiveness.

Acknowledgments. This work is supported, in part, by the National Science Foundation under Grant 2229873 (AI4ExceptionalEd) and U.S. Department of Education under Grant R305C240046 (CELaRAI). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

1. Cheatham, J.P., Allor, J.H.: The influence of decodability in early reading text on reading achievement: a review of the evidence. Read. Writ. **25**, 2223–2246 (2012)
2. EdenBD: Children stories text corpus (2021). https://www.kaggle.com/datasets/edenbd/children-stories-text-corpus
3. He, T., et al.: Planning like human: a dual-process framework for dialogue planning. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, pp. 4768–4791. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.acl-long.262. https://aclanthology.org/2024.acl-long.262/
4. Koss, M.D.: Diversity in contemporary picturebooks: a content analysis. J. Child. Lit. **41**(1), 32–42 (2015)
5. Kumar, P.C., Chetty, M., Clegg, T.L., Vitak, J.: Privacy and security considerations for digital technology use in elementary schools. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2019)
6. Murphy Odo, D.: The use of decodable texts in the teaching of reading in children without reading disabilities: a meta-analysis. Literacy (2024)

7. Neuman, S.B., Celano, D.C.: Giving our Children A Fighting Chance: Poverty, Literacy, and the Development of Information Capital. Teachers College Press (2015)
8. Newton, S., Gamble, H., Su, Y., Zoski, J., Damico, D.: Examining the impact of amplify reading on student literacy in grades k-2. 2019 report. Online Submission (2019)
9. OpenAI: GPT-4 technical report (2023). https://cdn.openai.com/papers/gpt-4.pdf
10. Paris, D., Alim, H.S.: Culturally Sustaining Pedagogies: Teaching and Learning for Justice in a Changing World. Teachers College Press (2017)
11. Qiao, S., et al.: AutoAct: automatic agent learning from scratch for QA via self-planning. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, pp. 3003–3021. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.acl-long.165. https://aclanthology.org/2024.acl-long.165/
12. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
13. Renninger, K.A., Hidi, S.: The Power of Interest for Motivation and Engagement. Routledge (2015)
14. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. **28**(1), 11–21 (1972)
15. Tetzlaff, L., Schmiedek, F., Brod, G.: Developing personalized education: a dynamic framework. Educ. Psychol. Rev. **33**, 863–882 (2021)
16. Tikhonov, A., Samenko, I., Yamshchikov, I.P.: StoryDB: broad multi-language narrative dataset. arXiv preprint arXiv:2109.14396 (2021)
17. University of Oregon: DIBELS 8th Edition Administration and Scoring Guide. University of Oregon, Eugene, OR (2023). https://dibels.uoregon.edu/sites/default/files/2024-12/dibels8_admin_scoring_guide.pdf
18. Walkington, C., Bernacki, M.L.: Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions (2020)
19. Yelle, L.E.: The learning curve: historical review and comprehensive survey. Decis. Sci. **10**(2), 302–328 (1979)
20. Zhao, Y., Chen, L., Cohan, A., Zhao, C.: TaPERA: enhancing faithfulness and interpretability in long-form table QA by content planning and execution-based reasoning. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, pp. 12824–12840. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.acl-long.692. https://aclanthology.org/2024.acl-long.692/

How Do Learners Read the Content in a Multi-source Reading-to-Write Task? – A Multimodal Study

Debarshi Nath[1,2(✉)], Dragan Gašević[1], Yizhou Fan[3], and Ramkumar Rajendran[2]

[1] Centre for Learning Analytics at Monash, Monash University, Melbourne, Australia
debarshi.nath@monash.edu
[2] Centre for Educational Technology, Indian Institute of Technology Bombay, Mumbai, India
debarshi.nath@iitb.ac.in
[3] Graduate School of Education, Peking University, Beijing, China

Abstract. Text comprehension is an important skill that a student can develop, and reading behaviours like skimming and scanning can provide deep insights into how comprehension unfolds in realtime. These behaviours can be examined through deliberate eye movements, which offer a unique lens for investigating reading processes. Yet, most studies rely on low-level eye gaze metrics, such as fixations, as proxies for reading that do not distinguish reading from other tasks like video watching. More precise operationalisations of reading behaviours can provide deeper insights into learning when examined in real-world educational activities. In this study, we assessed the feasibility of a reading-skimming detection algorithm in a two-hour reading-to-write task. Our analysis identified three distinct reading behaviours—persistent reading, scanning, and skimming, demonstrating the potential of this algorithm for multimodal analysis. Our findings indicate that learners engage in thorough reading when encountering new content for the first time and tend to re-read only specific sections while drafting their essays. By integrating log data with reading behaviours, we identified several theoretical learning strategies and reconstructed a temporal narrative of reading-writing that aligns with contemporary models of reading comprehension.

Keywords: Reading comprehension · eyetracking · multimodal learning analytics · learning strategies

1 Introduction

Reading and comprehension play a central role in learning, which explains the significant attention it has received over the years [6,10]. Understanding how students engage with text under different task demands can provide valuable insights into their learning strategies. A promising approach involves eyetracking, which captures very granular information about a reader's gaze, providing a window into their reading comprehension and engagement [1].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 293–300, 2025.
https://doi.org/10.1007/978-3-031-98462-4_37

Eye movements associated with reading exhibit unique patterns that set them apart from other activities like video watching [5,19]. Variations exist within reading behaviours as well that give rise to distinct types of reading like *thorough reading*, *skimming*, or *scanning* [19]. Differences in these behaviours are reflected in their corresponding eye movement patterns. For example, thorough reading often involves more vertical saccades than eye movements during skimming [19]. While these differences may seem subtle, they can provide valuable theoretical insights into readers' cognitive and metacognitive processes [1]. However, most existing studies have focused on a few standard sets of lower-level eyegaze metrics based on fixations and saccades [10,19], which cannot capture the nuances of the reading behaviours and may risk misinterpretations [15]. Abstraction of reading behaviours like *skimming* and *persistent reading* can act as a bridge between low-level metrics and higher-level learning constructs. In the current study, we demonstrate how an algorithm capable of automatically detecting whether a learner is reading or skimming can explain the cognitive mechanics of learners in a 2-hour-long reading-writing task.

2 Background

2.1 The Mechanics of Comprehension

The mechanics of reading involve multiple cognitive and metacognitive processes working simultaneously. It requires readers to make deliberate eye movements, shift attention, and exert control as they construe the text and connect it to task demands [12,17]. Skilled comprehension goes beyond word recognition – it involves interpreting information across sentences and making inferences, while suppressing irrelevant or misleading details. Building on prior theories of reading comprehension, Knight and Horsley [12] proposed a framework to explain the mechanics of reading. It posits that at a macro level, an orienting attention network [17] helps readers orient to tasks through structural scannings, location scannings, and alignment scannings. At a micro level, the executive attention network [17] guides purposeful reading and text decoding. As reading progresses, readers coordinate task demands and distractions by scanning, re-scanning, reading, and rereading in search of alignment and confirmation.

These theoretical models offer a testbed for evaluating learning strategies in natural settings, where reading is guided by other goals such as writing an essay. Such endeavours require operationalising key constituent behavioural constructs of these models, such as *first-reading*, *re-reading*, *scanning*, and *skimming* [12, 16]. Ongoing evaluation of these models across varied contexts is also essential for their refinement and broader applicability [12,16].

2.2 Eye-Tracking Studies to Investigate Reading

Major reviews have highlighted fixations as a key low-level metric for characterising reading [1,10]. This focus stems from the eye-mind hypothesis [9], which suggests that the eyes fixate on what the mind is processing. Visual attention

leads to fixations, which are then used to infer higher-level processes [1,10]. However, such interpretations often rely on simple rule-based assumptions – if fixations are present, reading has occurred [10,19]. Such operationalisations may fall short in capturing deeper cognitive strategies like *perceived relevance of information* [3] or *satisficing* [5], which depend on abstracting reading behaviours like *skimming*. Further, mere presence of fixations on screen cannot distinguish reading from activities like video-watching [18], and can lead to validity threats [15]. Therefore, reading behaviours need to be better operationalised, rather than assuming that the presence of gaze on a screen equates to reading.

Interaction logs are widely used to study learning processes [6]. However, during periods of reading, learners typically make fewer physical interactions with the system compared to processes like writing, leaving gaps in the temporal traces [14]. These blind spots can limit our understanding of how people learn. Multimodal studies that supplement log data with eye-tracking offer a way to overcome this challenge [6]. However, most eye-tracking studies still rely primarily on fixation counts and durations, which cannot accurately operationalise reading. Existing algorithms that can detect reading and skimming behaviours have largely been developed for highly controlled environments, typically involving a single static screen [2–4]. These algorithms remain largely experimental, and their potential to advance theoretical models of comprehension will only be realised through application in real-world learning contexts [12].

One of the first attempts to automatically detect reading-skimming behaviours from eyegaze data is that of Campbell and Maglio [4] who proposed an algorithm that characterised eye movements and quantified them to build a pooled evidence-based method. However, this algorithm could only detect whether a person was reading or not. This limitation was overcome by Buscher et al. [3], who managed to distinguish skimming from reading. This class of algorithms is typically rule-based and characterises *mid-level features* [18] such as horizontal shifts of eye gaze, which are relevant for left-to-right language scripts. Another class of reading detection algorithms applies machine learning methods to relevant eyegaze features like saccade length and saccade amplitudes [2,18]. Real-world application of such experimental algorithms can help us find differences in how learners *read for the first time* or *re-read* – some of the foundational sub-constructs of theoretical models of reading comprehension (Sect. 2.1). Such explorations can also potentially unravel a host of information regarding higher-level learning behaviours like how learners strategically organise their readings and re-readings as per the requirements of the task [13].

Our primary research goal is to understand the mechanics of comprehension when learners are reading to write an essay, by distinguishing key behaviours like *reading* and *re-reading* (see Sect. 2.1):

RQ1: *Do the reading behaviours of the learners differ when they read content for the first time vs when they re-read it?*

Answers to RQ1, when analysed in conjunction with the temporal progression of reading and writing processes, could potentially unravel more insights:

RQ2: *What insights do the overall temporal progression of the reading and writing processes of the learners offer in the 2-h activity?*

3 Methods

3.1 Participants, Study Procedure and the Learning Environment

The dataset used for this study was collected from 135 undergraduate students. The participants were non-native English speakers from different disciplinary backgrounds. Ethics approval was obtained from the ethics committee before the commencement of the data collection, and consent was taken from all the students. After accounting for the loss of data in both the eyetracking and log data channels, we eventually performed our data analysis on 55 learners.

Participants were asked to write an essay in 200–400 words that provides a projected state of education in the year 2035. They were provided reading materials covering three topics: (1) artificial intelligence, (2) differentiation in education, and (3) scaffolding in education. The time allotted was 120 min. After the 120 min session, the students underwent a 15-min self-assessment session, following which they were given one hour to revise their essays. For our analysis, we consider only the first two-hour learning task, excluding the revision period.

The learning environment comprised several pages of reading content and was equipped with tools such as *planner*, the *timer*, *annotation tool*, *notes panel*, and *search tool*. The essay writing panel could be opened and closed at any time. We collected trace data of the learners from their timestamped logs, clickstream data and keystrokes during the activity. We also set up a Tobii Eyetracker sampled 120 Hz to capture the eyegaze data from the learners.

3.2 Algorithm to Identify Reading Behaviours During the Reading-Writing Activity and Further Data Processing

We employed an algorithm developed by Buscher et al. [3], which can detect whether a person is reading or skimming text from their eye gaze data. The algorithm uses *mid-level* features like rightward shifts in fixations and regressions [18], typical in scripts of languages going left to right, which contributes to its high explanatory potential. It has also demonstrated robustness across different text styles and participants [4], and has shown good performance [2,4].

Two separate detectors (reading detector and skimming detector) analyze the sequences of features. The scores for each type of eye movement feature are provided in Table 1. *Reading* or *skimming* is detected when the set threshold score is reached by one of the detectors (their respective thresholds being 30 and 20 [3]). If both detectors exceed their thresholds, the more plausible behaviour is determined by choosing the detector that accumulated the higher score.

Obtaining Three Reading Behaviours– *Persistent Reading*, *Scanning*, *Skimming*. The output *reading* sequences of the algorithm detailed in Sect. 3.2 were further categorized into *persistent reading* (reading thoroughly line by line)

Table 1. Types of transitions in fixations and their detector-specific scores

Distance and direction in letter spaces	Feature	Reading detector score (s_r)	Skimming detector score (s_s)
$0 < x \leq 11$	Read forward	10	5
$11 < x \leq 21$	Skim forward	5	10
$21 < x \leq 30$	Long skim jump	−5	8
$-6 \leq x < 0$	Short regression	−8	−8
$-16 \leq x < 6$	Long regression	−5	−3
$x < -16$	Reset jump	5 and line delimiter	5 and line delimiter
All other movements	Unrelated move	Line delimiter	Line delimiter

and *scanning* (reading quickly to find specific details within the text) [8]. This categorisation was based on whether the *reading* sequences detected by the algorithm occur in proximity. If more than 3 *read* lines were situated close to each other on the screen, the sequence was categorised into *persistent reading*, otherwise *scanning*. Three was chosen as the threshold because an average sentence on the screen in our learning environment ran for 3 lines. These definitions and categorisations were inspired by literature [12]. Thus, the sequences of fixations from our data are classified into *persistent reading*, *scanning*, *skimming* or none.

Aligning and Aggregating Eyegaze Behaviours With log Data. The following steps were performed on the timestamped fixation and saccade metrics exported using Tobii Pro Lab.

- **Measuring character sizes**: The reading-skimming algorithm depends on character sizes. We used PyTesseract [7], a Python wrapper for Google's Tesseract-OCR Engine, to measure the character sizes and spacings.
- **Obtaining reading behaviours combining log and eyegaze data**: The temporally-ordered fixations throughout the learning task were classified into *persistent reading*, *scanning* and *skimming* using the algorithm described previously, which were then aggregated over the logs of *FIRST_READING* and *RE_READING*, i.e., when reading occurs for the first time and repeated readings respectively. A full definition of the actions captured in our learning environment is provided as a supplementary document.
- **Statistical tests to analyse differences in reading behaviours**: The differences between reading behaviours during *FIRST_READING* and *RE_READING* were inspected using Welch's t-test. We check for differences in the number and proportions of *persistent readings* and *scannings*.

Visualising Temporal Progression of Cognitive Activities. To answer RQ2, we use a raincloud plot to visualise the temporal progression of the learners' reading and writing actions [14].

4 Results

4.1 RQ1: Contrasting First-Reading and Re-reading Behaviours

Statistical tests reveal that the number of *persistent_readings* was significantly higher during *FIRST_READINGS*(m = 1.608, sd = 4.066) than in *RE_READINGS* (m = 0.488, sd = 1.913) ($p < 0.0001$, t = 7.792, d = 0.352). Also, the proportion of *persistent_readings* during *FIRST_READINGS* (m = 0.093, sd = 0.208) was significantly higher than that in *RE_READING* (m = 0.038, sd = 0.142) ($p < 0.0001$, t = 6.507, d = 0.314). But, the proportion of scannings during *RE_READINGS* (m = 0.733, sd = 0.353) was found to be significantly higher than that during *FIRST_READINGS* (m = 0.691, sd = 0.326) (p = 0.019, t = 2.357, d = 0.124). The results indicate that the learners tended to read more persistently when they read the content for the first time, and scan lines selectively during re-readings.

4.2 RQ2: Mechanics of Reading and Writing Processes of the Learners

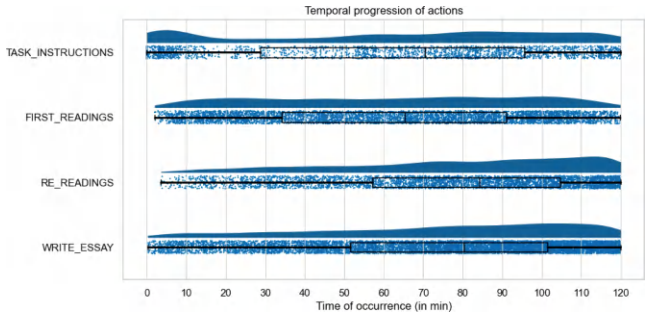


Fig. 1. Temporal progression of reading and writing actions of the learners

The temporal progression of the reading and writing actions is depicted in Fig. 1. The raincloud plots indicate that the learners started the task by going through the *TASK_INSTRUCTIONS*. They engaged in *re-reading* and *writing the essay* in parallel towards the later part of the task, when they also kept revisiting the task instructions from time to time. Combining these insights with our answers to RQ1, we conclude that the learners read more persistently during *FIRST_READING*, and they then strategically did local readings while writing the essay, revisiting the task objectives now and then.

5 Discussion

Our results from RQ1 highlight the differences in the reading behaviours of the learners, while the findings from RQ2 provide a deeper understanding of how

these reading behaviours are influenced by, and in turn lead to other cognitive processes. A prior study had indicated the behaviour of *selective* or *strategic re-reading* in learners when they were reading to write [14]. Our current study augments those findings by indicating that not only do learners re-read pages selectively, but also read lines of text selectively during re-reading. It had earlier been indicated that readers tend to engage in strategic re-readings when given the opportunity to read expository texts [13], and our study reiterates these findings. One form of strategic re-reading involves (re)organising the text representations mentally, where the reader might perform additional integration of explicit ideas, generate more inferences, and make long-distance connections [13], which our current learners also seem to perform.

A consolidation of the results of RQ1 and RQ2 demonstrates reading comprehension mechanics at work. The Construction-Integration Model [11] and Attention network theory [17] explain that readers start their reading by orienting towards the task requirements, similar to what we identified. At the executive attention network level [17], readers employ micro-level reading mechanics to execute control over their cognitive processes by keeping the task requirements in view. We too observe that our learners started by visiting the task requirements, and engaged in thorough and persistent reading when they encountered the content for the first time (Fig. 1). Towards the later part of the task, the learners performed strategic re-readings of the source material to confirm the alignment of their understanding and the requirement of the task while they wrote the essays in parallel (Fig. 1). Combining the results of RQ1 and RQ2, our findings indicate that learners engaged in strategic re-reading by scanning specific portions of the text they deemed relevant while connecting information across pages and generating inferences for essay writing [13].

The local re-reading mechanics can also be explained by *satisficing*, which assumes that readers set a threshold of satisfaction and linearly read through a unit of text as long as the rate of information gain remains above this threshold [5]. When a unit of text has been read, or satisfaction drops below the threshold of satisfaction, readers move to the next action. This strategy was effective in a reading-to-memorise task in literature, where the text was spread across multiple pages [5]. We see similar behaviour in our reading-to-write activity, indicating that satisficing may be a strategy that learners adopt in wider contexts.

Our study is also a response to the call by prior researchers who stressed the need for such eye-tracking studies in larger, real-world settings [12]. As comprehension varies across contexts [12,13], similar studies should be replicated to test, refine, and expand learning theories. However, text characteristics like complexity, structure, and learner characteristics like prior knowledge may impact how the learners engage with the content. We have not considered these aspects in our study, which is a limitation and should be investigated in the future.

References

1. Alemdag, E., Cagiltay, K.: A systematic review of eye tracking research on multimedia learning. Comput. Educ. **125**, 413–428 (2018)
2. Biedert, R., Hees, J., Dengel, A., Buscher, G.: A robust realtime reading-skimming classifier. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 123–130 (2012)
3. Buscher, G., Dengel, A., van Elst, L.: Eye movements as implicit relevance feedback. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2008, pp. 2991–2996. Association for Computing Machinery, New York (2008)
4. Campbell, C.S., Maglio, P.P.: A robust algorithm for reading detection. In: Proceedings of the Workshop on Perceptive User Interfaces, pp. 1–7 (2001)
5. Duggan, G.B., Payne, S.J.: Skim reading by satisficing: evidence from eye tracking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1141–1150 (2011)
6. Fan, Y., et al.: Improving the measurement of self-regulated learning using multichannel data. Metacogn. Learn. **17**(3), 1025–1055 (2022)
7. Hoffstaetter, S.: Pytesseract (2024). https://pypi.org/project/pytesseract/
8. IDP: How to skim and scan for keywords in your IELTS test (2018). https://ielts.idp.com/prepare/article-how-to-skim-scan-for-keywords
9. Just, M.A., Carpenter, P.A.: A theory of reading: from eye fixations to comprehension. Psychol. Rev. **87**(4), 329 (1980)
10. Ke, F., Liu, R., Sokolikj, Z., Dahlstrom-Hakki, I., Israel, M.: Using eye-tracking in education: review of empirical research and technology. In: Educational Technology Research and Development, pp. 1–36 (2024)
11. Kintsch, W.: Comprehension: a paradigm for cognition (1998)
12. Knight, B.A., Horsley, M.: A new approach to cognitive metrics: analysing the visual mechanics of comprehension using eye-tracking data in student completion of high-stakes testing evaluation. In: Current Trends in Eye Tracking Research, pp. 287–296. Springer (2013)
13. Millis, K.K., King, A.: Rereading strategically: the influences of comprehension ability and a prior reading on the memory for expository text. Read. Psychol. **22**(1), 41–65 (2001)
14. Nath, D., Gasevic, D., Fan, Y., Rajendran, R.: CTAM4SRL: a consolidated temporal analytic method for analysis of self-regulated learning. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, LAK 2024, pp. 645–655 (2024)
15. Orquin, J.L., Holmqvist, K.: Threats to the validity of eye-movement research in psychology. Behav. Res. Methods **50**, 1645–1656 (2018)
16. Petersen, S.E., Posner, M.I.: The attention system of the human brain: 20 years after. Annu. Rev. Neurosci. **35**(1), 73–89 (2012)
17. Posner, M.I., Petersen, S.E.: The attention system of the human brain. Annu. Rev. Neurosci. **13**(1), 25–42 (1990)
18. Srivastava, N., Newn, J., Velloso, E.: Combining low and mid-level gaze features for desktop activity recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(4), 1–27 (2018)
19. Strukelj, A., Niehorster, D.C.: One page of text: eye movements during regular and thorough reading, skimming, and spell checking. J. Eye Mov. Res. **11**(1) (2018)

Efficacy of a Computer Tutor that Models Expert Human Tutors

Andrew M. Olney[1(✉)], Sidney K. D'Mello[2], Natalie Person[3], Whitney Cade[4], Patrick Hays[1], Claire W. Dempsey[1], Blair Lehman[5], Betsy Williams[3], and Art Graesser[1]

[1] University of Memphis, Memphis, TN 38152, USA
{aolney,dphays,mcwllams,graesser}@memphis.edu
[2] University of Colorado Boulder, Boulder, CO 80309, USA
sidney.dmello@colorado.edu
[3] Rhodes College, Memphis, TN 38112, USA
{person,sandersb}@rhodes.edu
[4] American Institutes for Research, Arlington, VA 22202, USA
wcade@air.org
[5] Brighter Research, Thousand Oaks, CA 91362, USA
blehmann@brighter-research.com

Abstract. The contribution of expertise to tutoring effectiveness is unclear and continues to be debated. We conducted a 9-week learning efficacy study of an intelligent tutoring system (ITS) for biology modeled on expert human tutors with two control conditions: human tutors who were experts in the domain but not in tutoring and a no-tutoring condition. All conditions were supplemental to classroom instruction, and students took learning tests immediately before and after tutoring sessions as well as delayed tests 1–2 weeks later. Analysis using logistic mixed-effects modeling indicates significant positive effects on the immediate post-test for the ITS ($d = .71$) and human tutors ($d = .66$) which are in the 99th percentile of meta-analytic effects, as well as significant positive effects on the delayed post-test for the ITS ($d = .36$) and human tutors ($d = .39$). We discuss implications for the role of expertise in tutoring and the design of future studies.

Keywords: intelligent tutoring systems · expert tutor · dialogue · biology

1 Introduction

Meta-analyses of decades of research support the effectiveness of human tutoring for promoting learning [7,9,14,24,27]. The median effect size (ES) across these meta-analyses is .4 over conventional instruction, which is equivalent to an improvement from the 50th percentile to the 66th percentile. While these meta-analyses don't contrast the effectiveness of tutors by expertise, untrained tutors (.36 ES) are about as effective as trained tutors (.41 ES) [7], and peer tutors

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 301–308, 2025.
https://doi.org/10.1007/978-3-031-98462-4_38

(.52 ES) are about as effective as adults (.54 ES) [9]. Indeed across these meta-analyses, the median effect size when children are tutors (.4 ES) is comparable to the median effect size when adults are tutors (.38 ES). Thus these meta-analyses bring into question the intuition that training and age should increase tutoring effectiveness, leading some to argue that tutor experience is not important as long as the tutor is sufficiently knowledgeable and interactive [27].

Bloom's landmark 2-sigma paper [5] describes tutoring effects that are five times the median tutoring effect above (an improvement from the 50th percentile to the 98th percentile) and has been viewed as evidence supporting the importance of tutoring expertise and promoting "super tutors." Bloom's paper summarizes the work of two of his students [1,6] who contrasted tutoring, mastery learning (students could not move on until achieving criterion mastery), and conventional instruction. Bloom's 2 ES effect finding, however, obfuscates that his students' studies included a mastery learning criterion in the tutoring condition of 90% vs. 80% in the mastery learning condition [11,27]. Assuming no interaction between mastery learning and tutoring, we can subtract the mastery effect from the tutoring effect in these studies (1.91 ES – 1.06 ES = .85 ES) [1,6,10]. While .85 is a larger effect of tutoring than might be expected from meta-analyses generally, it is consistent with a meta-analysis .83 ES reported for human tutoring studies that create their own learning outcome measures when children are tutors [7]. Moreover, the Bloom studies used undergraduate students as tutors, with one study describing tutor training as lasting one week [6]. Altogether, a thorough inspection suggests that the 2-sigma tutors were not "super tutors" after all, yet the 2-sigma paper has been influential in shaping the development of ITSs moving forward, cf. [2].

Intelligent tutoring systems (ITS) compare favorably to human tutors in terms of effectiveness. A meta-analysis of ITS effectiveness found a median .66 ES compared to conventional instruction, which is equivalent to an improvement from the 50th percentile to the 75th percentile [11]. While a median .66 ES for ITS is larger than the median .4 ES for human tutoring described above, the meta-analysis also found that the ES depended heavily on the type of test used to measure learning. When the test was developed by researchers, the ITS effect size was .73 ES, but when a standardized test was used, the ITS effect size was .13 ES. Tests that combined researcher and standardized items had an intermediate .45 ES. Test type was found to be the single most important predictor of ITS effect size, such that when test type was held constant, no other study features influenced the size of the effect. In summary, human tutors may be more effective than ITS when test type is considered, so direct comparisons of ITS and human tutors are essential for understanding their relative effectiveness.

Our research builds on the intuition that expertise contributes to tutoring effectiveness beyond subject matter expertise. This intuition is informed by our observations and analyses of expert human tutors [21,22]. These expert tutors had 5+ years of tutoring experience, a teaching license, a degree in the tutored subject, and reputations for being effective tutors. In contrast to novice tutors,

these expert tutors were more interactive, diagnostic, and gave more discriminating feedback, all of which have been cited as theoretical reasons for the effectiveness of human tutors (for a review see [27]). Our goal was to build an ITS modeled on these expert human tutors, both to increase the effectiveness of the ITS and to provide further evidence on the role of expertise in tutoring. In this paper, we present the results of a 9-week study that compared the ITS to subject matter expert tutors and a classroom control. Our research questions are: (1) is learning in the ITS condition different from the classroom control and (2) is learning in the ITS condition different from subject matter expert tutors?

2 Guru: An ITS Modeled on Expert Human Tutors

We developed an intelligent tutoring system (ITS) for high school biology called Guru. Guru is a dialogue-based ITS in the AutoTutor ITS family [17]. Like AutoTutor, an animated tutor and the student interact using natural language dialogue as well as a multimedia workspace that displays and animates images relevant to the conversation. Student responses are interpreted using natural language understanding to provide formative feedback and tailor the dialogue to individual students' knowledge levels. In contrast to AutoTutor, Guru's pedagogical and motivational strategies are informed by in-depth observation and computational modeling of expert tutoring sessions [8,21,23]. The Guru animated pedagogical agent is also more sophisticated than the standard AutoTutor agent and uses motion capture data to produce realistic gestures and pointing. Guru was designed to cover Tennessee Biology I Curriculum Standards using curriculum maps provided by the state. In addition to covering topics in tutorial dialogue sessions, Guru presents interactive tasks such as generating summaries, completing concept maps [15,16], and cloze tasks [26]. The structure of a Guru tutoring session has been described in previous publications [8,19–21].

3 Method

3.1 Participants

Thirty-four tenth graders from an urban high school in the U.S. participated in the study. All students were enrolled in Biology I and had the same teacher for that course. Once a week (for nine weeks) students participated in the study during another class period. Of note, all students were required to pass the state-mandated end-of-course assessment for Biology I to graduate from high school.

3.2 Design and Knowledge Assessments

The study used a three-condition repeated-measures design in which each student interacted with both Guru (ITS) and a human tutor (Human) in addition to their regular classroom instruction (Class). The tutoring topics (for both

ITS and Human) always lagged behind what the biology teacher covered in the classroom by one week. In a particular week, students would receive either ITS or Human tutoring (e.g., ITS week 1, Human week 2, or vice versa). The nine-week study included three, 3-week cycles. Each cycle addressed four topics, where two were tutored (A, B) and two were not tutored (X, Y); this enabled comparison of tutoring vs. classroom instruction only. For the first two weeks of a cycle, the tutoring conditions completed immediate pre- and post-tests for both tutored and non-tutored topics (AX or BY). On the third week of a cycle, students took a delayed post-test covering all four topics (ABXY).

The knowledge assessments were multiple-choice tests where items were either obtained from previous state standardized tests across the U.S. or were created by a researcher who was blind to condition. The ratio of researcher-created to standardized items was approximately 2:1. Twelve item pre- and post-tests were administered at the beginning and end of each tutoring session for both the ITS and Human conditions to assess prior knowledge and immediate learning gains, respectively. Half of the items on each test were on a tutored topic and the other half on an untutored topic (Class condition). Test items and order of presentation were randomized across pre- and post-tests across students. On the third week of each cycle, students completed a 48-item delayed post-test. Half of the items on this test were previously seen by students (e.g., on the immediate pre- or post-test) and half the items were new but on the same topics.

Our analysis of the first 3 weeks of the study used repeated measures ANOVA [20]. However, the design and authentic setting of the study were a challenge for ANOVA, particularly since students increasingly missed sessions. Additionally, test items had varying difficulties, and test items were randomly assigned to tests for each participant. Logistic mixed models addressed these concerns by easily handling missing data and modeling the variability of test item correctness clustered by participant (cf. ability) and clustered by item (cf. difficulty). The contrast of interest for the effectiveness of ITS relative to Human and Class conditions is therefore the condition by test interaction using correctness of response on each test item as the dependent variable.

3.3 Procedure

Each tutorial session involved (a) students completing the pretest for 10 min (b) tutoring with the ITS or the human tutor for 35 min, and (c) the immediate post-test for 10 min. The delayed posttest occurred one week after all tutoring was complete. Interactions with each human tutor occurred in groups of two to four, which does not appear to reduce the effectiveness of typical tutors [13]. The tutors were undergraduate biology majors; one was a recent graduate. The tutors were provided with the topic to be tutored, the list of ITS concepts for the topic, and the class biology textbook. Students interacted with the ITS one-on-one using researcher-provided laptops.

Table 1. N per week

Cycle	Week 1	Week 2	Week 3
1	33	31	30
2	25	29	26
3	25	28	26

Note: $N = 34$

Table 2. Assessment descriptive statistics.

Condition	Class		Human		ITS	
	M	SD	M	SD	M	SD
Sessions	2.79	.59	2.68	.73	2.35	.85
Researcher items	.71	.09	.78	.14	.63	.18
Standardized items	.31	.04	.24	.14	.40	.16
Pre-test	.44	.50	.43	.50	.41	.49
Post-test	.49	.50	.63	.48	.64	.48
Delayed test	.40	.49	.56	.50	.55	.50

4 Results

Generally 75–90% of students attended sessions in a given week, as shown in Table 1. Table 2 shows assessment variable means, standard deviations, and item proportions. We fit a logistic mixed model and conducted statistical tests to answer our first two research questions. The model with fixed effects was $correctness \sim condition * test$, where condition was Class, Human, or ITS, and test was Pre-test, Post-test, or Delayed. Maximal random effects that resulted in model convergence [3] yielded an intercepts-only model: $correctness \sim condition *$ test $+ (1|participant) + (1|item)$. A Type III ANOVA analysis of the converging model [25] revealed a significant main effect of condition, $\chi^2(2) = 10.83, p = .004$, a significant main effect of test, $\chi^2(2) = 117.13, p < .001$, and a significant interaction between condition and test, $\chi^2(4) = 43.50, p < .001$. To answer our first two research questions, we conducted contrasts within the interaction using Tukey's p-value adjustment [12]. Contrasts of test scores within conditions were conducted to reveal learning effects. The ITS condition had significant differences between pre- and post-test, $z = 7.60, p < .001$, pre- and delayed test, $z = 3.90, p < .001$, and post- and delayed test, $z = -3.88, p < .001$, such that the ITS condition was more likely to correctly answer questions on the post- than the pre-test, $OR = 3.13, d = .71$, and on the delayed test than the pre-test, $OR = 1.77, d = .36$, but less likely to correctly answer questions on the delayed test than the post-test, $OR = .56, d = -.35$. The Human condition likewise had a significant difference between pre- and post-test, $z = 7.71, p < .001$, pre-test and delayed test, $z = 4.34, p < .001$, and post- and delayed test, $z = -3.12, p = .005$, such that the Human condition was more likely to correctly answer questions on the post- than the pre-test, $OR = 2.90, d = .66$, and on the delayed test than the pre-test, $OR = 1.86, d = .39$, but less likely to correctly answer questions on the delayed test than the post-test, $OR = .63, d = -.27$. The Class condition had no significant differences between pre-test and post-test, $z = 2.27, p = .06$, or pre- and delayed test, $z = -1.86, p = .150$. However, there was a significant difference between the delayed test and the post-test, $z = -3.57, p = .001$, such that the Class condition was less likely to correctly

answer questions on the delayed test than the post-test, $OR = .62, d = -.30$. Contrasts between conditions across tests similarly followed the pattern of $[ITS = Human] > Class$. There were no significant differences at pretest across conditions, and there were no significant differences at post- or delayed test between ITS and Human conditions. However, at post-test ITS was more likely to answer questions correctly than Class, $z = 4.24, p < .001, OR = 2.24, d = .50$, as was the Human condition, $z = 3.84, p < .001$, OR $= 2.04, d = .44$. Similarly, at delayed test, ITS was more likely to answer questions correctly than Class, $z = 4.24, p < .001, OR = 2.04, d = .45$, as was the Human condition, $z = 4.42, p < .001$, OR $= 2.11, d = .47$.

4.1 Discussion

Our research questions in this study were (1) is learning in the ITS condition different from the classroom control and (2) is learning in the ITS condition different from subject matter expert tutors? For the first question, the ITS condition improved learning on both the immediate post- and delayed-test, but the Class showed no significant improvement in learning. These results suggest that the ITS condition is more effective than the Class condition at promoting learning. For the second question, both ITS and Human conditions had approximately the same effects on learning. The ITS had a slightly larger effect on pre- to post-test ($d = .71$) compared to Human ($d = .66$), and Human had a slightly larger effect on pre- to delayed test ($d = .39$) compared to ITS ($d = .36$), but the ITS and Human conditions were not significantly different at post- or delayed tests. Additionally, the pre- to post-test effect for the ITS that used both standardized and researcher-created test items was in the 99th percentile [11], and the pre- to post-test effect for the human tutors was in the 99th percentile of meta-analytic effects [7,9,14,24,27] when the 2-sigma studies confounding mastery learning with tutoring were removed. These results suggest that the ITS and Human conditions were both highly and equally and effective at promoting learning.

Our study has several limitations. Since the Class condition was based on a single teacher, a better (or worse) teacher would have affected pre-test scores and corresponding comparisons within and between conditions. Additionally, we were unable to analyze test performance on standardized test items separately because these were both unevenly distributed across topics and randomly assigned across tests, so it is possible we would see a different pattern of results across only standardized items. Finally, the greatest limitation of the study is the lack of contrast between Human and ITS conditions. Our intuition for this research is that expertise in tutoring makes a contribution to tutoring effectiveness beyond subject matter expertise. It was not practical for us to include either a novice ITS matched for content or expert human tutors with the experience criteria described in Sect. 1 as comparison conditions, but these contrasts would have further informed the role of expertise in tutoring.

One of the challenges in expert tutoring research is the construction of assessments. The expert tutors we have studied base instruction on student needs

dynamically [21], but standard research practice is to control for content and use the same assessments across conditions. We argue that controlling for content across conditions effectively forces the use of a curriculum script and would create a handicap for expert human tutor effectiveness. Recent advances in generative AI for assessment suggest that large language models can dynamically produce multiple choice questions and that these questions have comparable quality and psychometric properties to human-authored questions on the same topic [4,18]. If effective learning assessments can be dynamically generated by AI, then future studies could study both human tutors who were experts in the domain but not in tutoring and expert human tutors in their natural contexts, across topics. Such research would better inform our understanding of the role of expertise in tutoring: not just what expert human tutors do, but how those differences translate into effectiveness.

Acknowledgments. This research was supported by the National Science Foundation (NSF) (HCC 0834847, DRL 1108845, DUE 1918751) and Institute of Education Sciences (IES), U.S. Department of Education (DoE), through Grant R305A080594. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF, IES, or DoE.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anania, J.: The effects of quality of instruction on the cognitive and affective learning of students. Ph.D. thesis (1981)
2. Anderson, J.R., Boyle, C.F., Reiser, B.J.: Intelligent tutoring systems. Science **228**(4698), 456–462 (1985). https://doi.org/10.1126/science.228.4698.456
3. Barr, D.J., Levy, R., Scheepers, C., Tily, H.J.: Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. **68**(3), 255–278 (2013). https://doi.org/10.1016/j.jml.2012.11.001
4. Bhandari, S., Liu, Y., Kwak, Y., Pardos, Z.A.: Evaluating the psychometric properties of ChatGPT-generated questions. Comput. Educ. Artif. Intell. **7**, 100284 (2024). https://doi.org/10.1016/j.caeai.2024.100284
5. Bloom, B.S.: The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. Educ. Res. **13**(6), 4–16 (1984)
6. Burke, A.J.: Students' potential for learning contrasted under tutorial and group approaches to instruction. Ph.D. thesis (1983)
7. Cohen, P.A., Kulik, J.A., Kulik, C.: Educational outcomes of tutoring: a meta analysis of findings. Am. Educ. Res. J. **19**, 237–248 (1982)
8. D'Mello, S.K., Olney, A.M., Person, N.: Mining collaborative patterns in tutorial dialogues. J. Educ. Data Mining **2**(1), 1–37 (2010)
9. Hartley, S.S.: Meta-analysis of the effects of individually paced instruction in mathematics. Ph.D. thesis (1977)
10. Kulik, C., Kulik, J.A., Bangert-Drowns, R.L.: Effectiveness of mastery learning programs: a meta-analysis. Rev. Educ. Res. **60**(2), 265–299 (1990). https://doi.org/10.3102/00346543060002265

11. Kulik, J.A., Fletcher, J.D.: Effectiveness of intelligent tutoring systems. Rev. Educ. Res. **86**(1), 42–78 (2016). https://doi.org/10.3102/0034654315581420
12. Lenth, R.V.: emmeans: Estimated Marginal Means, aka Least-Squares Means (2025). https://rvlenth.github.io/emmeans/
13. Ma, W., Adesope, O.O., Nesbit, J.C., Liu, Q.: Intelligent tutoring systems and learning outcomes: a meta-analysis. J. Educ. Psychol. **106**(4), 901–918 (2014). https://doi.org/10.1037/a0037123
14. Mathes, P.G., Fuchs, L.S.: The efficacy of peer tutoring in reading for students with mild disabilities: a best-evidence synthesis. Sch. Psychol. Rev. **23**(1), 59–80 (1994). https://doi.org/10.1080/02796015.1994.12085695
15. Novak, J.D.: Concept mapping: a useful tool for science education. J. Res. Sci. Teach. **27**(10), 937–49 (1990)
16. Novak, J.D., Canas, A.J.: The theory underlying concept maps and how to construct them. Technical report, Institute for Human and Machine Cognition (2006)
17. Nye, B.D., Graesser, A.C., Hu, X.: AutoTutor and family: a review of 17 years of natural language tutoring. Int. J. Artif. Intell. Educ. **24**(4), 427–469 (2014). https://doi.org/10.1007/s40593-014-0029-5
18. Olney, A.M.: Generating multiple choice questions from a textbook: LLMs match human performance on most metrics. In: Moore, S., et al. (eds.) Proceedings of Empowering Education with LLMs - the Next-Gen Interface and Content Generation. CEUR-WS.org (2023). https://ceur-ws.org/Vol-3487/paper7.pdf
19. Olney, A.M., et al.: Efficacy of a computer tutor that models expert human tutors (2025). https://arxiv.org/abs/2504.16132
20. Olney, A.M., et al.: Guru: a computer tutor that models expert human tutors. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 256–261. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30950-2_32
21. Olney, A.M., Graesser, A.C., Person, N.K.: Tutorial dialog in natural language. In: Nkambou, R., Bourdeau, J., Mizoguchi, R. (eds.) Advances in Intelligent Tutoring Systems, Studies in Computational Intelligence, vol. 308, pp. 181–206. Springer-Verlag, Berlin (2010)
22. Olney, A.M., Person, N.K., Graesser, A.C.: Guru: designing a conversational expert intelligent tutoring system. In: McCarthy, P., Boonthum-Denecke, C., Lamkin, T. (eds.) Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches, pp. 156–171. IGI Global, Hershey, PA (2012)
23. Person, N.K., Lehman, B., Ozbun, R.: Pedagogical and motivational dialogue moves used by expert tutors: presented at the 17th Annual Meeting of the Society for Text and Discourse, Glasgow, Scotland (2007)
24. Ritter, G.W., Barnett, J.H., Denny, G.S., Albin, G.R.: The effectiveness of volunteer tutoring programs for elementary and middle school students: a meta-analysis. Rev. Educ. Res. **79**(1), 3–38 (2009). https://doi.org/10.3102/0034654308325690
25. Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M.S.: afex: analysis of Factorial Experiments (2023). https://cran.r-project.org/package=afex, r package version 1.3-0
26. Taylor, W.L.: "Cloze procedure": a new tool for measuring readability. J. Q. **30**(4), 415–433 (1953)
27. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. **46**(4), 197–221 (2011)



An Agentic Framework for Real-Time Pedagogical Plot Generation

S. Thomas Christie[1,2], Anna N. Rafferty[2], Zack Lee[3], Ella Cutler[1], Yu Tian[1], and Husni Almoubayyed[3(✉)]

[1] Digital Harbor Foundation, Baltimore, MD, USA
[2] Carleton College, Northfield, MN, USA
[3] Carnegie Learning, Pittsburgh, PA, USA
halmoubayyed@carnegielearning.com

Abstract. Plotting functions on the Cartesian plane is a critical pedagogical aid for conveying concepts in K-12 mathematics education. We report on the integration of LiveHint AI with the graphing calculator Desmos to generate plots in real-time in response to student messages using LLM function calling. We characterize the system's behavior, discussing its reliability and readiness for use with students. We also describe recurring issues and discuss possible architectural improvements to mitigate them. Finally, we describe opportunities for future improvements to address issues such as defining appropriate plot bounds and optimizing the frequency at which the tool generates plots.

Keywords: Conversational agents · LLM agents · LLM function calling · Math education · Multi-modal · Automated graph generation

1 Introduction

One-on-one tutoring is a time-tested and effective method for assisting student learning in mathematics. With the recent availability of Large Language Models (LLMs), technology providers have begun producing electronic simulacra of human tutors in which students can seek help from artificial agents via a chat interface. AI-powered tutors are scalable, inexpensive, and available on-demand in ways that traditional human tutors are not.

Whereas human tutors often illustrate concepts from algebra or geometry on paper, tutoring via LLM-driven text-based chats restricts explanations to be text-only. In middle school mathematics, the relationship between algebraic equations and their representation on Cartesian coordinates is a core part of the curriculum and difficult to meaningfully discuss without visual aids. What Works Clearinghouse reports that graphical representations and the use of number lines are two important strategies in assisting students struggling with mathematics [4–6,17] based on dozens of experiments.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-98462-4_39.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 309–316, 2025.
https://doi.org/10.1007/978-3-031-98462-4_39

A few organizations such as Wolfram|Alpha [9] and more recently Merlyn Mind [1] have incorporated plots alongside text-based solutions. However, the inclusion of graphs or diagrams into interactive chats remains relatively rare.

Carnegie Learning is developing an LLM-powered AI tutor called LiveHint AI as part of its suite of educational tools. LiveHint AI provides step-by-step guidance to students working on math problems. Following prior research, evidence-based recommendations, and school feedback, we seek to investigate whether real-time contextually relevant, pedagogically helpful, and functionally reliable plots could be incorporated into LiveHint AI. We describe a framework for integrating this functionality into LiveHint AI, and conducting both automated and human evaluations for nearly 200 problems from Carnegie Learning's Algebra I curriculum. We share design and implementation details of generative plot integration to LiveHint AI, along with results of thorough automated and human evaluations of the plotting system. As evaluating LLM-driven technical systems is still a novel process, especially in the educational domain, we hope our process serves as a motivating example for educational technology developers.

2 Methods

While working on a math problem within Carnegie Learning products such as the digital book or MATHstream, students can click the LiveHint AI icon for on-demand, step-by-step, chat-based help. Plotting functionality is made available to the model via 'function calling' [10], also called 'tool calling' or 'structured outputs'. During relevant LLM calls, a set of function definitions are included during the invocation that describes the plotting capabilities available to the model. The model 'generates a plot' by returning one or more JSON-formatted function calls, which are parsed and converted to validated Desmos expressions. The expression is forwarded to the browser and rendered using Desmos.

Desmos [16] is a flexible plotting tool available at http://www.desmos.com. Desmos plots can be embedded in a webpage using the Desmos JavaScript API, which is the mechanism we used to include plots in LiveHint AI. Desmos plots are interactive and allow students to edit generated plots. Plot expressions can be parameterized, for example by mapping slope and intercept parameters of a line `y=mx+b` to sliders for `m` and `b` that students can manipulate. 'Key points' can also be added to the plot by extending the expression list, enabling the system to explicitly highlight e.g., intercepts or extreme values. Finally, Desmos allows multiple functions to be graphed in a single plot window. Our implementation supports the automated use of each of these features.

Function definitions are treated as context tokens by LLM providers and application developers are charged accordingly. To reduce LLM invocation costs, we implemented a multi-step plotting architecture using LangGraph [7], a tool designed for agentic workflows. The graph topology we used is detailed in Fig. 1(b). Most completion requests are handled by a conversation node, which has access to only a single placeholder function called `generate_plot`. This function includes a single argument called `plot_description`, a natural language description of the plot the LLM would like to produce. When the node 'decides'

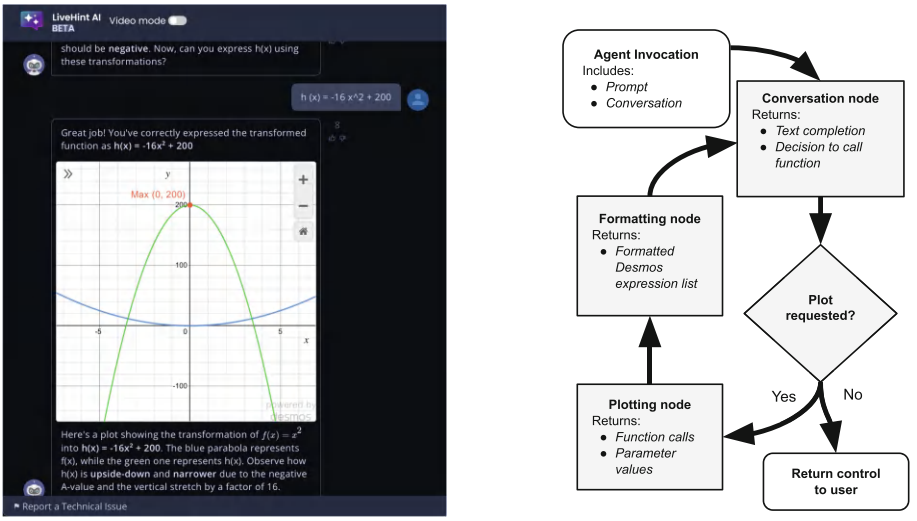


Fig. 1. (a) Screenshot of LiveHint AI automatically generating an appropriate plot. (b) LangGraph agent control flow.

a plot is pedagogically useful, it calls `generate_plot` and control is passed to a plotting node, which makes an additional LLM call with descriptions of all available functions, and then control is returned to the conversation node. The sequence experienced by the student is as follows:

1. `Student` Plot request
2. `LiveHint AI` Text content <may be omitted>
3. `LiveHint AI` One or more Desmos plots
4. `LiveHint AI` Text content describing what was just plotted
5. `Student` Next turn

Figure 1(a) shows an example of this sequence. This multi-step process significantly reduces costs associated with function calling by allowing the majority of LLM calls to be made without a long list of function definitions in the context.

3 Characterization and Evaluation

Standards for characterizing and evaluating LLM performance are rapidly developing [14,15], as evidenced by the recent proliferation of evaluation tools and benchmark datasets [2,8,12,13]. In evaluating LiveHint AI with plotting, we created a list of design specifications and a corresponding list of technical metrics to identify the degree to which these specifications were met. Desired outcomes included being responsive to plot requests; generating plots without syntactic or semantic errors; creating plots that were pedagogically appropriate; and consistency between the description in step one of plotting and the generated plot in

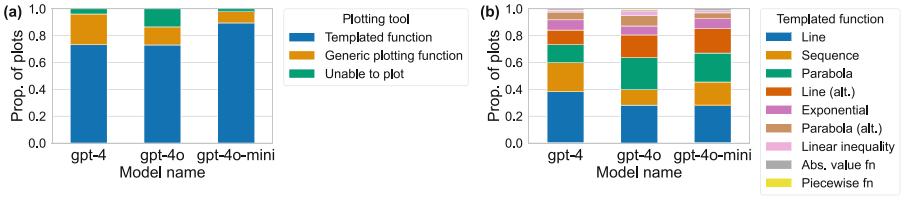


Fig. 2. (a) Most plotting was executed via templated functions. (b) Relative usage of each templated function.

step two of plotting. We carried out these tests with three OpenAI models: GPT-4 (230613), GPT-4o(240806), and GPT-4o-mini (240718). We discuss specifics and insights from both human and automated evaluations below.

Automated Evaluation via FlexEval. We simulated 600 conversations between LiveHint AI and a simulated student (powered by GPT-4o with injections of plot requests at the end of some student turns). Each conversation included six conversational turns for the simulated student and the agent. In each conversation, the student asked for help with a real Algebra 1 problem from the Carnegie Learning curriculum, the details of which were represented to LiveHint AI in the system prompt. Results below include evaluations of the system using OpenAI models GPT-4, GPT-4o, and GPT-4o-mini (200 conversations per model). We used FlexEval [3] to extract information from each thread and deploy rubric-based evaluations, in which both Python functions and a grader LLM quantified properties of the conversations.

On average, conversations consist of approximately 20,000 total prompt and completion tokens. Using two-step plotting provided an estimated 40% cost savings: since the prompt for the second step of two-step plotting, with the function definitions, contains an average of > 2.5 times more context tokens than the average non-plotting prompt.

The system can either call templated functions (e.g., by providing `m` and `b` for the function `y=mx+b`), or provide a full LaTeX-compatible string. As shown in Fig. 2(a), all three models used the templated functions for the vast majority of their function calls, demonstrating good coverage of the types of plotting needs encountered in the problems. In a limited number of situations, the system was unable to create a plot specified in step 1 (see supplementary materials for details)[1]. Cases where the system provided LaTeX directly had mixed results: while some uses were successful, in other cases the function was used to try to mimic non-Cartesian plot types like bar plots, leading to a visual representation that had some correct features but was also missing important information.

Almost all plotting attempts were deemed "pedagogically appropriate" (Fig. 3(a)), and in most cases, the plot generated by step 2 matched or somewhat matched the description from step 1 (Fig. 3(b)). We did not find evidence that

[1] https://osf.io/2hx63/?view_only=977b11f8c80141bcb64f51c30269c6b0.

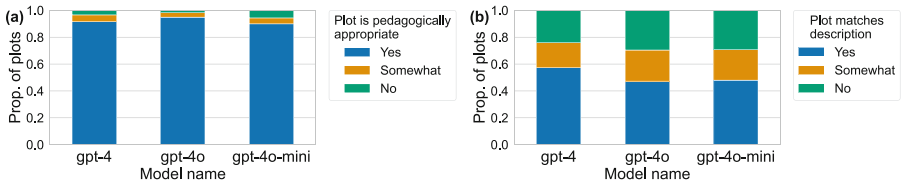


Fig. 3. (a) Almost all plots were deemed pedagogically appropriate. (b) The majority of plots generated by step 2 matched or somewhat matched the description provided in step 1 of the plot generation process.

these metrics varied across models (appropriateness: $\chi^2(4, N = 786) = 6.45$, $p = .17$; matching: $\chi^2(4, N = 771) = 6.86$, $p = .14$). Examples of mismatch included missing annotations; inappropriate axis ranges; and most seriously, mathematical errors in some cases (see supplement for further details).

Human Evaluation. While automated evaluations allow us to do more systematic and large-scale evaluations, some issues are more reliably identified by human reviewers. We conducted a human evaluation (with some process variations compared to the automated evaluations) with internal staff from Carnegie Learning. In total, 166 sessions were reviewed by 12 members of staff with GPT-4 and GPT-4o sessions. Each session consisted of 14 turns between LiveHint AI and a GPT-4o instance acting as a student.

Evaluators were provided with transcripts viewed within the LiveHint AI interface and asked to rate the graphing feature by choosing a 'good', 'bad', or 'mixed' rating. 'Mixed' ratings were indicative of minor issues that fell into some 'technically correct but potentially awkward' category. An example of a mixed rating was instances where LiveHint AI would create a correct plot to explain a concept, but not use numbers specific to the current problem. Evaluators were also asked to flag one or more of the following issues in each session: There was a technical error (hereafter, 'technical error'), text description by tutor did not match plot ('unmatching desc.'), student **reasonably** asked for a plot but was not given ('plot not generated'), plot was generated by request but it was unhelpful ('unhelpful'), plot was generated but it was not correct ('incorrect').

As shown in Fig. 4 (left), performance was similar across models, and we found no significant differences in terms of ratings overall ($\chi^2(2, N = 166) = 0.89$, $p = .641$). For the most common types of function calls, GPT-4o seems to have a small edge for Generic function plots ($\chi^2(2, N = 66) = 6.87, p = .032$) and, directionally but not significantly, Line plots ($\chi^2(2, N = 45) = 4.50, p = .105$). Figure 4 (right) shows how often each issue was flagged per function type. While sample sizes were small, differing error rates suggest targeted improvements. Only Generic function plots had 'unhelpful' and 'plot not generated' sessions, while Line plots had a larger fraction of 'incorrect' plots.

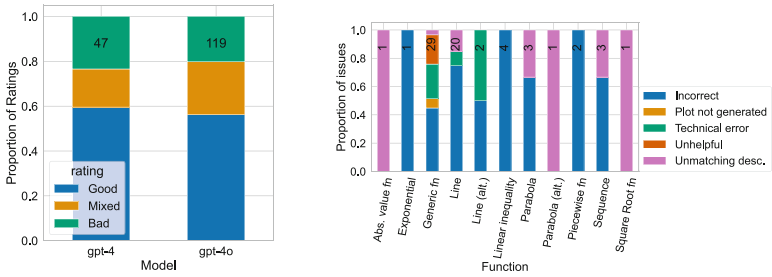


Fig. 4. The ratings from the human evaluations broken down by model (left), and proportion of issues by function type (right).

4 Discussion and Future Work

In early development, we attempted to generate plots by asking LLMs to generate code for the plotting tools JSXGraph and Penrose, with very high failure rates. Fine-tuning using examples from the model documentation increased failure rates. Including the full API specification in the chat context was helpful but expensive, and this investigation led to the idea to use function templates.

The agentic two-step plotting process allows us to record the model's intentions in natural language and provides that description of a plot as context during the subsequent plot generation step. However, providing only a generic description of plotting capabilities to the conversation node does not sufficiently constrain the types of attempted plots: as the results show, some plots desired in Step 1 were not supported by Step 2. Explicitly enumerating supported plotting types in the system prompt for Step 1 may mitigate this issue. Since beginning this project, LLMs have begun supporting structured outputs [11] with 'strict' adherence to specified schemas; this is a promising future direction.

Self-consistency between the chat context, the plot description, and the resulting plot is critical for correct and relevant mathematical explanations. Recently released reasoning models like OpenAI's o3-mini may improve the reliability of the second plotting step. An important benefit of the agentic architecture is that separate models can be used for each node in the agent graph. A persistent issue that surfaced during human evaluations was the appropriateness of the bounds for the plot window (which Desmos does not set automatically). Plots sometimes show mismatches between axis scales and described text, especially when the horizontal axis involves years. We are creating success metrics to quantify these observations and others to guide future iterations.

Once the system is sufficiently reliable, we plan to perform a randomized controlled trial to evaluate the impact of plot accessibility, frequency, and interactivity on student learning, engagement, and persistence.

Acknowledgments. This work was supported by the Learning Engineering Virtual Institute (LEVI) (http://learning-engineering-virtual-institute.org) and fiscally sponsored by the Digital Harbor Foundation.

References

1. Bulusu, A., Man, B., Jagmohan, A., Vempaty, A., Mari-Wyka, J.: An automated graphing system for mathematical pedagogy. In: NeurIPS'23 Workshop on Generative AI for Education (GAIED), New Orleans, LA, USA (2023)
2. Chen, Z., et al.: T-eval: evaluating the tool utilization capability of large language models step by step. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 9510–9529 (2024)
3. Christie, S.T., Moreau-Pernet, B., Tian, Y., Whitmer, J.: Flexeval: a customizable tool for chatbot performance evaluation and dialogue analysis. In: Paaen, B., Epp, C.D. (eds.) Proceedings of the 17th International Conference on Educational Data Mining, pp. 903–908. International Educational Data Mining Society, Atlanta (2024). https://doi.org/10.5281/zenodo.12729993
4. Fuchs, L.S., Compton, D.L., Fuchs, D., Paulsen, K., Bryant, J.D., Hamlett, C.L.: The prevention, identification, and cognitive determinants of math difficulty. J. Educ. Psychol. **97**(3), 493–513 (2005). https://doi.org/10.1037/0022-0663.97.3.493
5. Fuchs, L.S., Fuchs, D.: Principles for the prevention and intervention of mathematics difficulties. Learn. Disabil. Res. Pract. **16**(2), 85–95 (2001). https://doi.org/10.1111/0938-8982.00010
6. Jitendra, A.K., Nelson, G., Pulles, S.M., Kiss, A.J., Houseworth, J.: Is mathematical representation of problems an evidence-based strategy for students with mathematics difficulties? Except. Child. **83**(1), 8–25 (2016). https://doi.org/10.1111/0938-8982.00010
7. LangGraph (2025). https://langchain-ai.github.io/langgraph/. Accessed 18 Feb 2025
8. LangSmith: LangSmith (2025). https://docs.smith.langchain.com. Accessed 18 Feb 2025
9. LLC, W.A.: Wolfram Alpha. https://www.wolframalpha.com/. Accessed 18 Feb 2025
10. OpenAI: Function Calling in GPT-4 and GPT-3.5 (2023). https://platform.openai.com/docs/guides/function-calling. Accessed 19 Feb 2025
11. Pokrass, M.: Introducing structured outputs in the api. https://openai.com/index/introducing-structured-outputs-in-the-api/, blog post
12. PromptFoo: PromptFoo (2025). https://www.promptfoo.dev. Accessed 18 Feb 2025
13. Saini, H., Laskar, M.T.R., Chen, C., Mohammadi, E., Rossouw, D.: LLM Evaluate: an industry-focused evaluation tool for large language models. In: Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pp. 286–294 (2025)
14. van Schaik, T.A., Pugh, B.: A field guide to automatic evaluation of llm-generated summaries. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2832–2836 (2024)
15. Shankar, S., Zamfirescu-Pereira, J., Hartmann, B., Parameswaran, A., Arawjo, I.: Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, pp. 1–14 (2024)

16. Studio, P.D.: Desmos graphing calculator. https://www.desmos.com/calculator. Accessed 18 Feb 2025
17. What Works Clearinghouse: Assisting students struggling with mathematics: Intervention in the elementary grades: Educator's practice guide. Practice Guide WWC 2021006, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (2021)

Predicting Self-regulated Learning Support Needs During Learning

Rick Dijkstra[1(✉)], Max Hinne[2], Eliane Segers[1,3], Josh Ring[1], and Inge Molenaar[1,3]

[1] Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands
{rick.dijkstra,eliane.segers,inge.molenaar}@ru.nl,
josh.ring@donders.ru.nl
[2] Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands
max.hinne@donders.ru.nl
[3] Dutch National Education Lab AI, Radboud University, Nijmegen, The Netherlands
https://www.ru.nl/en/nolai

Abstract. Adaptive Learning Technologies generate data traces as children interact with them, offering a unique opportunity to estimate self-regulated learning (SRL) support needs. Providing timely, data-driven support for these needs may enhance children's ability to self-regulate their learning and improve learning outcomes. While previous work has identified different levels of SRL support needs using Bayesian non-parametric clustering, these classifications were determined after completing a learning session, delaying potential support. In this study, we present a novel method for identifying children's SRL support needs *during* a learning session, utilizing a Dirichlet-process Gaussian-process mixture model (DPGP). The model clustered based on children's response count and response time, resulting in a Matthews' correlation coefficient of 0.75 after children solve 33 problems out of an average of 71 - less than half the session. Our findings demonstrate that real-time identification of SRL support needs is feasible and effective. This work opens new possibilities for enhancing personalized, online learning experiences by enabling timely, data-driven support tailored to each child's needs.

Keywords: Clustering · Gaussian Processes · Self-Regulated Learning · Prediction, Personalized support

1 Introduction

The daily use of adaptive learning technologies (ALT) in primary schools [25] creates opportunities to use artificial intelligence (AI) to optimize children's learning with personalized support [2,17]. However, the current focus is on improving children's cognitive abilities [10], whereas children also need self-regulated learning (SRL) support [26]. Recent studies have shown that data collected by ALTs can

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 317–324, 2025.
https://doi.org/10.1007/978-3-031-98462-4_40

be leveraged to both accurately measure children's SRL [22,29] and detect their SRL support needs [5] after learning has been concluded. Children have been shown to benefit from personalized SRL support during learning [6,19]. Therefore, this paper will focus on predicting the SRL support needs during learning, to allow for real-time, data-driven SRL support. According to SRL theory, learning is a goal-directed process that involves conscious decision-making to achieve the desired outcomes [1]. Cognitive and metacognitive actions are together indicative of self-regulated learning, which has been associated with improved learning outcomes [27]. Still, children do not apply SRL naturally and need training and personalized support to improve their SRL [13,18].

Personalized SRL support relies on accurate measurements of SRL, which can be inferred from trace data from ALTs [5] collected during learning [20]. Another indicator of a child's SRL support needs is the time they spend on solving a problem (response time). However, limited research has explored the relationship between response time and SRL processes [12,24].

In this study, we investigated whether it was possible to predict SRL support needs from children in real time, while using trace data from ALTs that are focussed on estimating cognitive abilities. We constructed ability curves from children working on arithmetic problems based on the number of problems solved (response count) and the time spent solving the problems (response time). We analysed how these ability curves can be clustered and predicted by extending the model from [5]. We evaluated their predictive performance, and finally estimated the SRL support needs of these clusters. A longer version of this paper, with a more thorough comparison with different model settings and with a formal description of the model can be found on Github[1].

2 Methods

2.1 Data

The data have been collected in an experiment in a classroom setting. In total, 194 children between the ages of 10–13 in grade 5 of elementary school were asked to work with the ALT, which they work with daily [7]. After instruction, they practised three learning goals of increasing difficulty – one day per learning goal. On the fourth day they could select which of the goals to further practice on the fourth day. A pre-test was administered at the start of the first day, and a post-test was administered at the end of the fourth day. A subset of these data has previously been used by [5] to obtain a post-hoc clustering of learners.

2.2 Clustering Model

The foundation of our SRL support-needs prediction algorithm is a probabilistic clustering model known as a Dirichlet process-Gaussian process (DPGP; [5,16]) model. The DPGP model identifies clusters of sequential data, i.e., the ability

[1] https://github.com/Dijkstra7/AIED2025_Appendix.

curves collected via the ALT, without pre-determining the number of clusters or their properties (e.g., slope, variance).

The DPGP model is a mixture model, meaning that each cluster is a distribution represented by a combination of mean and covariance functions [21]. The mean function μ_c for each cluster c describes the expected ability level progression. It incorporates the assumption that children might not initially know the topic, but once they do, they will grow towards a certain ability level with a certain speed. The kernel function K_c for each cluster incorporates the assumption that inputs that are close together will have more similar outputs. Both functions depend on parameters, which are collectively represented as θ_c. The probability of observing an ability curve a, with input x_a and ability scores y_a given a cluster c is $p\left(y_a \mid \theta_c, x_a\right) = \mathcal{N}\left(\mu_c\left(x_a \mid \theta_c\right), K_c\left(x_a \mid \theta_c\right) + \sigma_c^2 I\right)$.

To infer the number of clusters, we assign a weight w_c on each cluster and place a Dirichlet process prior over these weights $\boldsymbol{w}$. Using Bayes' rule, we can now calculate the probability that a curve a belongs to cluster c (that is, $P\left(z^{(a)} = c\right)$) instead of the other clusters using the weighted proportion of the probability of observing data from the ability curve if it is part of cluster c, that is:

$$P\left(z^{(a)} = c \mid x_a, y_a, \boldsymbol{\theta}, \boldsymbol{w}\right) = \frac{w_c p\left(y_a \mid \theta_c, x_a\right)}{\sum_{c'=1}^{C} w_{c'} p\left(y_a \mid \theta_{c'}, x_a\right)} \quad . \tag{1}$$

Inferring the distribution of the values for the weights ($\boldsymbol{w}$) and the clustering parameters ($\boldsymbol{\theta}$) is achieved using adaptive tempered sequential Monte Carlo sampling [4,8,9].

2.3 Predicting Cluster Assignment

A partial ability curve $a_{t<h}$ contains all data points for ability curve a up until time point h, the latest time point at which data came in during real-time learning. We classify the cluster with the highest probability as the predicted cluster for the partial ability curve using Eq. (1) with only the observations available up until that point; $p\left(z_a = c \mid a_{t<h}, \boldsymbol{\theta}, \boldsymbol{w}\right)$.

We evaluate each cluster's performance using Matthews correlation coefficient (MCC) [15], which is a reliable measure for unbalanced data sets [3,28].

2.4 Interpreting Clusters

After obtaining the clusters, we link them to SRL support needs following [5] by comparing the learning process measures, learning outcomes and the cluster's mean ability curve characteristics of the children assigned to each cluster. The learning process measures are accuracy, effort. Accuracy is measured as the percentage of problems solved correctly, and effort is calculated as the number of problems solved. The learning outcomes are the children's scores on the pre-tests and post-tests, as well as the relative gain between the two tests, which we measure with the normalised learning change [14].

To see whether the learning outcomes and process measures of the children differ significantly between clusters, we run a linear mixed model in R using lmerTest [11] and, if so, t-tests with corrections using Satterthwaite's method [23] to determine which clusters differ significantly from each other on which measures.

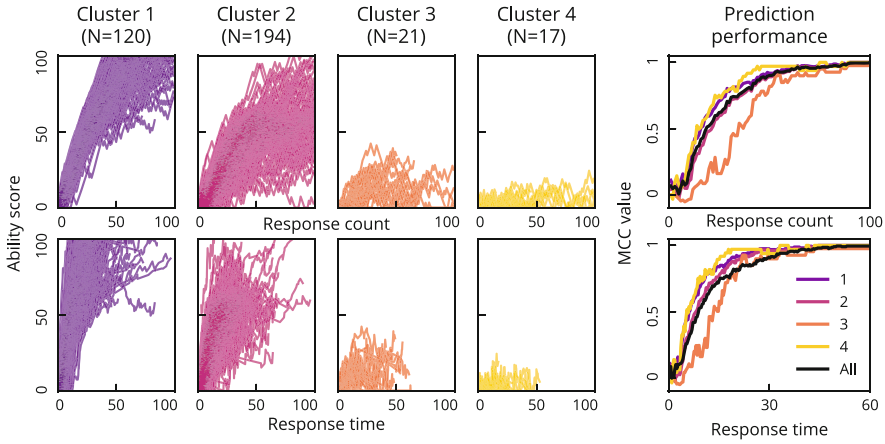


Fig. 1. On the left: the clusters that were found with the assigned ability curves for each cluster. Clusters were sorted on eventual ability score reached. On the right: predictive performance for all clusters, with in black the predictive performance for the whole model.

3 Results

The DPGP model has, based on response count and response time, identified 4 clusters. These clusters are visualized in Fig. 1, showing the ability curves assigned to these clusters. The clusters are sorted by their average ability score after solving 75 problems. Note that since there are two input dimensions (response count and response time), we visualize each cluster and the predictive performance for both dimensions, as this is clearer than a 3-dimensional plot.

The clusters with the best predictive performance are cluster 1 and 4. These clusters have similar MCC values for the first few problems. Afterwards, Cluster 1 is performing slightly better than cluster 4. This is becoming clear at an MCC value of 0.75 which it reaches after 24 problems – out of the average of 72 problems that students work on each topic – whereas cluster 4 needs 28 problems. Cluster 2 is performing similar to the mean of the model. It needs 35 problems to obtain an MCC value of 0.75, slightly more than half of the problems. The model itself on average performs a little better and needs 33 problems to reach an MCC value of 0.75.

Cluster 3 is the hardest to predict and needs up to 43 problems before it reaches an MCC value of 0.75. The reason for the lower performance for cluster 3, is that with less data most of the partial ability curves from cluster 3 get assigned to either cluster 2 or cluster 4. Cluster 2 is a wider cluster and is thus more likely to get curves assigned to it even if the complete ability curve gets assigned to cluster 3. Cluster 4 gets assigned all ability curves that keep the lowest ability scores, so even if children perform better later on, causing their ability curve to be assigned to cluster 3, the initial partial ability curve will be assigned to cluster 4 The performance in prediction over response time gives similar results. The only difference is that cluster 2 now outperforms the model on average. This is possible due to children assigned to cluster 2 giving more responses on average than the rest of the children.

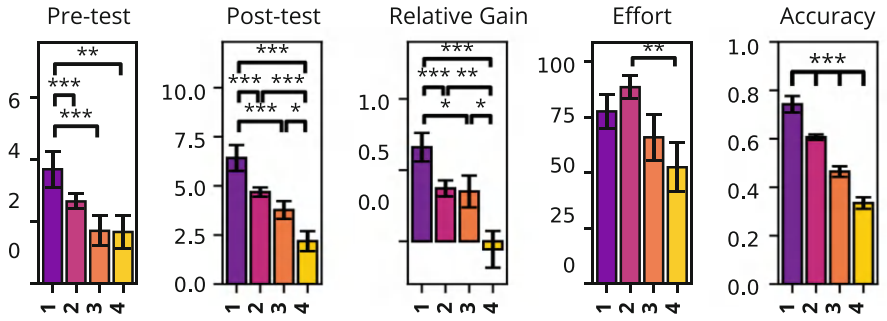


Fig. 2. The learning metrics for the clusters identified by the clustering model. Significance level is indicated with asterisks: * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$.

Relating the clusters to the learning metrics (Fig. 2) shows there is a clear and significant difference in accuracy between all clusters and significant differences in the other metrics between some of the clusters. Detailed results of the experiments are available on GitHub[2].

Cluster 1 demonstrates the highest accuracy of all clusters. Its pre-test and post-test scores are significantly higher than all other clusters, showing that children in this cluster had mastered the skill to a large extent before learning began. The ability curves assigned to this cluster reach the system's highest possible ability level. We call this cluster the *Masters* cluster. Children in this cluster appear to need minimal SRL support, as they either mastered the skill before practice or can master it independently.

Cluster 2 has the second-best accuracy of the clusters, with the post-test scores and relative gains being significantly higher than those of cluster 4 but lower than cluster 1. It also exhibits significantly higher effort compared to cluster 4. Generally, the ability curves assigned to this cluster show an upward trend in ability score, but do not reach the maximum score. Therefore, we refer to the

[2] https://github.com/Dijkstra7/AIED2025_Appendix.

children in this cluster as *Risers*. Risers may require moderate SRL support to achieve mastery.

Children in cluster 3 have significantly lower pre-test scores, post-test scores, and relative gains than those in cluster 1, as well as a significantly lower accuracy than children both in cluster 1 and 2. However, they achieve significantly higher post-test scores, relative gains, and accuracy than children in cluster 4. Additionally, children in this cluster do not exceed above half of the maximum possible ability level. Therefore, we refer to these children as *Strugglers*, and they may need substantial SRL support.

Finally, the children in cluster 4 have significantly lower post-test scores, relative learning gains, and accuracy than all other clusters. They also exhibit significantly lower pre-test scores than the children in cluster 1 and significantly lower effort than those in cluster 2. Additionally, their ability score does not exceed 20% of the maximum possible ability level. This stagnating ability score shows they are *Trailers* and may need extensive SRL support for this learning goal.

4 Discussion

In the current study, we investigated the use of a Dirichlet process Gaussian process model to identify clusters related to children's SRL support needs, to evaluate the model's predictive power in the context of real-time data. Our findings showed that the model's predictive performance is best when incorporating both the response count and the response time. This model resulted in four clusters that indicate different SRL support needs, ranging from the *Masters* cluster that seems to need no SRL support to the *Trailers* cluster, which seems to require extensive support. Notably, these clusters can be accurately predicted after 33 responses or 15 min (out of an average of 71 responses or 35 min in a lesson).

A limitation of this study is that response time is calculated as the interval between solving two problems. However, this interval only indirectly reflects response time. The child could have been receiving additional instructions, have been writing solutions on a scratch path to understand the problem further, have been distracted, or there could have been other reasons why they were not actively engaged in the task. Further work could be done to determine which response times are outliers and which are valid.

The predictive performance after 15 min enables early intervention in children's learning: If we predict that a child may not reach their goal, the ALT can adjust the SRL support to their needs. This support can provide more monitoring information, suggest control strategies, or take over regulation by fine-tuning beyond the current adaptivity of the difficulty level used to select the following problem. If effective, this support may impact their regulation of practice behaviour and improve learning outcomes.

Acknowledgement. This work was supported by the European Research Council [grant number 948786].

References

1. Azevedo, R.: Understanding the complex nature of self-regulatory processes in learning with computer-based learning environments: an introduction. Metacogn. Learn. **2**(2), 57–65 (2007). https://doi.org/10.1007/S11409-007-9018-5
2. Baker, R.S.: Artificial intelligence in education: bringing it all together. In: Digital Education Outlook: Pushing the Frontiers with AI, Blockchain, and Robots, pp. 43–54 (2021). https://doi.org/10.1787/589b283f-en
3. Chicco, D., Jurman, G.: The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min. **16**(1), 4 (2023). https://doi.org/10.1186/s13040-023-00322-4
4. Chopin, N., Papaspiliopoulos, O., et al.: An introduction to sequential Monte Carlo, vol. 4. Springer, Heidelberg (2020). https://doi.org/10.1007/978-3-030-47845-2
5. Dijkstra, R., Hinne, M., Segers, E., Molenaar, I.: Clustering children's learning behaviour to identify self-regulated learning support needs. Comput. Hum. Behav., 107754 (2023). https://doi.org/10.1016/j.chb.2023.107754
6. Greene, J.A., Azevedo, R.: The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. Educ. Psychol. **45**(4), 203–209 (2010). https://doi.org/10.1080/00461520.2010.515935
7. Gynzy: Rekenen. Educational software (2025). https://www.gynzy.com/en. Accessed 17 Feb 2025
8. Herbst, E., Schorfheide, F.: Sequential monte carlo sampling for dsge models. J. Appl. Economet. **29**(7), 1073–1098 (2014). https://doi.org/10.1002/jae.2397
9. Hinne, M.: An introduction to sequential monte carlo for bayesian inference and model comparison–with examples for psychology and behavioral science. Behav. Res. Methods **57**(5), 125 (2025)
10. Klinkenberg, S., Straatemeier, M., Van Der Maas, H.: Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. Comput. Educ. **57**, 1813–1824 (2011). https://doi.org/10.1016/j.compedu.2011.02.003
11. Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B.: lmerTest package: tests in linear mixed effects models. J. Stat. Softw. **82**(13), 1–26 (2017). https://doi.org/10.18637/jss.v082.i13
12. Li, S., Zheng, J., Poitras, E., Lajoie, S.: The allocation of time matters to students' performance in clinical reasoning. In: Nkambou, R., Azevedo, R., Vassileva, J. (eds.) ITS 2018. LNCS, vol. 10858, pp. 110–119. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91464-0_11
13. Li, T., et al.: Analytics of self-regulated learning scaffolding: effects on learning processes. Front. Psychol. **14** (2023).https://doi.org/10.3389/fpsyg.2023.1206696
14. Marx, J.D., Cummings, K.: Normalized change. Am. J. Phys. **75**(1), 87–91 (2007). https://doi.org/10.1119/1.2372468
15. Matthews, B.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Struct. **405**(2), 442–451 (1975). https://doi.org/10.1016/0005-2795(75)90109-9
16. McDowell, I.C., Manandhar, D., Vockley, C.M., Schmid, A.K., Reddy, T.E., Engelhardt, B.E.: Clustering gene expression time series data using an infinite Gaussian process mixture model. PLoS Comput. Biol. **14**(1) (2018). https://doi.org/10.1371/journal.pcbi.1005896

17. Molenaar, I.: Towards hybrid human-AI learning technologies. Eur. J. Educ. **57**(4), 632–645 (2022). https://doi.org/10.1111/ejed.12527
18. Molenaar, I., Horvers, A., Dijkstra, R., Baker, R.S.: Personalized visualizations to promote young learners' SRL: the learning path app. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, pp. 330–339 (2020). https://doi.org/10.1145/3375462.3375465
19. Molenaar, I., Järvelä, S.: Sequential and temporal characteristics of self and socially regulated learning. Metacogn. Learn. **9**(2), 75–85 (2014). https://doi.org/10.1007/s11409-014-9114-2
20. Plass, J., Pawar, S.: Toward a taxonomy of adaptivity for learning. J. Res. Technol. Educ. **52**, 275–300 (2020). https://doi.org/10.1080/15391523.2020.1719943
21. Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning. MIT press, Cambridge (2006). https://doi.org/10.7551/mitpress/3206.001.0001
22. Saint, J., Fan, Y., Gašević, D., Pardo, A.: Temporally-focused analytics of self-regulated learning: a systematic review of literature. Comput. Educ. Artif. Intell. **3**, 100060 (2022). https://doi.org/10.1016/j.caeai.2022.100060
23. Satterthwaite, F.E.: An approximate distribution of estimates of variance components. Biometr. Bull. **2**(6), 110–114 (1946). http://www.jstor.org/stable/3002019
24. Taub, M., Azevedo, R., Bradbury, A.E., Millar, G.C., Lester, J.: Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment. Learn. Instr. **54**, 93–103 (2018). https://doi.org/10.1016/j.learninstruc.2017.08.005
25. van Wetering, M., Booij, E., van Bruggen, W.: Education in an artificially intelligent world. Technical report, Kennisnet (2020)
26. Winne, P.H.: Learning analytics for self-regulated learning. In: Handbook of Learning Analytics, pp. 241–249 (2017). https://doi.org/10.18608/hla17.021
27. Winne, P.H., Hadwin, A.F.: Studying as self-regulated learning. In: Metacognition in Educational Theory and Practice, pp. 277–304. The educational psychology series, Lawrence Erlbaum Associates Publishers, Mahwah (1998). https://doi.org/10.4324/9781410602350-19
28. Yao, J., Shepperd, M.: Assessing software defection prediction performance: why using the matthews correlation coefficient matters. In: Proceedings of the 24th International Conference on Evaluation and Assessment in Software Engineering, pp. 120–129 (2020). https://doi.org/10.1145/3383219.3383232
29. Zhang, T., Taub, M., Chen, Z.: A multi-level trace clustering analysis scheme for measuring students' self-regulated learning behavior in a mastery-based online learning environment. In: LAK22: 12th International Learning Analytics and Knowledge Conference, pp. 197–207. ACM, Online USA (2022). https://doi.org/10.1145/3506860.3506887

Examining the Influence of Students' Personality on Self-regulatory Behaviors and Learning Outcomes During Complex Learning with MetaTutor

Cameron Marano[1(✉)], Megan Wiedbusch[1], Annamarie Brosnihan[1], Matthew Moreno[1], Milla Sherman[2], Maral Karimi[1], Tara Delgado[1], and Roger Azevedo[1]

[1] School of Modeling, Simulation, and Training, University of Central Florida, Orlando, FL 32826, USA
{cameron.marano,megan.wiedbusch,annamarie.brosnihan, matthew.moreno,maral.karimi,tara.delgado,roger.azevedo}@ucf.edu
[2] Department of Psychology, Rollins College, Winter Park, USA
msherman@rollins.edu

Abstract. The utilization of self-regulated learning (SRL) requires enacting complex and cyclical cognitive, affective, metacognitive, motivational, and social processes that may be influenced by individual differences such as personality. The frequency and types of SRL strategies enacted by learners are impacted by individual differences (e.g., motivation orientation, demographics), however current literature has yet to fully examine the influence of learners' personality traits on SRL processes and learning outcomes while using intelligent tutoring systems (ITSs). In this study, we investigated the relationship between undergraduate students' ($N = 110$) personality traits and their SRL processes and learning outcomes while using MetaTutor, an ITS designed to provide adaptive scaffolding while learning about the human circulatory system. Results indicated significant differences in participants' learning outcomes across different personality clusters, as well as differences in the use of note taking, an effective SRL strategy predictive of learning gains. However, there was not a significant interaction or effect on the combination of SRL engagement, learning outcomes, and personality clusters. These results highlight the importance of individual differences in understanding how personality traits influence students' self-regulation and strategy use with ITSs during complex learning. Future ITSs can leverage personality traits through a comprehensive multimodal integration approach that begins with sophisticated assessment methods combining personality inventories, self-reports, and trace data to adapt their content delivery, feedback and scaffolding methods, based on personality traits and other individual differences.

Keywords: Self-regulated Learning · Intelligent Tutoring Systems · Personality

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 325–331, 2025.
https://doi.org/10.1007/978-3-031-98462-4_41

1 Introduction and Current Study

Accurately and consistently deploying self-regulated learning (SRL) strategies, such as setting learning goals, strategizing, deploying, and modifying methods to reach self-set goals, has been shown to be challenging for learners, given the complexity of the cyclical process [1]. This is especially apparent for students learning about complex and challenging STEM topics with advanced learning technologies such as intelligent tutoring systems (ITSs) [2]. To address this challenge, ITSs, such as MetaTutor, have been designed as a scaffolding tool to guide and foster students' SRL processes [1], however such tools have not been able to successfully promote learning for certain students [3].

Previous studies have found that individual differences such as motivation [4] can influence learning outcomes within MetaTutor. The influence of personality traits on SRL and learning outcomes have been explored within educational contexts (e.g., [5–8]). However, research of personality traits' influence on ITSs, such as MetaTutor, has had less focus, but with some promising results (e.g., [9–11]). As such, this study aims to further investigate how and when personality traits as individual differences can impact the deployment and use of learning process with an ITS.

The current study aims to understand how various clusters of personality types impact the learning outcomes and SRL processes while using the ITS, MetaTutor. To address the major gaps in literature, we address the following research questions:

(1) *What kind of personality clusters can be identified across experimental conditions?*
(2) *Are there significant differences of student's normalized learning gains between personality clusters?*
(3) *Is there a significant difference in the user-initiated self-regulated learning strategies with the SRL palette while learning, between personality clusters?*
(4) *Is there a significant difference in normalized learning gains based on the interaction between personality clusters and the use of self-regulated learning strategies that were identified to be different between personality clusters found from RQ3?*

2 Methodology

2.1 Participants and Experimental Conditions

110 undergraduate students ($Mage = 20.3$) from three North American Universities participated in a two-day experimental study with MetaTutor, a hyper-media ITS designed to support learning about the human circulatory system [1]. Participants were randomly assigned to one of two experimental conditions for the learning task: the control condition or the prompt and feedback (PF) condition. Participants in the PF condition ($N = 54$) received scaffolding from the PAs to engage in SRL strategies and received feedback on the quality of their deployment of the SRL strategies. Participants in the control condition ($N = 56$) did not receive prompts to engage in SRL strategies; however, they could self-initiate SRL strategies, but they did not receive any feedback from the PAs on the quality of their engagement.

2.2 MetaTutor

MetaTutor is a hypermedia-based ITS designed to teach students challenging STEM content, such as the human circulatory system, while simultaneously helping them develop and apply self-regulation skills while they are learning [1]. This is achieved with the assistance of embedded PAs that guide students through the system content and encourage the use of SRL strategies [1]. The PAs within MetaTutor (i.e., Sam the Stratregizer, Mary the Monitor, Gavin the Guide, Pam the Planner) each serve a distinct purpose in helping students navigate the system through prompting and scaffolding metacognitive processes, which enable students to deploy the necessary strategies needed to improve their learning gains.

The MetaTutor interface includes an SRL palette, a tool designed to help participants engage in a range of SRL strategies throughout the learning session [1]. The activation of SRL strategies can be self-initiated via the SRL palette or suggested by one of the PAs, which will provide feedback on their performance. The SRL strategies that can be initiated include planning (PLAN), summarizing (SUMM), taking notes (TN), monitoring progress toward goals (MPTG), reviewing and revisiting prior material (RR), coordinating or integrating information sources (COIS), activating prior knowledge (PKA), judgment of learning (JOL), feeling of knowing (FOK), content evaluation (CE), and making inferences (INF) [12].

2.3 Experimental Procedure and Materials

Following the completion of informed consent on day one, participants completed a battery of self-report measures, including a 30-item domain knowledge assessment derived by subject matter experts to evaluate prior knowledge and performance. Participants were also administered the International Personality Item Pool (IPIP) [13], a 50-item self-report measure grounded within the Five Factor Model (FFM). The IPIP includes five subsections, each measuring the major personality constructs (e.g., openness, conscientiousness, extraversion, agreeableness, and neuroticism). On day two, participants completed the learning task with MetaTutor for 60 min, followed by the post-test questionnaires. Participants were debriefed and compensated ($10/hr.) for their participation.

2.4 Coding and Scoring

Learning Gains. To assess students' performance, we used the normalized learning gains (NLGs) equations by Marx & Cummings [14]. These metrics evaluate learning progress by comparing a participant's pre- and post-test scores, normalized to the total possible points and adjusted for prior knowledge.

Log Files. Log files were collected from each participant after their completion of the learning task. The log files provided the frequency counts for participant's self-initiated use of the SRL palette for each strategy. We excluded RR, COIS, PKA, FOK, and CE from our analyses due to an insignificant number of participants self-initiating those strategies. The remaining SRL strategies included in the analysis (i.e., SUMM, PLAN, TN, MPTG,

JOL, INF) were transformed to binary variables (i.e., Did the participants self-initiate the SRL strategy: Yes/No) due to the variation in the frequency of user-initiation with each strategy across participants.

Personality Scoring. The 50-item IPIP [13] scores of each participant were calculated from their pre-test responses. Each subscale is comprised of 10 questions that require a response from 1 (very inaccurate) to 5 (very accurate) and scored by the total sum of each subscale. Each personality type score can range from 10 (low) to 50 (high). We used each participant's subscale score for each of the five subscales to create clusters of participants (see Research Question 1).

3 Results

3.1 Research Question 1: What Kind of Personality Clusters Can Be Identified Across Experimental Conditions?

We conducted a k-means cluster analysis, calculated using Euclidean distance, to determine what kind of personality clusters could be identified across experimental conditions. The following methods proposed three as the best number of clusters: Ball-Hall, 95.41; Scott-Symons, 79.85; Ratkowski-Lance, 0.30; Trace W, 24.39; and Trace of the Within-cluster covariance matrix (TrCovW), 2121.66. Examination of the clusters revealed that Cluster 1 was high in conscientiousness in openness, and low in neuroticism and lower (than the other clusters) in extraversion and agreeableness; Cluster 2 was highest in openness, neuroticism, agreeableness, and extraversion, and the lowest in conscientiousness; Cluster 3 was highest in conscientiousness, and relatively high in neuroticism, extraversion, and agreeableness, and lowest in openness.

We then conducted a post-hoc chi-square of independence test to ensure there was not a significant difference in the distribution of conditions across the 3 clusters identified. Results indicated there was not a significant difference in the distribution of conditions across clusters, X^2 $(2, N = 110) = 5.50, p > 0.05$ (Table 1).

Table 1. Descriptives for Research Question 1

Cluster	Conditions *N*		Personality Subscales				
	Prompt & Feedback	Control	Extraversion	Agreeableness	Neuroticism	Conscientiousness	Openness
1	24	15	28.15	33.74	32.23	28.36	40.38
2	12	23	38.46	38.37	39.89	24.66	40.63
3	18	18	30.19	37.22	36.19	33.28	32.81

3.2 Research Question 2: Are There Significant Differences of Student's Normalized Learning Gains Across Personality Clusters?

We ran a non-parametric Kruskal-Wallis test to determine if there were significant differences of student's normalized learning gains across the personality clusters identified

in RQ1. Normalized learning gains were not normally distributed, so we ran the non-parametric version of a one-way ANOVA. The Kruskal-Wallis test indicated there was a significant difference of learner's normalized learning gains across personality clusters, χ^2 $(2, N = 110) = 7.12, p < 0.05$. To determine how student's normalized learning gains differed across personality clusters, we conducted three post-hoc Mann-Whitney tests to examine each cluster's NLGs against each other. The first Mann-Whitney indicated that the normalized learning gains of Cluster 1 was significantly higher than those in Cluster 2, $p = 0.01$. The second Mann-Whitney indicated that the normalized learning gains of Cluster 1 was significantly higher than those in Cluster 3, $p < 0.05$. However, the third Mann-Whitney indicated that there was no significant difference between the normalized learning gains of those in Cluster 2 and Cluster 3, $p > 0.05$. In sum, these findings indicate that learners high in openness and conscientiousness and lower in neuroticism had significantly higher NLGS than their peers with varied personality distributions.

3.3 Research Question 3: Is There a Significant Difference in the User-Initiated Self-Regulated Learning Strategies with the SRL Palette While Learning Between Personality Clusters?

We ran 6 separate chi-square of independence tests to determine if there was a significant difference in the use of user-initiated SRL strategies with the SRL palette while learning between personality clusters. There was a significant difference in user-initiated note taking between personality clusters, X^2 $(2, N = 110) = 5.82, p < 0.05$. Those in Cluster 1 were more likely to take notes (user-initiated $N = 27$, did not initiate $N = 12$) than those in Cluster 2 (user-initiated $N = 15$, did not initiate $N = 20$) and Cluster 3 (user-initiated $N = 23$, did not initiate $N = 13$). However, the other user-initiated SRL strategies did not appear significantly different between personality clusters, SUMM χ^2 $(2, N = 110)$ $= 0.62, p > 0.05$; PLAN χ^2 $(2, N = 110) = 0.98, p > 0.05$; MPTG χ^2 $(2, N = 110) =$ $1.58, p > 0.05$; JOL χ^2 $(2, N = 110) = 0.47, p > 0.05$; INF χ^2 $(2, N = 110) = 1.52, p$ > 0.05. Overall, these findings indicate that learners in Cluster 1 (high in openness and conscientiousness and lower in neuroticism) were significantly more likely to self-initiate note taking during the learning task than then their peers with more varied personality distributions.

3.4 Research Question 4: Is There an Effect on Normalized Learning Gains Based on the Interaction Between Personality Clusters and the Use of Self-Regulated Learning Strategies that Were Identified to Be Different Between Personality Clusters Found from RQ3?

We ran a linear regression to find if there was an effect on normalized learning gains based on the interaction between personality clusters and the use of SRL strategies that were identified to be different between personality clusters found from RQ3. Taking notes was the only SRL strategy that was found to show significant differences between personality clusters, therefore our linear regression only included normalized learning gains, personality clusters, and taking notes. Results indicated that there were no main effects or interactions of normalized learning gains personality clusters and taking notes, and that the overall model was a not a significant fit, F $(5, 104) = 0.03, p > 0.05$.

These results suggest neither personality cluster, whether a participant took notes, or the interaction between the two factors significantly influenced how much a participant learned using MetaTutor.

4 Conclusion

Previous literature has linked personality types to SRL and metacognition [5], however, there is a lack of literature investigating personality types and its influence on SRL processes and learning outcomes with ITSs. In sum, our study sought out to address the major gaps in literature regarding how and when personality impacts SRL processes and learning outcomes with MetaTutor, an ITS designed to teach students about the human circulatory system. Our results indicated significant differences between personality types and their learning outcomes as well as their note taking, an SRL strategy enacted during the learning process. However, our results did not indicate a significant difference in normalized learning gains based on the interaction between personality types and their note taking. Future research is needed to investigate why there was no interaction or effect between personality types, learning gains, and note taking. Additionally, future research investigating personality types, SRL processes, and learning outcomes within ITSs should integrate a multimodal approach, that allows for real-time feedback and adaptation based on users' prior knowledge, trace data, and personality types.

Acknowledgments. This research was funded by the National Science Foundation (DRL #1916417). The authors would like to thank the members of UCF's SMART Lab for their numerous contributions.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Azevedo, R., et al.: Lessons learned and future directions of metatutor: leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system. Front. Psychol. **13**, 813632 (2022)
2. Azevedo, R., Wiedbusch, M.: Theories of metacognition and pedagogy applied to AIed systems. In: Handbook of Artificial Intelligence in Education, pp. 45–67. Edward Elgar Publishing (2023)
3. Moos, D.C., Azevedo, R.: Self-regulated learning with hypermedia: the role of prior domain knowledge. Contemp. Educ. Psychol. **33**(2), 270–298 (2008)
4. Duffy, M.C., Azevedo, R.: Motivation matters: interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system. Comput. Hum. Behav. **52**, 338–348 (2015)
5. Dorrenbacher, L., Perels, F.: Self-regulated learning profiles in college students: their relationship to achievement, personality, and the effectiveness of an intervention to foster self-regulated learning. Learn. Individ. Diff. **51**, 229–241 (2016)
6. Bidjerano, T., Dai, D.Y.: The relationship between the big-five model of personality and self-regulated learning strategies. Learn. Individ. Diff. **17**(1), 69–81 (2007)

7. Chamorro-Premuzic, T., Furnham, A.: Personality predicts academic performance: evidence from two longitudinal university samples. J. Res. Pers. **37**(4), 319–338 (2003)
8. Eilam, B., Zeidner, M., Aharon, I.: Student conscientiousness, self-regulated learning, and science achievement: an explorative field study. Psychol. Sch. **46**(5), 420–432 (2009)
9. Harley, J.M., et al.: Examining the predictive relationship between personality and emotion traits and students' agent-directed emotions: towards emotionally- adaptive agent-based learning environments. User Model. User-Adap. Interact. **26**, 177–219 (2016)
10. Lallé, S., Taub, M., Mudrick, N.V., Conati, C., Azevedo, R.: The impact of student individual differences and visual attention to pedagogical agents during learning with metatutor. In: Artificial Intelligence in Education: 18th International Conference, AIED 2017, Wuhan, China, June 28–July 1, 2017, Proceedings, vol. 18, pp. 149–161 (2017). Springer
11. Conati, C., Barral, O., Putnam, V., Rieger, L.: Toward personalized XAI: a case study in intelligent tutoring systems. Artif. Intell. **298**, 103503 (2021)
12. Taub, M., Azevedo, R., Bouchet, F., Khosravifar, B.: Can the use of cognitive and metacognitive self-regulated learning strategies be predicted by learners' levels of prior knowledge in hypermedia-learning environments? Comput. Hum. Behav. **39**, 356–367 (2014)
13. Goldberg, L.R.: The development of markers for the big-five factor structure. Psychol. Assess. **4**(1), 26 (1992)
14. Marx, J.D., Cummings, K.: Normalized change. Am. J. Phys. **75**(1), 87–91 (2007)



Leveraging a Multi-agent LLM-Based System to Educate Teachers in Hate Incidents Management

Ewelina Gajewska(✉), Michal Wawer, Katarzyna Budzynska, and Jaroslaw A. Chudziak

Warsaw University of Technology, 00-661 Warsaw, Poland
ewelina.gajewska.dokt@pw.edu.pl

Abstract. Computer-aided teacher training is a state-of-the-art method designed to enhance teachers' professional skills effectively while minimising concerns related to costs, time constraints, and geographical limitations. We investigate the potential of large language models (LLMs) in teacher education, using a case of teaching hate incidents management in schools. To this end, we create a multi-agent LLM-based system that mimics realistic situations of hate, using a combination of retrieval-augmented prompting and persona modelling. It is designed to identify and analyse hate speech patterns, predict potential escalation, and propose effective intervention strategies. By integrating persona modelling with agentic LLMs, we create contextually diverse simulations of hate incidents, mimicking real-life situations. The system allows teachers to analyse and understand the dynamics of hate incidents in a safe and controlled environment, providing valuable insights and practical knowledge to manage such situations confidently in real life. Our pilot evaluation demonstrates teachers' enhanced understanding of the nature of annotator disagreements and the role of context in hate speech interpretation, leading to the development of more informed and effective strategies for addressing hate in classroom settings.

Keywords: Artificial Intelligence · LLMs · Multi-Agent System · Teacher Education · Hate Speech

1 Introduction

Early researchers and educators recognised the potential of artificial intelligence (AI) to revolutionise learning and training. In recent years, virtual classrooms, AI-powered tutoring systems, and interactive learning platforms have become commonplace. Despite these advancements, the widespread adoption of AI in educational training remains limited. Recent advancements in language technologies have introduced new opportunities to create accessible and highly interactive educational experiences. We propose to leverage these technologies to create new educational opportunities for teachers to learn hate incident management skills. The high prevalence of hate in recent years has highlighted the urgent need for

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 332–339, 2025.
https://doi.org/10.1007/978-3-031-98462-4_42

designing effective intervention strategies [19], and teachers, who play a pivotal role in shaping the social and emotional environment of their classrooms, are often on the front lines when such incidents occur.

Consequently, it is important to provide teachers with comprehensive training on how to identify, address, and prevent hate incidents to create a safe and inclusive environment for students. This training should not only equip teachers with the necessary skills to manage these situations effectively but also foster a deeper understanding of the underlying issues that contribute to hate-based behaviour. Trained on massive amounts of data, LLMs harness the power of deep learning to generate human-like responses, provide accurate information, and help individuals complete complex tasks across a variety of domains [1]. A recent review of the transformative potential of AI emphasises LLMs's ability to enhance task proficiency, foster creativity in writing, and augment student engagement during learning sessions [2,3]. LLMs have the potential to revolutionise also medical education by acting as interactive virtual tutors [13].

A number of articles recognise the potential of using AI tools like ChatGPT in teacher training programs to address the disconnection between teachers' theoretical knowledge and real-world teaching applications, often in the form of case-based reasoning [9,15]. To this end, we explore the potential of LLM-based systems as practice-oriented spaces in which teachers can learn to effectively manage hate incidents. Specifically, we propose a multi-agent LLM-based system equipped with a custom personalisation module that is designed to identify hate incidents while being sensitive to cultural differences. We aim to demonstrate how such systems can complement traditional training programs, emphasising the importance of integrating human expertise and ethical considerations. Through this comprehensive analysis, we seek to offer valuable insights into the integration of innovative technologies in educational settings.

2 Managing Hate Incidents

Psychological studies link experience of hateful attacks to both short and long-term effects such as mood swings, emotional harm and erosion of social trust [19]. However, identification of hateful speech can be highly subjective by evoking different reactions and interpretations depending on the individual identity [16,17]. These differences are reflected in the lack of uniform standards of management of hate speech cases as well as differing definitions and policies on hate incidents. On the other hand, from 25% to over 50% of students witness at least one hate incident in school over the period of one year and over 10% of students reported being a victim of hate [8]. Furthermore, over 60% of teachers witnessed a situation where students made publicly offensive comments about the skin colour, sexual orientation, religion, or gender of a group of people. Yet, only 40% of the teachers expressed the intention to intervene in such situations. Moreover, less professionally experienced teachers have been found to employ different intervention strategies than more experienced teachers [4]. Schools provide few educational opportunities for teachers on how to handle hate incidents, leaving them unprepared to address these situations effectively.

Despite the advancements in educational technology, the use of AI in teacher training remains relatively underexplored. Traditional methods such as seminars and in-person training with human experts still dominate the field. These methods, while valuable, have limitations in terms of scalability, personalisation, and continuous learning opportunities. To this end, we design a multi-agent LLM-based platform with the purpose of supporting teachers in managing hate speech in school settings. The system is particularly valuable for understanding the dynamics of hate incidents in order to program effective intervention strategies, providing continuous, scalable, and personalised training opportunities - the key features of educational AI systems [14]. We pose four research questions that regard the development of such a platform: **RQ1.** How does a machine compare to humans in annotating hate speech? **RQ2.** How does the personalisation of agents affect the annotation of hate speech? **RQ3.** How can the proposed platform be utilised in educational workshops, and in which areas does this platform have the greatest potential for utilisation? **RQ4.** What are the criteria for evaluating the usefulness and effectiveness of such a platform in an educational context?

To assess the efficacy of our system, we devise a two-fold evaluation strategy, encompassing both the capability of AI agents to recognise hate speech and the effectiveness of integrating these agents into an interactive system that supports teachers in real time. First, state-of-the-art LLM, GPT-o1-mini, is selected as the basis of our platform. Its performance is assessed with two benchmark datasets: HateXplain [12], which includes explicit hate speech examples, and Latent Hatred [6], which presents more nuanced, implicit, and context-dependent hate speech cases. This evaluation allows us to answer RQ.1 and 2. The second part of our evaluation investigates the practical utility of embedding AI agents into an interactive system designed for teachers. This evaluation aims to determine whether such a system enhances teachers' preparedness and decision-making in real-world scenarios. The evaluation is conducted with a system where teachers can input descriptions of hate incidents and receive structured responses. The agents analyse each case, simulating diverse student perspectives, and provide explanations regarding the nature of hate speech, potential escalation risks, and suggested intervention strategies. This allows us to answer RQ.3 and 4.

3 ARISE: Agent Resource for Incident Support and Education

This section presents the architecture of the proposed ARISE platform. We provide details on the LLMs employed in the study, characteristics of agents, the hate speech analysis workflow, and evaluation of the classification performance.

3.1 Architecture

To explore the impact of external knowledge on model performance, we employed two configurations. The first utilised Retrieval-Augmented Generation (RAG)

[11], providing supplementary materials such as academic articles, case studies, and official definitions of hate speech to enhance contextual understanding. The second configuration relied solely on the intrinsic capabilities of the pre-trained models. Furthermore, we examined the influence of agent collaboration [7] by employing two experimental setups. In the single-agent setup, three independent AI agents, each simulating a university student from different academic disciplines, performed separate analyses. In contrast, the multi-agent setup introduced a manager agent, simulating an academic professor, to facilitate collaboration among the student agents. The results from these configurations provided insights into the models accuracy in detecting hate speech and the potential benefits of multi-agent collaboration.

The system uses CrewAI framework, which implements student-role and domain-specific advisors AI agents and a conversational graphical interface to ensure dynamic and adaptive interaction with user. At the core of the system, a manager agent orchestrates the flow of information and decision-making, ensuring structured analysis. Teachers interact with the system through the manager agent (see Fig. 1), which coordinates responses from different sub-agents specialised in distinct aspects of hate speech analysis and intervention. The primary components include: **(1) Manager Agent** acts as a supervisor, facilitating collaboration among student agents. Modelled as an professor, it aggregates insights, resolves conflicts, and refines recommendations for intervention strategies. **(2) Student Agents** simulate students with different academic backgrounds, including psychology, pedagogy, and cognitive science. Each student agent analyses incidents from its respective disciplinary perspective, ensuring diverse interpretations of hate speech cases. **(3) Advisory Agents** analyses incidents through diverse cultural lenses, examining how different cultural backgrounds and lived experiences might affect the interpretation of potentially hateful content. At the end, these agents provide feedback to the user. **(4) RAG Module** – We incorporated dynamic retrieval of relevant contextual materials, such as case studies, legal definitions, and academic research. **(5) Conversational Interface** – Teachers interact with the system through a user-friendly chat interface, where they can describe incidents, ask questions, and receive real-time feedback.

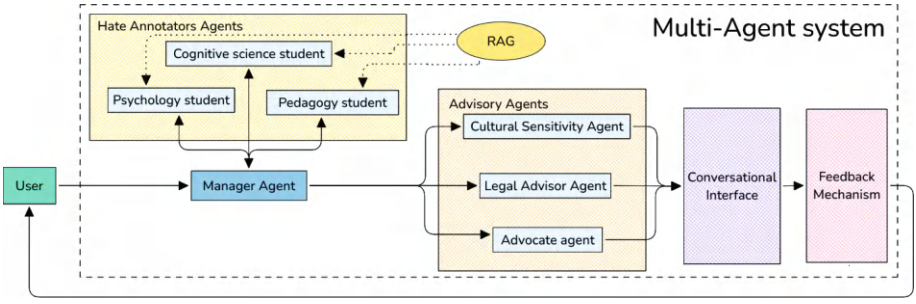


Fig. 1. Architecture of the ARISE platform.

3.2 LLM-Based Agents

Our system employs three specialised student agents, each with a unique backstory and perspective that influences their approach to hate speech detection. These personas are designed to leverage different academic disciplines and personal experiences, enabling a more comprehensive analysis of potential hate incidents. Backstories are not merely narrative elements but are integrated into each agent's prompt engineering [10,18], influencing how they process and analyse potential hate incidents. By maintaining consistent personas across interactions, the agents provide reliable and specialised perspectives that complement each other in the analysis process [5]. This approach allows for a multi-faceted examination of hate speech incidents, combining emotional intelligence, educational insight, and technical analysis in a single system.

3.3 Evaluation of Classification Performance

We use 100 samples from each dataset to evaluate the performance of different agent configurations and models. Results are presented in Table 1. The accuracy of detecting explicit hate speech cases from the HateXplain dataset ranges from 72% to 79%. The implementation of RAG improves performance across most configurations, with an average increase of 4.5 p.p.. The multi-agent setup combined with RAG achieves the highest accuracy (79%), suggesting that collaborative analysis can enhance detection accuracy and outperform single-agent setups. The Latent Hatred dataset presents a more complex challenge, requiring models to classify implicit forms of hate speech into seven distinct categories: grievance, incitement, stereotypes, inferiority, irony, threats and the other category. In this task, the performance of GPT-o1-mini ranges from 47% to 60%, substantially lower than for detecting explicitly hateful speech. Again, the multi-agent setup consistently outperforms single-agent configurations, reaching 60% accuracy. The RAG implementation provides improvements of 1–7 p.p. in classification accuracy.

Table 1. Performance of GPT-o1-mini across different student types and setups.

gpt-o1-mini	Psychology student	Pedagogy student	Cognitive science student	Mixture of agents
HateXplain w/o RAG	72%	76%	74%	78%
HateXplain w/ RAG	79%	78%	79%	79%
Implicit-Hate w/o RAG	51%	49%	47%	55%
Implicit-Hate w/ RAG	55%	56%	53%	60%

4 Case Studies: Application of ARISE in Teacher Education

In this section, we delve into the usability and applicability of the ARISE platform within the realm of teacher education to answer RQ.3 and 4. To this end, we design a scenario of a workshop, the SafeSchools Initiative, aimed at educating teachers about challenges and effective management of hate incidents, with semi-structured interviews conducted with a group of professional school teachers after such a workshop session with the system. As part of this initiative, educators recognised a need for enhanced training in effectively managing and addressing these incidents when they occur. Teachers faced challenges in knowing how to respond appropriately when hate incidents occurred, partly due to a lack of specific training and the complex nature of these situations. Traditional training methods were time-consuming and limited in the number of training scenarios. To this end, the ARISE platform was designed to simulate diverse contexts and provide personalised guidance using real-world examples.

Results of the study confirm the scarcity of precise guidelines for managing hate incidents in schools, although teachers feel confident in their ability to intervene effectively. Some teachers report that interventions should focus on educating students that words matter and the language they use can harm others even if employed humouristically. Overall, the system's analysis of hate incidents is rated as satisfactory, although one individual noted overconfidence and overinterpretation in some cases as the system's weakness. Teachers acknowledged the representation of a multifaceted nature of hate, including the various underlying causes and impacts on students the system provides. At the same time, teachers raised the need for some kind of comparative analysis of hate instances against a "legal norm" of the country or school. We will explore adding such a module in future studies with the system as it necessitates the inclusion of a larger team of researchers who would evaluate the accuracy of this functionality.

5 Conclusions

This study highlights the transformative potential of LLMs in enhancing teachers' skills of effective hate incident management. While the application of artificial intelligence in educational environments has been limited, our multi-agent LLM-based system demonstrates a promising approach to simulating realistic scenarios that educators may encounter. By addressing the complexities associated with hate speech—such as context dependence, varying definitions, and annotator disagreements—our system equips teachers with the necessary tools to navigate these challenges effectively. The pilot evaluation indicates that teachers gained a deeper understanding of the nuances of interpreting hate speech cases, increasing their skills in managing such sensitive situations. As educational institutions continue to strive for inclusivity and safety, integrating advanced AI technologies like the proposed ARISE platform can play a crucial role in preparing educators to confront and mitigate hate incidents proactively.

Acknowledgment. We would like to acknowledge that the work reported in this paper has been supported in part by the Polish National Science Centre, Poland (ChistEra IV) under grant 2022/04/Y/ST6/00001.

References

1. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Raiaan, M.A.K., et al.: A review on large language models: architectures, applications, taxonomies, open issues and challenges. IEEE Access **12** (2024)
3. Alshahrani, K., Qureshi, R.J.: Review the prospects and obstacles of ai-enhanced learning environments: the role of chatgpt in education. Int. J. Mod. Educ. Comput. Sci. **16**(4), 71–86 (2024)
4. Bilz, L., Fischer, S.M., Kansok-Dusche, J., Wachs, S., Wettstein, A.: Teachers' intervention strategies for handling hate-speech incidents in schools. Soc. Psychol. Educ., 1–24 (2024)
5. Cinkusz, K., Chudziak, J.A.: Towards LLM-augmented multiagent systems for agile software engineering. In: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, pp. 2476–2477. Association for Computing Machinery. New York (2024). https://doi.org/10.1145/3691620.3695336. isbn: 9798400712487
6. ElSherief, M., et al.: Latent hatred: a benchmark for understanding implicit hate speech. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 345–363. Association for Computational Linguistics, Online and Punta Cana (2021). https://aclanthology.org/2021.emnlp-main.29
7. Harbar, Y., Chudziak, J.A.: Simulating oxford-style debates with llm-based multiagent systems. In: Proceedings of the ACIIDS 2025 : 17th Asian Conference on Intelligent Information and Database Systems (2025)
8. Kansok-Dusche, J., et al.: A systematic review on hate speech among children and adolescents: definitions, prevalence, and overlap with related phenomena. Trauma Viol. Abuse **24**(4), 2598–2615 (2023)
9. Kim, J.: Leading teachers' perspective on teacher-ai collaboration in education. Educ. Inf. Technol. **29**(7), 8693–8724 (2024)
10. Kostka, A., Chudziak, J.A.: Synergizing logical reasoning, knowledge management and collaboration in multi-agent llm system. In: Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation. Association for Computational Linguistics (2024)
11. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks (2021). https://arxiv.org/abs/2005.11401
12. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: Hatexplain: a benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 14867–14875 (2021)
13. Mondal, H., Marndi, G., Behera, J.K., Mondal, S.: Chatgpt for teachers: practical examples for utilizing artificial intelligence for educational purposes. Indian J. Vasc. Endovasc. Surg. (2023)
14. Oyekoya, O., Urbanski, J., Shynkar, Y., Baksh, A., Etsaghara, M.: Exploring first-person perspectives in designing a role-playing vr simulation for bullying prevention: a focus group study. Front. Virtual Real. **2**, 672003 (2021)

15. Pitura, J., Kaplan-Rakowski, R., Asotska-Wierzba, Y.: The vr-ai–assisted simulation for content knowledge application in pre-service efl teacher training. In: TechTrends, pp. 1–11 (2024)
16. Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the reliability of hate speech annotations: the case of the European refugee crisis. In: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication/Social Media, pp. 6–9. Ruhr-Universitat Bochum (2016)
17. Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., Smith, N.A.: Annotators with attitudes: how annotator beliefs and identities bias toxic language detection. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5884–5906. Association for Computational Linguistics, Seattle (2022). https://doi.org/10.18653/v1/2022.naacl-main.431. https://aclanthology.org/2022.naacl-main.431/
18. Zhao, P., Jin, Z., Cheng, N.: An in-depth survey of large language model-based artificial intelligence agents. arXiv:2309.14365v1 [cs.CL] (9 2023)
19. Ștefăniță, O., Buf, D.M.: Hate speech in social media and its effects on the LGBT community: a review of the current research. Romanian J. Commun. Public Relat. **23**(1), 47–55 (2021)



Leveraging Lecture and Student Knowledge for Improved AI Chatbot Responses in Education

Atsushi Shimada(✉)

Kyushu University, Fukuoka, Japan
atsushi@ait.kyushu-u.ac.jp

Abstract. AI-powered chatbots are increasingly used in education to provide personalized support, especially in large-scale and online learning environments. However, concerns remain regarding the factual accuracy of generative AI responses and potential over-reliance by students. This study implemented a Retrieval-Augmented Generation (RAG)-based chatbot in a Digital Signal Processing course, leveraging both lecture materials and student-contributed study articles. A comparative experiment with four configurations, using lecture knowledge, student knowledge, both, or neither, revealed that the hybrid approach consistently produced the most accurate and contextually relevant responses. Our findings demonstrate that combining structured academic content with student-contributed contextual insights enhances chatbot performance. This study offers empirical guidance for designing AI-enhanced educational tools that promote engagement and deeper understanding.

Keywords: AI chatbot · Retrieval-Augmented Generation · Learning support · Knowledge integration

1 Introduction

Artificial intelligence (AI) chatbots have become essential tools for providing personalized support in higher education [7,16,28]. Particularly in large-scale lectures and online environments, they enable immediate responses and timely feedback [11]. As noted by [8], AI contributes to enhancing learning experiences, improving administrative efficiency, and supporting grading and assessment.

However, concerns remain. Beyond pedagogical issues [3,13,14] like decreased student engagement and over-reliance on AI, there are risks regarding privacy, data security, and especially the accuracy and contextual appropriateness of AI-generated responses. Large language models (LLMs), while powerful, often produce incorrect or misleading information [1,27], posing challenges in fields requiring precise, domain-specific knowledge.

To improve reliability, Retrieval-Augmented Generation (RAG) enhances LLMs by incorporating external knowledge bases [6,24]. RAG-based chatbots have been applied successfully in programming [12,15], mathematics [10], and MOOCs [18,19], improving factual grounding. However, prior studies often rely

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 340–347, 2025.
https://doi.org/10.1007/978-3-031-98462-4_43

solely on lecture or external content, lack empirical validation in content-based learning, and seldom explore the effects of combining multiple knowledge types.

To address these gaps, we developed a RAG-based chatbot that integrates two sources of knowledge: lecture materials to ensure factual accuracy and alignment, and student-authored articles offering contextualized insights and diverse perspectives.

This study addresses two key questions: (RQ1) "How effective is RAG in improving chatbot responses in educational settings?" and (RQ2) "What type of knowledge base best supports learning through AI-powered chatbots?" We empirically examine how combining structured and experiential knowledge improves chatbot performance and informs future AI-enhanced educational systems.

2 Related Work

RAG-based educational chatbots have been studied in diverse domains. Burgan et al. [2] developed RamChat, combining adaptive LLMs and LangChain for accurate university handbook responses. Joseph et al. [25] reported improved student retention in writing assignments with RAG-based tutors.

In programming education, Lang et al. [15] applied a RAG chatbot to an online R course, enhancing response quality for advanced topics. Ko et al. [12] also used RAG for Python support, improving engagement and automated feedback. For mathematical Q&A, Owen et al. [10] showed that textbook-based retrieval improved accuracy and reduced hallucinations. A balanced combination of retrieval and generative reasoning was found most effective.

In the context of Massive Open Online Courses (MOOCs), which provide large-scale, open-access online learning environments, RAG was used to enhance discussion responses [18,19] and course recommendations via the RAMO system [22], supporting personalized learning paths.

RAG's impact in higher education was further shown in OwlMentor [26], which supports literature comprehension, and a chatbot for prospective students [23], which addressed retrieval-efficiency trade-offs. Overall, these studies show that RAG-enhanced chatbots improve factual accuracy, response relevance, and learner engagement. Yet, challenges remain in optimizing retrieval strategies, balancing clarity and grounding, and ensuring ethical alignment.

3 Method

3.1 Overview

Figure 1 shows an overview of our question-and-answer (Q&A) system, based on the Retrieval-Augmented Generation (RAG) framework [17]. The knowledge base integrates two sources: digital textbooks used in lectures and study articles submitted by students who previously took the same course. By combining factual, curriculum-aligned materials with student-generated experiential insights, the system aims to enhance both the accuracy and contextual relevance of chatbot responses, providing more personalized learning support.

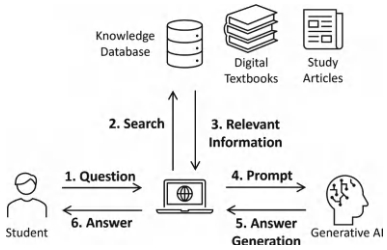


Fig. 1. An overview of the proposed Q&A system.

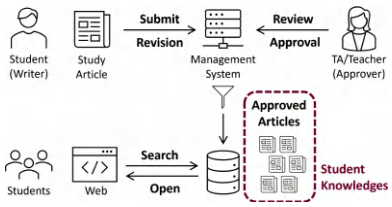


Fig. 2. A submission flow of study articles. In this study, the approved articles were used for a student knowledge database in RAG.

3.2 Digital Textbooks as Lecture Knowledge

The authors' university provides digital textbooks via LMS and content delivery systems [20]. In this study, lecture materials used in the course were employed as the RAG knowledge base. This ensures that chatbot responses are aligned with course content and improves factual accuracy. However, relying solely on lecture materials limits flexibility, making it difficult to handle questions that go beyond what is explicitly covered in the texts.

3.3 Study Articles as Student Knowledge

Study articles were collected from students via a dedicated repository [21](see Fig. 2), where each student ("Writer") submitted insights from their learning experiences. Submissions were reviewed by teaching assistants or instructors ("Approvers") for factual accuracy, clarity, and consistency with lecture content. Revisions were requested when necessary, and only approved articles were published on a course-specific website. These curated articles, which captured diverse student perspectives beyond lecture materials, served as the second knowledge base in RAG.

3.4 Implementation

We developed a Q&A system using a no-code framework, Dify [4], and implemented it in a university-level Digital Signal Processing (DSP) course to evaluate the effectiveness of combining lecture and student-contributed knowledge. As shown in Fig. 3, the workflow begins with a student query ("START"), followed by retrieval from two knowledge bases ("KNOWLEDGE RETRIEVAL") and response generation via GPT-4o ("LLM").

Lecture knowledge was built from eleven digital signal processing textbooks, preprocessed into approximately 20 chunks per file using 500-token segmentation. For student knowledge, 812 study articles collected over three years were curated, with only approved articles included. Personally identifiable information was



Fig. 3. A workflow of chatbot with RAG.

Table 1. Correspondence between System Configurations and Student Groups

System Number	System Configuration	Student Group
S1	Lecture knowledge + Student knowledge	G1
S2	Lecture knowledge only	G2
S3	Student knowledge only	G3
S4	Without RAG (No local knowledge used)	G4

removed, and contents were exported in JSON format for integration. Student submissions yielded about 400 chunks.

The prompt used for the LLM is shown below (originally written in the course language and translated into English):

> "Provide a straightforward and accurate answer to the student's question about digital signal processing based on the retrieved knowledge. The goal is to help the student better understand the question. Answers should be factual, avoid bias or stereotypes, and use appropriate terminology with clarifications if needed."

4 Experiments

4.1 Experimental Settings

The experiment involved both teacher and student evaluations to assess the quality and effectiveness of chatbot responses. We designed experiments using four different system configurations (see Table 1). To ensure fairness in response generation, both lecture and student-contributed knowledge sources were used equally in the retrieval process for systems where both were available, avoiding prioritization of either source.

The teacher, who was the sole course instructor, evaluated system responses using 98 questions generated from the lecture materials and categorized into four types: basic, applied, advanced, and key concepts. Each response was rated on a five-point Likert scale according to three criteria:

- **Q1: Match the Intent.** Does the response match the intent of the question?
- **Q2: Match the Contents.** Is the response consistent with course content?
- **Q3: Correctness.** Is the response factually accurate?

Table 2. Teacher Evaluation Results by Question Category

	Basic				Applied				Advanced				Key Points			
	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Q1: Match the intent	**4.4**	4.3	4.3	3.4	**4.4**	4.2	3.8	3.4	**4.3**	3.8	4.0	3.1	**4.3**	3.8	4.2	3.1
Q2: Match the contents	**4.4**	4.2	4.0	3.2	**4.1**	3.7	3.5	3.1	**4.1**	3.9	3.8	2.9	4.3	**4.5**	4.0	3.0
Q3: Correctness	**4.4**	4.3	4.1	3.9	**4.3**	**4.3**	4.1	3.8	**4.1**	4.0	4.0	3.6	4.3	4.2	**4.4**	3.8

Student evaluations were conducted over one week after the course ended (August 6–13, 2024), ensuring participants had an understanding of the course. This timing helped avoid misjudgment due to partial knowledge. Students were informed of the study, and consent was obtained via an online form. Before using the system, students were instructed to formulate questions with at least five keywords and carefully evaluate the system's responses based on intent match and factual correctness, referring to course materials as needed.

65 students were randomly assigned to one of four groups (G1–G4), each using a different system configuration (S1–S4), with attention paid to balancing their prior academic performance and course engagement level to minimize bias. They were not informed about the differences between the system configurations, ensuring a single-blind condition. Students who did not follow instructions were excluded, resulting in 40 valid participants (10 per group).

In addition to Q1–Q3, students also rated the following two criteria:

- **Q4: New Knowledge.** Whether the response offered new insights.
- **Q5: Usefulness.** Whether the system was helpful for learning.

4.2 Teacher Evaluation of System Responses

Table 2 summarizes the teacher's evaluation across four question types (basic, applied, advanced, key points) and three criteria (Q1–Q3). S1 (Lecture + Student knowledge) achieved the highest scores in all categories, indicating that combining structured and experiential knowledge improves both quality and relevance. S2 (Lecture only) performed well for basic and applied questions, likely due to its alignment with course content. S3 (Student only) offered flexible and nuanced responses but lacked consistency, while S4 (No RAG) scored the lowest across all metrics.

In the basic category, S1, S2, and S3 performed similarly, suggesting that foundational questions can be addressed using either knowledge type. For applied and advanced questions, S1 had a clear advantage, benefiting from the complementary nature of its hybrid knowledge base. In key point questions, S1 outperformed others overall, with S3 excelling in intent and correctness, and S2 in content alignment, supporting the idea that hybrid knowledge offers balanced and comprehensive learning support.

Table 3. Average of questionnaire scores(five scales)

	G1 (N = 9)	G2 (N = 10)	G3 (N = 9)	G4 (N = 9)
Q1: Match the intent	**4.4**	**4.4**	4.1	4.1
Q2: Match the contents	**4.8**	4.6	4.2	4.0
Q3: Correctness	**4.3**	4.2	4.0	3.7
Q4: New knowledge	**4.3**	3.6	3.2	3.8
Q5: Usefulness	**5.0**	4.6	4.2	4.1

4.3 Student Evaluation of System Responses

Table 3 shows the average questionnaire scores from the four groups (G1–G4), each using a different system. G1, which used both lecture and student knowledge, consistently assigned the highest ratings, particularly for Q5 (Usefulness, 5.0) and Q2 (Content Match, 4.8), indicating strong satisfaction with response quality and relevance. Although G3 and G4 (which lacked lecture knowledge) gave lower scores in Q3 (Correctness) and Q4 (New Knowledge), all group averages exceeded 3.0, suggesting a generally favorable impression of the chatbot. This implies that while all systems were positively received, integrating lecture content was important for factual accuracy and deeper insights.

Overall, the hybrid system (S1) met student expectations most effectively, highlighting the value of combining structured and student-contributed knowledge for educational chatbot support. To assess the significance of group differences, we conducted Mann-Whitney U tests with Holm-Bonferroni correction. G1 showed significantly higher ratings than G2–G4 in Q5 (Usefulness), and also in Q4 (New Knowledge) compared to G3. Although not all differences were statistically significant, G1 consistently received the highest scores across all items.

5 Conclusion

Regarding RQ1, our results confirm that RAG-based systems effectively support content learning. Systems incorporating RAG (S1–S3) received consistently higher evaluations than the baseline (S4), based on both teacher and student assessments. Regarding RQ2, the hybrid approach combining lecture materials and student-contributed articles proved most effective. While lecture content ensured factual accuracy, student articles enriched responses with diverse perspectives. This integration led to superior performance across all question types, balancing precision and adaptability for robust learning support.

This study was limited to a single course and a short-term evaluation. Broader validation across disciplines and extended deployments are needed to assess long-term educational impact. Future work should explore long-term classroom integration and evaluate advanced retrieval techniques such as GraphRAG [5] and DeepRAG [9] to further enhance contextual accuracy.

Acknowledgements. This work was supported by JST CREST Grant Number JPMJCR22D1 and JSPS KAKENHI Grant Number JP22H00551, Japan.

References

1. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021)
2. Burgan, C., Kowalski, J., Liao, W.: Developing a retrieval augmented generation (rag) chatbot app using adaptive large language models (llm) and langchain framework. In: Proceedings of the West Virginia Academy of Science (2024)
3. Chang, D.H., Lin, M.P.C., Hajian, S., Wang, Q.Q.: Educational design principles of using ai chatbot that supports self-regulated learning in education: goal setting, feedback, and personalization. Sustainability **15**(17) (2023)
4. Dify Project: Dify: The Innovation Engine for GenAI Applications (2023). https://dify.ai/. Accessed 01 May 2025
5. Edge, D., et al.: From local to global: a graph rag approach to query-focused summarization. ArXiv arxiv:2404.16130 (2024). https://api.semanticscholar.org/CorpusID:269363075
6. Gao, Y., et al.: Retrieval-augmented generation for large language models: a survey (2024)
7. Gill, S.S., et al.: Transformative effects of chatgpt on modern education: emerging era of ai chatbots. Internet Things Cyber-Phys. Syst. **4**, 19–23 (2024)
8. Gökçearslan, Ş, Tosun, C., Erdemir, Z.: Benefits, challenges, and methods of artificial intelligence (ai) chatbots in education: a systematic literature review. Int. J. Technol. Educ. **7**, 19–39 (2024)
9. Guan, X., et al.: Deeprag: thinking to retrieval step by step for large language models (2025)
10. Henkel, O., Levonian, Z., Li, C., Postle, M.: Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. In: Proceedings of the 17th International Conference on Educational Data Mining, pp. 315–320 (2024)
11. Huang, W., Hew, K.F., Fryer, L.K.: Chatbots for language learning - are they really useful? A systematic review of chatbot-supported language learning. J. Comput. Assist. Learn. **38**(1), 237–257 (2022)
12. Ko, H.T., Liu, Y.K., Tsai, Y.C., Suen, S.: Enhancing python learning through retrieval-augmented generation: a theoretical and applied innovation in generative ai education. In: Innovative Technologies and Learning, pp. 164–173 (2024)
13. Kooli, C.: Chatbots in education and research: a critical examination of ethical implications and solutions. Sustainability **15**(7) (2023)
14. Labadze, L., Grigolia, M., Machaidze, L.: Role of ai chatbots in education: systematic literature review. Int. J. Educ. Technol. High. Educ. **20** (2023)
15. Lang, G., Gürpinar, T.: Use of a retrieval-augmented generation (rag) chatbot in an online r programming course. In: Proceedings of the 2024 ISCAP Conference (2024)
16. Lee, Y.F., Hwang, G.J., Chen, P.Y.: Impacts of an ai-based chabot on college students' after-class review, academic performance, self-efficacy, learning attitude, and motivation. Educ. Technol. Res. Dev. **70** (2022)

17. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (2020)
18. Miladi, F., Psyché, V., Lemire, D.: Comparative performance of gpt-4, rag-augmented gpt-4, and students in moocs. In: Breaking Barriers with Generative Intelligence. Using GI to Improve Human Education and Well-Being, pp. 81–92. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-65996-6_7
19. Miladi, F., Psyché, V., Lemire, D.: Leveraging gpt-4 for accuracy in education: a comparative study on retrieval-augmented generation in moocs. In: Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, pp. 427–434 (2024)
20. Ogata, H., et al.: e-book-based learning analytics in university education. In: IEEE International Conference on Consumer Electronics (2015)
21. Okai, S., Minematsu, T., Okubo, F., Taniguchi, Y., Uchiyama, H., Shimada, A.: A system to realize time- and location-independent teaching and learning among learners through sharing learning-articles. In: Towards a Collaborative Society Through Creative Learning, pp. 475–487 (2023)
22. Rao, J., Lin, J.: Ramo: retrieval-augmented generation for enhancing moocs recommendations (2024)
23. Shanthakumar, A.K., Fassihi-Tash, F., Lotfi, A., Bird, J.J.: Retrieval augmented large language model chatbots in00*a*0*highereducation* : *astudyon*00a0university open days. In: Advances in Computational Intelligence Systems, pp. 32–44 (2024)
24. Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., Nanayakkara, S.: Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. Trans. Assoc. Comput. Linguist. **11**, 1–17 (2023)
25. Slade, J.J., Hyk, A., Gurung, R.A.R.: Transforming learning: assessing the efficacy of a retrieval-augmented generation system as a tutor for introductory psychology. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting (2024)
26. Thüs, D., Malone, S., Brünken, R.: Exploring generative ai in higher education: a rag system to enhance student engagement with scientific literature. Front. Psychol. **15** (2024)
27. Weidinger, L., et al.: Ethical and social risks of harm from language models (2021)
28. Yang, S., Evans, C.: Opportunities and challenges in using ai chatbots in higher education. In: Proceedings of the 2019 3rd International Conference on Education and E-Learning, ICEEL '19, pp. 79–83 (2020)

Note-Driven RAG for Learner Performance Estimation via Controlling LLM Knowledge

Tsubasa Minematsu[1(✉)] and Atsushi Shimada[2]

[1] Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Kitakyushu, Japan
minematsu@ai.kyutech.ac.jp

[2] Faculty of Information Science and Electrical Engineering, Kyushu University, Kitakyushu, Japan
atsushi@ait.kyushu-u.ac.jp

Abstract. Learner performance estimation is a critical topic in education, offering the potential to model learner understanding. Large language models (LLMs) have recently gained attention for student modeling and performance estimation. However, because LLMs are pretrained on vast amounts of information, they often possess knowledge that exceeds that of individual learners. As a result, an LLM may correctly answer questions without relying on the content explicitly provided by the learner.We propose a learner response estimation method based on the learner's note-driven Retrieval Augmented Generation (RAG) to discuss how we can model LLMs as student models. In an on-demand classroom environment, we experimentally collect participants' notes as learners acquire knowledge. We evaluate whether notes are helpful in LLM-based learner performance estimation and investigate the necessary adjustments for LLMs.

Keywords: large language model · performance estimation

1 Introduction

Estimating learners' performance and knowledge is critical for providing appropriate feedback in learning support systems such as Intelligent Tutoring Systems (ITS). The student model, representing learners' progress and knowledge, plays a crucial role in deciding tutoring strategies and ITS design [14]. Numerous methodologies have been developed to estimate learners' performance, including machine learning-based approaches [15,18] and statistical models [4].

Large Language Models (LLMs) have recently gained attention due to their adaptability to diverse tasks, driven by their high accuracy in contextual understanding. In the field of performance estimation, advances in research and development related to automated essay scoring [11] and knowledge tracking (KT) [16] continue to progress. Unlike traditional performance estimation models, LLMs can adapt dynamically based on the input text. Through in-context learning [2], they can flexibly adjust to various tasks and learner inputs without changing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 348–355, 2025.
https://doi.org/10.1007/978-3-031-98462-4_44

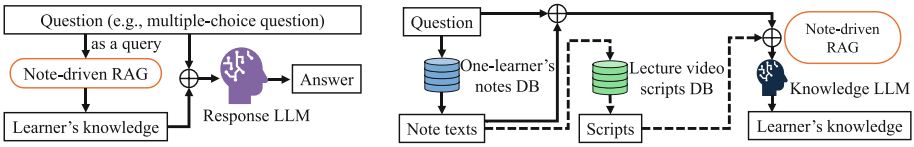
LLM's parameters. For example, a knowledge-restricted LLM by the prompt engineering approach acts as a student model [6]. Inserting accurate knowledge into the LLM has improved the correction of multiple-choice question (MCQ) responses from incorrect to correct. The finding suggests that integrating learners' acquired knowledge into LLMs enhances their ability to reflect learners' understanding. However, there is little discussion of several key challenges, such as determining the optimal type of knowledge to insert, developing LLM knowledge restriction techniques, and identifying effective insertion methodologies.

We propose a Question and Answering method based on Retrieval-Augmented Generation (RAG) [9] as our baseline model to tackle these challenges. This approach aims to integrate learners' progress data into LLMs. We evaluate the proposed method by comparing inserted data, strategies, and techniques for restricting LLM knowledge. This study uses learners' notes as the primary data source to reflect their learning progress. Because notes can be gathered without specialized systems, they are a practical choice for developing baseline models. In addition, note-taking is recognized as an effective learning strategy [1]. The completeness of notes shows a positive correlation with academic performance [3]. Therefore, learners' notes are valuable data sources reflecting their acquired knowledge. Our research questions of this study are "RQ1: To what extent does using learners' notes improve the accuracy of LLM-based performance estimation?" and "RQ2: How do variations in the types of inserted data, insertion methods, and LLM knowledge restriction techniques affect estimation accuracy?".

2 Related Work

Performance Estimation. Performance estimation can be realized through a machine-learning approach and statistical models such as Bayesian Knowledge Tracing [4], Item Response Theory utilizing deep learning [18], and Automated Essay Scoring by natural language processing [11]. These methods estimate learners' knowledge states and performance based on response data and test results. On the other hand, some approaches estimate performance directly from learners' learning activities. An LLM-based method, LLMKT, has recently been introduced to estimate knowledge states from dialogues between tutors and learners [16]. However, research on using LLMs to estimate knowledge from natural language-based learning activities remains insufficient, and there is a lack of empirical validation. Furthermore, while LLMKT is a supervised approach, our study adopts an unsupervised approach. Our method eliminates the need for pre-collected training data, making it more broadly applicable and scalable.

Student Model in ITS. Before the emergence of LLMs, ITSs such as AutoTutor [5] inferred learners' misconceptions through dialogues with the system. To directly construct a student model by learner's operations, Betty's Brain [8] implemented a teachable agent model where learners taught concepts, and the system assessed knowledge based on constraints about the learning concepts. However, these ITSs faced high authoring costs [7], requiring the development of dedicated systems for each learning topic. In contrast, LLMs allow for



(a) Overview (b) RAG details

Fig. 1. Note-driven RAG.

prompt-based scenario setting and role configuration, making it easier to establish chat-based learning environments with minimal modifications [13]. Teachable Agent by LLM is especially relevant to our study. Jin et al. developed LLM-based Teachable Agents in programming education, demonstrating that LLMs can have knowledge constraints and update knowledge states accordingly [6]. However, this study focused on specific topics (e.g., binary search) and has not yet explored how LLM-based ITSs can be applied to semester-long university courses covering multiple topics. Moreover, further discussion is needed on prompt design strategies to enforce knowledge restriction in such scenarios.

3 Learner Response Estimation via Note-Driven RAG

As shown in the overall structure in Fig. 1a, our proposed method adopts a mechanism where a question is input into the LLM, and the LLM outputs a response. To extract the learner's understanding, we use RAG using the learner's notes. A separate database for RAG is constructed for each learner. Switching databases according to the learner makes it possible to simulate the corresponding learner's responses to various questions. The solid-line flow in Fig. 1b illustrates the specific steps of RAG. Before inference, the learner's notes are divided into chunks, and a vector store is prepared as a database for RAG using text embedding. The question text is received as an RAG query during inference, and its embedding vector is computed similarly to when constructing the vector store. Then, related notes are retrieved from the vector store using cosine similarity. If the retrieved notes are used as they are, the LLM may struggle to grasp key points if the text is excessively long. Therefore, another LLM (knowledge LLM) is introduced to extract knowledge from the retrieved learning notes related to the question. The question and the retrieved note text are input into the knowledge LLM, generating sentences containing relevant knowledge.

A version that uses course material information as a knowledge source for RAG is also possible, as illustrated by the dotted-line flow in Fig. 1b. In this study, scripts of lecture videos viewed by learners were used as the extra database. Since learners watch lecture videos, their notes may already include lecture content. However, some learners may take notes in bullet points or only write down keywords. Therefore, after retrieving learner notes, each note's text is used as queries to retrieve relevant lecture video scripts using the same procedure. The lecture video scripts are also chunked in advance and stored in a vector store.

The knowledge obtained through the above steps is combined with the question and input into an LLM (Response LLM) that simulates the learner's response. When using a well-trained LLM like GPT-4o as Response LLM, it may provide answers based on its pre-existing knowledge rather than relying on RAG, making it challenging to reflect the learner's understanding accurately [6]. To address this, we restrict LLM's pre-existing knowledge. By the restriction, the LLM cannot correctly answer questions about concepts the learner has not yet studied.

To specifically restrict the LLM's knowledge, either a persona-based approach or a fine-tuning-based approach is used. As the persona-based approach, Inspired by AlgoBo [6], prompts are designed under the assumption that the LLM lacks knowledge of the key topics covered in the lecture. Specifically, phrases such as "You do not know #topic" are embedded in the prompt. In the fine-tuning-based approach, we use a modified version of the detuning method [12]. LLM is retrained via incorrect sentences about learning materials in the detuning method. The training data is generated automatically by a LLM. In this study, we use two training strategies: (1) next token prediction task using incorrect sentences and (2) input: MCQ + correct/incorrect explanation, output: the correct/incorrect choice.

4 Experiment

This study recruited participants to take an on-demand lecture. The participants viewed slide-based lecture materials and corresponding lecture videos. They were instructed to write their understanding in their notes during this process. As a prompt for note-taking, a worksheet was provided, incorporating questions inspired by Bloom's taxonomy [10], such as "Try explaining what you have learned". Participants were required to bring their notes when taking a comprehension test.

Collected Data. The participants were Kyushu university students and were paid for participation. The topic of the on-demand lecture was an Introduction to Generative AI, covering four lecture materials, including the mechanisms of generative AI, deep learning, and legal regulations. The four lecture videos were edited to be approximately 20 min each, with transcripts of recorded lectures converted into machine-generated voice. The test consisted of 29 multiple-choice questions (MCQs). This study analyzed only machine-readable notes (Word, Excel, PDF), excluding handwritten notes. Finally, the analysis was conducted using the notes and test answer data of 31 participants.

Models. Knowledge restriction was applied to GPT-4o (gpt-4o-2024-08-06) using either the Persona-based or Fine-tuning methods.[1] For fine-tuning, the dataset contained 565,132 characters of generated reasoning texts. These were trained using OpenAI's fine-tuning platform. For text embedding in the RAG, OpenAI's text-embedding-3-large was used. The chunking process ensured that context was not interrupted by splitting text into approximately 400-character segments. The top 3 most relevant passages were retrieved for each RAG search. In this experiment, the Response LLM used the knowledge-restricted LLM, while the Knowledge LLM was GPT-4o.

Evaluation Setting. A task was conducted to predict whether learners would answer test questions correctly. The test questions were input into the LLM, which predicted the learners' answers. If a learner's answer was correct or incorrect and the prediction matched this outcome, it was considered a True Positive or True Negative, respectively. The dataset contained 627 positive (correct) and 272 negative (incorrect) answers, making it highly imbalanced. For the ablation study, different conditions were compared to evaluate prediction accuracy.

Experimental Results. The accuracy rate of GPT-4o without knowledge constraints was 82.8%. When using fine-tuning as a knowledge restriction method, the accuracy dropped to 0%. The persona-based approach resulted in an accuracy of 86.2%, failing to lower the accuracy rate. The results of the persona-based approach did not match those of AlgoBo [6]. This discrepancy is likely due to differences in the granularity of instructions regarding missing knowledge provided by the persona. In this study, the LLM's knowledge was restricted based on the overalls of lecture materials, whereas AlgoBo used highly specific topic designations (e.g., binary search). We observed that the model could not select the correct answer when more pinpointed knowledge restrictions were applied to each problem using a persona. This finding indicates that if specific topic constraints can be set, knowledge restrictions can also be applied using a persona-based approach. However, further discussion is needed regarding the development of topic-specific estimation methods and the impact on learner performance predictions. Additionally, when we inserted the necessary knowledge for answering the MCQ into the fine-tuned LLM, the accuracy rate improved to 86.2%. This result suggests that if the correct knowledge can be inserted, similar to a teachable agent, the fine-tuned LLM has the potential to answer correctly.

Table 1 shows the results of the ablation study. The persona-based approach was an overestimation of correct answers. Compared to fine-tuning, the recall was higher, but the FPR was also high. Using note-driven RAG improved the macro F-measure by approximately 0.03 to 0.05. However, since performance

[1] Persona prompt. Your persona: You are a university student. You have the following problems: - You do not know how generative AI works. - You do not have knowledge of deep learning and machine learning. - You do not know the legal regulations related to generative AI. You might give incorrect answers or fail to explain things properly. You will respond according to this persona.

Table 1. Performance comparison of Persona and Finetuning approaches.As evaluation metrics, precision (P), recall (R), false positive rate (FPR), false negative rate (FNR), and macro F-measure (FM) were used.

		Note	Script	KLLM	P ↑	R ↑	FPR ↓	FNR ↓	Macro FM↑
Persona	1				0.702	0.868	0.849	0.132	0.492
	2	✓			0.718	0.845	0.765	0.155	0.536
	3	✓	✓		0.711	0.836	0.783	0.164	0.520
	4	✓		✓	0.719	0.850	0.765	0.150	0.538
	5	✓	✓	✓	0.721	0.850	0.757	0.150	0.543
Finetuning	1				0.000	0.000	0.000	1.000	0.232
	2	✓			0.771	0.118	0.081	0.882	0.335
	3	✓	✓		0.743	0.166	0.132	0.834	0.364
	4	✓		✓	0.792	0.402	0.243	0.598	0.508
	5	✓	✓	✓	0.752	0.381	0.290	0.619	0.479

degradation could not be achieved with the persona-based approach, reflecting the learner's understanding status was challenging even if we used notes. In fine-tuning, precision was similar to that of the persona-based approach, but the recall was lower, FPR was lower, and FNR was higher. This result suggests that while fine-tuning makes detecting a lack of understanding easier, it struggles to grasp knowledge that the learner has already acquired comprehensively. When comparing Fine-tuning 1 with Fine-tuning 2–3, the improvement in performance due to the learner's notes was evident. Furthermore, comparing Fine-tuning 2–3 with Fine-tuning 4–5, the introduction of Knowledge LLM allowed additional filtering of sentences related to the questions from the retrieved note and script texts, making it easier for the fine-tuned LLM to answer correctly.

5 Conclusion

We proposed note-driven RAG to estimate their performance using a LLMs. The answers to RQs are summarized as follows. **RQ1:Ans.** When knowledge restriction was adequately applied, the effectiveness of utilizing learners' notes was significantly observed (Finetuning 1 v.s. 2, 3, 4, 5). **RQ2:Ans.** In this study, persona-based knowledge restriction was challenging when restricting various knowledge. The restricted LLM overestimated correct answers. From the perspective of learning support, this finding suggests that the system might fail to detect learners' lack of understanding. On the other hand, knowledge restriction through finetuning proved to be effective, exhibiting a higher sensitivity to detecting learners' lack of understanding. This approach could be beneficial as an initial state for interactive systems like ITS. The use of scripts did not lead to significant performance improvements (Finetuning 3 v.s. 5). Furthermore, while directly inputting learners' notes led to some performance improvement, summarizing the notes resulted in even greater accuracy gains (Finetuning 2 v.s. 4).

This suggests that extracting and structuring information from notes and other knowledge sources is crucial. This finding aligns with general prompt design principles [17].

The limitations of our study are summarized as follows. We used learners' notes as the primary data source for estimating their knowledge. However, notes alone may not capture learners' comprehensive understanding. Moreover, this study focused only on multiple-choice questions for evaluation. Future research should extend the evaluation to include open-ended questions and computational problems to assess the robustness of our approach.

Acknowledgments. This work was supported by JST, PRESTO Grant Number JPMJPR236A, Japan.

References

1. Armbruster, B.B.: Taking notes from lectures (2008)
2. Brown, T., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
3. Cohn, E., Cohn, S., Bradley, J.: Notetaking, working memory, and learning in principles of economics. J. Econ. Educ. **26**(4), 291–307 (1995). http://www.jstor.org/stable/1182993
4. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User Model. User-Adap. Inter. **4**, 253–278 (1994)
5. Graesser, A., Lu, S., Mitchell, H., Ventura, M., Olney, A., Louwerse, M.: Autotutor: a tutor with dialogue in natural language. Behav. Res. Methods Instr. Comput. **36**, 180–192 (2004). https://doi.org/10.3758/BF03195563
6. Jin, H., Lee, S., Shin, H., Kim, J.: Teach ai how to code: using large language models as teachable agents for programming education. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI '24. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3613904.3642349
7. Jordan, P.W., Hall, B., Ringenberg, M., Cue, Y., Rosé, C.: Tools for authoring a dialogue agent that participates in learning studies. In: Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, p. 43–50. IOS Press, NLD (2007)
8. Leelawong, K., Biswas, G.: Designing learning by teaching agents: the betty's brain system. Int. J. Artif. Intell. Ed. **18**(3), 181–208 (2008)
9. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)
10. LW, A., et al.: A Taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives (2001)
11. Mansour, W.A., Albatarni, S., Eltanbouly, S., Elsayed, T.: Can large language models automatically score proficiency of written essays? In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 2777–2786. ELRA and ICCL, Torino (2024). https://aclanthology.org/2024.lrec-main.247/

12. Minematsu, T., Shimada, A.: Large language model detuning in learning content understanding. In: 21st International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA 2024, pp. 11–18. IADIS Press (2024)
13. Mollick, E., Mollick, L.: Assigning ai: seven approaches for students, with prompts. ArXiv arxiv:2306.10052 (2023). https://api.semanticscholar.org/CorpusID:259203456
14. Paladines, J., Ramirez, J.: A systematic literature review of intelligent tutoring systems with dialogue in natural language. IEEE Access **8**, 164246–164267 (2020). https://doi.org/10.1109/ACCESS.2020.3021383
15. Piech, C., et al.: Deep knowledge tracing. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc. (2015)
16. Scarlatos, A., Baker, R.S., Lan, A.: Exploring knowledge tracing in tutor-student dialogues using llms. In: The 15th International Learning Analytics and Knowledge Conference (2025)
17. Sondos Mahmoud Bsharat, Aidar Myrzakhan, Z.S.: Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. arXiv preprint arXiv:2312.16171 (2023)
18. Tsutsumi, E., Kinoshita, R., Ueno, M.: Deep-irt with independent student and item networks. In: Educational Data Mining (2021). https://api.semanticscholar.org/CorpusID:235672622



Exploring Student Identity in Adaptive Learning Systems Through Qualitative Data

Clara Belitz[1(✉)], HaeJin Lee[1], Nidhi Nasiar[2], Stephen E. Fancsali[3], Frank Stinar[1], Husni Almoubayyed[3], Steve Ritter[3], Ryan S. Baker[2], Jaclyn Ocumpaugh[2], and Nigel Bosch[1]

[1] University of Illinois Urbana–Champaign, Champaign, IL 61820, USA
cbelitz2@illinois.edu
[2] University of Pennsylvania, Philadelphia, PA 19104, USA
[3] Carnegie Learning, Inc., Pittsburgh, PA 15219, USA

Abstract. Adaptive learning systems are increasingly common in U.S. classrooms, but it is not yet clear whether their positive impacts are realized equally across all students. This study explores whether nuanced identity categories from open-ended self-reported data are associated with outcomes in an adaptive learning system for secondary mathematics. As a measure of impact of these social identity data, we correlate student responses for 3 categories: race and ethnicity, gender, and learning identity—a category combining student status and orientation toward learning—and total lessons completed in an adaptive learning system over one academic year. Results show the value of emergent and novel identity categories when measuring student outcomes, as learning identity was positively correlated with mathematics outcomes across two statistical tests.

Keywords: Student Identity · Equity · Adaptive Learning Systems

1 Introduction

AI-driven technologies are increasingly used to improve student learning outcomes and support teachers [14,22], and researchers have aimed to ensure that adaptive learning systems benefit all types of students equitably [9,28]. Choosing which inherent or acquired characteristics to compare when measuring these outcomes, however, is often glossed over in algorithmic fairness methods [1]. This paper explores those choices, examining a new way to define the groups against which equity and algorithmic bias are measured in educational settings.

Many current questions around equity in AI-driven systems utilize census-style categories, often from preexisting sources [18]. Such data offer practical ways to examine differences at scale, but can lack both nuance and breadth, obscuring some variables that may be needed to ensure equity [27]. For example, in classroom contexts, students' intellectual identities could significantly affect

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 356–363, 2025.
https://doi.org/10.1007/978-3-031-98462-4_45

how they approach a subject, as has been shown with math identity [8,23]. Both missing data (e.g., learning identities) and underspecified data (e.g., oversimplified race and gender categories) can make it harder to understand how best to support learning for all students.

To capture students' identities and characteristics in a more nuanced and accurate approach, this work asks the following questions:

- **RQ1:** How does a free-response measure of student identity compare to school-provided demographic information in both occurrence and content?
- **RQ2:** Do free-response measures of student identity capture predictive information for learning that is not otherwise found in school-provided demographic data?

This paper presents identity data from students using a validated measure, collecting qualitative (descriptive and free-form) responses from middle and high school students across two academic years (2021–2023). The presence of gender and race responses, commonly used demographic categories, was not correlated with total completed lessons in an AI-driven adaptive math system, while the presence of a novel category that emerged from the data, *learning identity*, had significant positive correlation across two tests.

2 Related Work

Educational research has long recognized that complex sociocultural factors influence whether student opportunities and outcomes are fair and equitable [25]. Education researchers studying artificial intelligence, data science, and learning analytics have examined ways to support marginalized students [24] and whether commonly used algorithms exhibit signs of bias [2]. They have also proposed frameworks for ensuring fairness and equity [3,15]. Similarly, researchers have investigated identities that may not be legal or protected classifications (e.g., math identity) but are nonetheless tied to power dynamics [11,21].

Although the complexity of ensuring equitable learning opportunities in AI-driven learning systems is well known, access to data that would support deep analysis is not always possible, if these data exist at all [16]. Categorization and classification fundamentally require making choices about which information is relevant and important [5]. Making choices is not, in and of itself, problematic, since categorization is crucial to retrieving and using information, but these processes and decisions are often glossed over in artificial intelligence pipelines [4, 19]. Categories like race and gender are frequently characterized as inflexible and superficial, though they are socially constructed and nuanced [13,17]. Surveys have tried to develop more flexible categorizations, but may still offer incomplete classifications [13].

3 Method

This project adapts the Twenty Statements Test [12,20], which uses a constructivist approach to developing categories, rather than choosing features a priori

to represent the classes of interest. In this paper, a subset of representative categories is described in depth, while the entire coding schema is available online.

3.1 Student Demographics and Learning Context

Our study examines data from middle and high school students in a small city in the northeastern United States. Students used MATHia as part of their regular instruction in 6th to 8th grade mathematics, geometry, and algebra. As an AI-driven adaptive system, MATHia provides personalized hints and just-in-time feedback as learners progress [22]. Given this capability, it is important to understand whether these lessons, hints, and feedback are working equally effectively for all students. A typical school year encompasses 90–120 lessons, depending on standards, individual teacher and administrator preferences, and other factors impacting customization. Lesson count is a more effective measure of student outcomes than total time spent with MATHia [10]. The mean lessons completed per student in our study was 104.7 (SD = 44.5). Lesson count completion per student had a non–normal distribution. The school district, which collects demographic data as part of its standard practice, shared these data under the guidance of the University of Illinois Institutional Review Board #IRB24-1190.

3.2 Twenty Statements Test

The Twenty Statements Test (TST) uses a free-form answer format to elicit self-concept by having individuals answer the question "Who am I?" up to 20 different ways in a short amount of time [20]. Responses are then coded by researchers both into well-established categories, such as the schema developed by Gordon and Gergen [12], and into categories that arise inductively via thematic analysis [6]. The survey design permits students to respond with multiple answers that fall into the same category (e.g., multiple nationalities). The survey was delivered online, where each response had its own text box. The survey was available in the four languages most commonly spoken at home by students in the school district. The survey was piloted with data from 53 students, with 19.7 responses on average per student (SD=1.4) in the 2020–2021 academic year. Full instructions regarding the development of the schema, interrater reliability, and inclusion criteria are available online. The rest of this paper presents results from students in the 2021–2023 academic years (n=118). Note that the responses reported in this paper are not verbatim, to respect student privacy, but are emblematic of the type of responses seen.

4 Results

Overall, student-provided demographic categories aligned with those provided by the school. A few students provided responses that actively disagreed with their demographic data, including students with ethnic identities from countries outside of the U.S. that did not map neatly to American racial boxes (n = 5).

Table 1. Student responses (rows) vs. school-provided categories (columns). The category of ≥ 1 *Race* $+ \geq 1$ *Ethnicity* refers to students who provided at least one racial identity as well as at least one ethnic identity. The categories of *Multiple Races* and *Multiple Ethnicities* refer to those who provided at least two racial identities but no ethnic identities, and at least two ethnic identities but no racial identities, respectively.

	African-American	Asian	Hispanic	White	Multi-Race, Non-Hispanic
One Race	6	1	2	1	0
Multiple Races	0	0	0	0	1
≥ 1 Race $+ \geq 1$ Ethnicity	9	0	0	1	1
One Ethnicity	9	0	1	4	1
Multiple Ethnicities	1	1	2	1	0
Ethnicity + American	1	1	0	0	0
Person of Color + Ethnicity	1	0	0	0	0
No Ethnic or Racial Response	43	0	7	22	2
Total (2021–2023)	70	3	12	29	5

However, most students did not relay this information at all, or—when they did—provided a more nuanced or specific category than the one provided by the school. Importantly, more students provided responses related to learning identity than related to gender or racial categories, which are commonly used demographic features. As such, learning identity was chosen for further examination of the importance of novel, emergent categories. Specifically, we compare student performances through lesson completion (a measure that indicates how far through the MATHia cuirriculum a student has progressed) across these categories.

RQ1: How does a free-response measure of student identity compare to school-provided demographic information in both occurrence and content?

In total, 118 students provided valid responses across the 2021–2022 ($n = 62, \mu = 19.8, SD = 0.81$), and 2022–2023 ($n = 56, \mu = 19.4, SD = 2.6$) academic years. Race and ethnicity data are combined here to reflect the school-provided demographic data, even as we recognize that they are distinct dimensions. Overall, 45 out of 118 students (38.1%) indicated a racial and/or ethnic identity. The students who did provide racial and ethnic responses were classified using an emergent schema of Middle Eastern or North African, Multiracial, East Asian, South Asian, Pacific Islander, Latino/a, White, and Other or Unclear. Though the school uses the group "African-American," the term "African" only appeared in this set of responses to explicitly identify a student as from a specific African country or to differentiate a specific relationship to place (e.g., "from North Africa"). Students instead broadly referred to themselves as Black or from specific locations. Students in other racial groups represented in this dataset exhibited a similar pattern, indicating specific ethnic origins like Portuguese, Italian, or Irish for White students and Japanese, Filipino, or Chinese for Asian students (see Table 1, which uses school-provided demographic language).

Table 2. Spearman's *rho* and Mann–Whitney *U* test comparing identity (presence/absence) to lessons completed. **Bold** indicates significance.

Category	*rho*	*p*	*U*	*p*
Gender	–.029	.753	1748.5	.753
Race/Ethnicity	–.034	.716	1708.5	.717
Learning Identity	**.227**	**.013**	**970.5**	**.014**

Just over half of respondents (65 of 118; 55.1%) provided gender responses. Nearly all of these (n=64) aligned with the school demographics (40 girls, 24 boys), but 1 nonbinary student did not. Gendered responses ranged from implicit (e.g., familial roles like son; pronouns like she/her) to explicit statements (e.g., "I'm a boy").

Learning identity emerged from two categories (i.e., intellectual concerns and student role), which have both appeared in prior work that investigates the effects of (a) performance beliefs and (b) interest in mathematics identity development [7]. Intellectual concerns emerged at roughly the same rate as gendered responses (62 of 118; 52.3%), while student role appeared at higher rates (79 of 118; 66.9%). Intellectual role was most likely to be positive (75.8% positive; 6.1% negative; 18.2% neutral), while student role (which often emerged as "I am a student") was most likely to be neutral (53.6% neutral; 41.9% positive; 4.5% negative). Notably, negatively valenced responses for student role and intellectual concerns were always accompanied by at least one positive valence response. Such responses for these two categories of student role and intellectual concerns frequently overlapped, with negligible qualitative difference. For example, one student reported both "I am an honors student," coded as student role for the explicit inclusion of school, as well as "I am good at math," coded as intellectual concerns for the relationship to self-conception. Therefore these two forms of response are treated as a single category in our analysis.

RQ2: Do free-response measures of student identity capture predictive information for learning that is not otherwise found in school-provided demographic data?

Because **RQ2** explores whether the presence of an identity is significantly related to MATHia outcomes, specific identity responses were binarized as *present* or *not present*. We performed two statistical tests: the correlation between total lessons completed in MATHia and the presence of each of the categories of interest as well as the difference in completion count across presence and absence of those categories (Table 2). The presence of gender and race responses had no significant correlation with total lessons completed in MATHia across the school year. In contrast, learning identity did ($rho = .227$, $p = .013$; $U = 970.5$, $p = .014$).

5 Discussion and Conclusion

Our results demonstrate the importance of emergent categories, like learning identity, when exploring the impact of adaptive learning systems. Similarly, we show (1) how students' TST responses differ from the demographic data their schools use to describe them and (2) how students' relationship to the schools emerges in the data. Our results show more students mentioned their learning identity than did their gender. Notably, many students grounded their learning identities in a highly specific fashion (e.g., specific subjects in which they excel or struggle). Such nuances are not captured by typical school demographic categories, but are likely meaningful to interest in learning and subsequent educational outcomes. As such, finding ways to expand beyond traditional demographics, perhaps by working to elicit a broader range of learning identity categories, may help researchers to better understand whether student success in adaptive learning systems is occurring in an equitable fashion.

Learning identities also emerged more often than racial or ethnic ones in both school years, alluding to the significance of this designation to students. Although the context of completing the TST while at school may have contributed to the higher rates for these categories, it is also likely that these are important to students' everyday lived experiences with adaptive learning systems. These results may raise questions about the relevance of the social identities by which (in)equity is measured for individuals in this specific context. We hypothesize that, as with math identity, learning identity may be an important factor for student performance [23,26]. Given that our results demonstrate the positive impact of learning identity on outcomes, future designs of AI-driven educational systems might consider how they can build up students' learning identities and math confidence. While this work has not yet explored all the possible emergent categories from the data, nor the intersections of multiple categories, this reflects the rich nature of qualitative work and will be pursued in future research. Though this type of qualitative work is time-intensive, the varied data may help to avoid missing differentiations that matter for detecting nuanced algorithmic bias in adaptive learning systems.

Though the TST takes more time and resources than using extant demographics, these ground-up approaches provide greater detail for understanding the contextual realities of algorithmic systems and their sociotechnical impacts on classrooms. As demonstrated by our results, a broader range of more nuanced categories is an important part of understanding how algorithmic bias emerges in AI systems.

Acknowledgments. This research was supported by NSF grant no. 2000638. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Abdu, A.A., Pasquetto, I.V., Jacobs, A.Z.: An empirical analysis of racial categories in the algorithmic fairness literature. In: 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, pp. 1324–1333. Association for Computing Machinery, Chicago (2023). https://doi.org/10.1145/3593013.3594083
2. Ashwin, T.S., Biswas, G.: Identifying and mitigating algorithmic bias in student emotional analysis. In: Artificial Intelligence in Education: 25th International Conference, AIED 2024, pp. 89–103. Springer, Cham (2024), https://doi.org/10.1007/978-3-031-64299-9_7
3. Belitz, C., et al.: Hierarchical dependencies in classroom settings influence algorithmic bias metrics. In: Proceedings of the 14th Learning Analytics & Knowledge Conference, pp. 210–218. ACM, Kyoto (2024). https://doi.org/10.1145/3636555.3636869
4. Belitz, C., Ocumpaugh, J., Ritter, S., Baker, R.S., Fancsali, S.E., Bosch, N.: Constructing categories: moving beyond protected classes in algorithmic fairness. J. Am. Soc. Inf. Sci. **74**(6), 663–668 (2022). https://doi.org/10.1002/asi.24643
5. Bowker, G.C., Star, S.L.: Sorting Things Out: Classification and Its Consequences. MIT Press, Cambridge (1999)
6. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa
7. Cribbs, J.D., Hazari, Z., Sonnert, G., Sadler, P.M.: Establishing an explanatory model for mathematics identity. Child Dev. **86**(4), 1048–1062 (2015). https://doi.org/10.1111/cdev.12363
8. Crossley, S., Liu, R., McNamara, D.: Predicting math performance using natural language processing tools. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, LAK '17, pp. 339–347. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3027385.3027399
9. Doroudi, S., Brunskill, E.: Fairer but not fair enough: on the equitability of knowledge tracing. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge, pp. 335–339. ACM, Tempe (2019). https://doi.org/10.1145/3303772.3303838
10. Fancsali, S.E., Ritter, S., Yudelson, M., Sandbothe, M., Berman, S.R.: Implementation factors and outcomes for intelligent tutoring systems: a case study of time and efficiency with Cognitive Tutor Algebra. Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference, pp. 473–478 (2016). https://cdn.aaai.org/ocs/12965/12965-57684-1-PB.pdf
11. Flores, A.: Examining disparities in mathematics education: achievement gap or opportunity gap? High Sch. J. **91**(1), 29–42 (2007). https://doi.org/10.1353/hsj.2007.0022
12. Gordon, C., Gergen, K.J.: Self-conceptions: Configurations of content. In: The Self in Social Interaction. John Wiley & Sons, Inc., New York (1968)
13. Guyan, K.: Queer Data: Using Gender, Sex and Sexuality Data for Action. Bloomsbury Academic, London (2022)
14. Heffernan, N.T., Heffernan, C.L.: The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. Int. J. Artif. Intell. Educ. **24**(4), 470–497 (2014). https://doi.org/10.1007/s40593-014-0024-x

15. Holstein, K., Doroudi, S.: Fairness and equity in learning analytics systems (FairLAK). In: Companion Proceedings of the Ninth International Learning Analytics & Knowledge Conference, p. 2. Society for Learning Analytics Research, Arizona (2019)
16. Jo, E.S., Gebru, T.: Lessons from archives: strategies for collecting sociocultural data in machine learning. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, pp. 306–316. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3351095.3372829
17. Kayyali, R.: US census classifications and Arab Americans: contestations and definitions of identity markers. J. Ethn. Migr. Stud. **39**(8), 1299–1318 (2013). https://doi.org/10.1080/1369183X.2013.778150
18. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. WIREs Data Min. Knowl. Disc. **12**(3), e1452 (2022). https://doi.org/10.1002/widm.1452
19. Lu, C., Kay, J., McKee, K.R.: Subverting machines, fluctuating identities: relearning human categorization. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, pp. 1005–1015. ACM, Seoul (2022). https://doi.org/10.1145/3531146.3533161
20. Kuhn, M., McPartland, T.: An empirical investigation of self-attitudes. Am. Sociol. Rev. **19**, 68–76 (1954)
21. Osborne, J.W.: Testing stereotype threat: does anxiety explain race and sex differences in achievement? Contemp. Educ. Psychol. **26**(3), 291–310 (2001). https://doi.org/10.1006/ceps.2000.1052
22. Ritter, S., Fancsali, S.E.: MATHia X: the next generation cognitive Tutor. In: Proceedings of the EDM 2016 Workshops and Tutorials, pp. 624–625. International Educational Data Mining Society, Raleigh (2016)
23. Slater, S., Ocumpaugh, J., Baker, R., Li, J., Labrum, M.: Identifying changes in math identity through adaptive learning systems use. In: Proceedings of the 26th International Conference on Computers in Education, p. 7. Asia-Pacific Society for Computers in Education, Philippines (2018)
24. Sloan-Lynch, J., Morse, R.: Equity-forward learning analytics: designing a dashboard to support marginalized student success. In: Proceedings of the 14th Learning Analytics & Knowledge Conference, LAK '24, pp. 1–11. Association for Computing Machinery, Kyoto (2024). https://doi.org/10.1145/3636555.3636844
25. Stobart, G.: Fairness in multicultural assessment systems. Assess. Educ. Principles Policy Pract. **12**(3), 275–287 (2005). https://doi.org/10.1080/09695940500337249
26. Tywoniw, R., Crossley, S.A., Ocumpaugh, J., Karumbaiah, S., Baker, R.: Relationships between math performance and human judgments of motivational constructs in an online math tutoring system. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12164, pp. 329–333. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52240-7_60
27. Westbrook, L., Saperstein, A.: New categories are not enough: rethinking the measurement of sex and gender in social surveys. Gend. Soc. **29**(4), 534–560 (2015). https://doi.org/10.1177/0891243215584758
28. Wise, A.F.: Designing pedagogical interventions to support student use of learning analytics. In: Proceedings of the Fourth International Conference on Learning Analytics & Knowledge, LAK '14, pp. 203–211. Association for Computing Machinery, Indianapolis (2014). https://doi.org/10.1145/2567574.2567588



On the Fairness of Ensemble Learning Methods in Student Dropout Prediction

Abdelghafour Aboukacem(✉), Loubna Mekouar, El Houcine Bergou, Youssef Iraqi, and Ismail Berrada

College of Computing, Mohammed VI Polytechnic University, Ben Guerir, Morocco
{abdelghafour.aboukacem,loubna.mekouar,elhoucine.bergou,youssef.iraqi,ismail.berrada}@um6p.ma

Abstract. Student Dropout Prediction (SDP) is a crucial task in Educational Data Mining (EDM). It aims to help institutions intervene early to improve retention. Despite notable advancements in the predictive performance of the developed models, the topic of fairness is usually overlooked. Algorithmic bias can lead to disparate impacts across demographic groups. In this study, we explore the use of ensemble learning, specifically bagging, boosting, voting, and stacking techniques to improve both the predictive performance as well as the fairness aspect. Using a real-world dataset that spans the whole K-12 system of Morocco with 14 sub-datasets, we evaluate various Machine Learning (ML) models trained using the previously mentioned ensemble learning techniques. Our study examines model fairness for several protected attributes, including gender, handicap, financial aid, and boarding school availability.

Keywords: Student dropout prediction · Fairness · Ensemble learning

1 Introduction

With the recent advancements in learning technologies [24], the education field has been undergoing substantial improvements in learning quality. Leveraging Educational Data Mining (EDM), data mining techniques are being applied to extract meaningful insight to further improve key objectives[17]. These advancements span from personalized and adaptive learning [25], to curriculum optimization and policy decision improvement [7]. However, a growing interest is being dedicated toward Student Dropout Prediction (SDP) [1,20], driven by the severity of the issue [5,6]. To this extent, Dropout Early Warning Systems are being developed to predict the possibility of students dropping out. These systems are based on Machine Learning (ML) Deep Learning (DL) models and are based on datasets that span the academic, socio-economic, and demographic aspects of the students. A variety of models, such as eXtreme Gradient Boosting (XGB), Light Generalized Linear Model (LGBM), are being used to predict students at risk [8]. Studies also explored even more complex models such as Large Language Models (LLMs) to predict and explain student dropouts [2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 364–371, 2025.
https://doi.org/10.1007/978-3-031-98462-4_46

However, despite their predictive performance, the topic of fairness, transparency, and ethical accountability in SDP remains underexplored. Biased dropout prediction models risk reinforcing systemic inequalities, particularly for marginalized and under-resourced students. The fairness of AI in education is not just a technical issue but an ethical imperative, influencing student agency, teacher interventions, and institutional decision-making.

On the other hand, ensemble learning [16] has been known to improve the performance of simple machine learning models. This method aggregates several weaker models to create a more performant model. The aggregation can be done is a variety of ways. The most common ones are bagging [4], boosting [9], stacking [23], and voting [22]. However, the question of the fairness of the ensembled classification models is still unexplored, especially in the case of SDP.

In our work, we aim to explore the use of ensemble learning techniques, specifically bagging, boosting, stacking, and voting, in improving both the predictive performance as well as minimizing the disparities for several protected groups. Our experiment is based on a real life dataset that spans the whole K-12 system of Morocco. Currently over three hundred thousand students dropout each year in the Moroccan K-12 system[1], this calls for accurate and fair predictive methods to mitigate the issue. To summarize, the main contributions of this paper are as follows:

- **First exploration of ensemble learning for fairness**: To the best of our knowledge, this is the first study that investigates the effects of ensemble learning on model fairness for student dropout.
- **Multiple protected groups**: We consider four protected groups in our study based on gender, mental and physical handicaps, availability of boarding school, and financial aid. This diversity in protected groups is rare among prior research.
- **Evaluations on real-life dataset**: We compare the predictive performance as well as the fairness of ensemble learning methods to those of ML models. We provide a comprehensive comparison through real-life dataset of the Moroccan educational K-12 system.

The rest of this paper is structured as follows, Sect. 2 showcases previous studies on fairness for educational settings, Sect. 3 provides details on our experiment detailing important information about the dataset used and performance metrics. Section 4 presents our findings comparing the ensemble learning methods to the base ML models in terms of predictive performance and fairness. Finally, Sect. 5 provides a comprehensive conclusion and future work opportunities.

2 Related Work

Prior research has explored the topic of fairness in educational machine learning models through a variety of metrics and bias mitigation strategies. Several works

[1] https://www.moroccoworldnews.com/2023/06/32538/over-300-000-students-drop-out-of-school-each-year-in-morocco/.

studies the fairness of these models in broader areas than dropout prediction such as in language assessment [13], success prediction [3,12,27], and form post classification [19] based on gender.

For SDP, studies have measured fairness between groups based on features such as gender and race [10,11]. A variety of metrics are being used to measure the fairness of these models which include Absolute Between-ROC Area (ABROCA) first introduced in [10]. This metric computes the absolute difference between the Area Under the ROC Curve (AUC) of each group for a protected attribute. Moreover, the authors in [11] base their evaluations on discrimination, which measures the difference in model performance across demographic groups, and consistency, which evaluates how similar predictions are for similar individuals across groups. The work in [12] utilizes demographic parity, which ensures that predictions are independent of sensitive attributes, and equality of opportunity which in turn ensures equal true positive rates across demographic groups. Other studies rely on classification performance metrics such as accuracy, recall, True Negative Rate (TNR) [26], and AUC [19].

To mitigate these disparities, studies used data resampling strategies [18,19] which aim to train predictive models on a balanced set based on the studied sensitive feature. Adjusting the training set's sample weights is another idea previously explored in [26].

In educational contexts, ensemble machine learning was explored to overcome a variety of issues. For example, ensembled K-Nearest Neighbors (KNN), Neural Network (NN), and other models were used in [14] for curriculum recommendation for higher education. Ensemble learning approaches are also used to perform cheat detection [28]. Moreover, combining models such as XGB, LGBM, Random Forest (RF), and more to predict student failure along with suggesting learning paths is explored in [21].

3 Experiment Setting

3.1 Dataset

The dataset is an anonymized collection from Morocco's Ministry of National Education, Preschool, and Sports[2], it contains academic, demographic, and socio-economic features. Notably, the dataset pertains specifically to public-sector institutions within the Fes-Meknes region. Furthermore, it covers student records from the 2015/2016 academic year to the 2020/2021 academic year of 14 different educational levels and study options. We split our dataset into train and test sets by considering records from the 2015/2016 to 2019/2020 school years as the training set, and the records from the 2020/2021 school year as the test set.

In our fairness analysis we consider four different sensitive features which we find in our dataset: gender, handicaps, boarding school availability, and financial aid. It is worth noting that in some educational levels, we do not find records

[2] https://www.men.gov.ma/.

of particular groups. This in turn, renders the usage of resampling techniques useless.

3.2 Experiment

To evaluate the fairness of ensembled models, we consider cases of the most common ensemble learning techniques. For Bagging, we evaluate bagging of Decision Trees (DT) classifier, considering lower max depth trees (depth of 2) and a high number of estimators (100 estimators) as suggested in the literature [15]. For Boosting, we evaluate the effectiveness of boosted DT, employing Adaptive Boosting (AB) as a booster model with 100 estimators. For Stacking, we evaluate stacking DT,Logistic Regression (LR), and Naïve Bayes (NB) models, using LR as the meta-learner. Finally, for Voting, we consider both hard and soft voting using the three base models: DT, LR, and NB. We compare the four methods to the base models by considering AUC, class-specific precision, and recall for model performance. Additionally, we measure the fairness of the models using the ABROCA metric, defined as the absolute difference between the AUCs of each group, as well as the Demographic Parity (DP) between each group pair.

4 Results and Discussions

Table 1 showcases the results of the predictive performance of the tested approaches. Table 2 provides details on model fairness measurements using the ABROCA metric.

The results from Table 1 indicate that stacking the predictive models as well as the base LR model performed the best, achieving the highest AUC of 0.80. These models were followed closely by the boosted and bagged trees with an AUC score of 0.79. These models provide the best balanced performance in identifying at risk students. In terms of dropout precision, denoted as D precision, boosted trees outperformed the other models with a precision 0.35, indicating that when the model predicts a student as a dropout, it is likely correct. On the other hand, hard voting achieved the highest score on non-dropout precision, denoted as ND precision, making it more reliable for students who will persist. As far as recall goes, soft voting and NB both achieved the highest score in detecting actual dropout cases with 0.81 and 0.83 dropout recalls respectively. However, boosted trees achieved the highest score, of 0.79, for non-dropout recall, denoted ND recall, making it effective at correctly identifying those who are likely to persist.

Overall, stacking and LR emerged as the best-performing models based on AUC, while boosted trees excelled in dropout precision and non-dropout recall. For dropout identification, soft voting provided the best rounded performance while providing an edge in dropout detection with the higher dropout recall value.

After examining the disparities for all four sensitive features using the ABROCA metric (lower values are better). From Table 2, the results can be summarized in the following:

Table 1. Performance comparison of different methods.

Method	AUC	D Precision	ND Precision	D Recall	ND Recall
Soft Voting	0.77	0.23	0.95	**0.81**	0.56
Hard Voting	0.73	0.27	**0.95**	0.79	0.66
Stacking	**0.80**	0.29	0.94	0.72	0.73
Boosting	0.79	**0.35**	0.91	0.61	**0.79**
Bagged Trees	0.79	0.32	0.94	0.76	0.69
DT	0.76	0.33	**0.93**	0.74	0.68
LR	**0.80**	**0.33**	0.93	0.71	**0.73**
NB	0.54	0.18	0.88	**0.83**	0.25

- **Boarding School**: Boosted trees and LR achieved the best results of an ABROCA average of 0.0190 and 0.0208 respectively, exhibiting the lowest bias and indicating more equitable predictions for students from different schooling backgrounds.
- **Gender**: NB was the fairest model in terms of gender with an averaged ABROCA of 0.0182 and a DP score of 0.0232, suggesting minimal prediction disparities across male and female students.
- **Handicap**: Bagged trees along with LR achieved the lowest bias with averaged ABROCA scores of 0.1019 and 0.1024 respectively. Bagged trees however, exhibit slightly better performance in DP. This indicates their fairer predictive ability for students with disability.
- **Financial Aid**: NB exhibited the lowest bias with an ABROCA score of 0.0119, closely followed by boosted trees with an ABROCA score of 0.0283, suggesting that these models are the least sensitive to students' financial aid status.

Table 2. Fairness comparison across different methods using ABROCA and DP.

Method	Boarding School		Gender		Handicap		Financial Aid	
	ABROCA	DP	ABROCA	DP	ABROCA	DP	ABROCA	DP
Soft Voting	0.0221	0.0350	0.0390	0.1806	0.1141	0.3261	0.0241	0.0266
Hard Voting	0.0230	0.0259	0.0354	0.2169	0.1302	0.1851	0.0305	0.0477
Stacking	0.0193	0.0255	0.0354	0.2077	0.1153	0.2148	0.0344	0.0611
Boosting	**0.0190**	0.0524	0.0323	**0.1078**	0.1284	0.1652	**0.0283**	**0.0105**
Bagged Trees	0.0282	**0.0248**	**0.0319**	0.2155	**0.1019**	**0.1265**	0.0308	0.0747
DT	0.0238	**0.0219**	0.0325	0.2114	0.1084	**0.1324**	0.0335	0.0850
LR	**0.0208**	0.03266	0.0358	0.2103	**0.1024**	0.1761	0.0378	**0.0176**
NB	0.0321	0.0358	**0.0182**	**0.0232**	0.1166	0.4962	**0.0119**	0.0282

4.1 Ensemble Learning and Fairness

The findings suggest that ensemble learning methods can enhance fairness. However, the evidence is not strong enough to conclude that ensemble methods inherently provide better fairness overall. Although methods like bagging and boosting for trees demonstrated lower bias in certain sensitive attributes, fairness outcomes varied across ensemble methods and were not consistently superior to certain models like LR. We can summarize the findings of our experiment in the following points:

1. Boosted and Bagged trees showed enhanced fairness for certain attributes, e.g. disability and financial aid, however, other ensemble methods such as voting and stacking did not consistently outperform simpler models in fairness. Moreover, LR, a non-ensemble learning model performed comparably to ensembles in terms of fairness across multiple sensitive attributes. This challenges the assumption that ensembles are always fairer.
2. Some ensembles such as hard voting exhibited the highest bias for students with handicaps, indicating that simply aggregating model decisions may reinforce biases from individual models. On the other hand, despite the best AUC of 0.80, stacking did not significantly improve the fairness. This indicates that even an advanced ensemble method does not automatically mitigate disparities.
3. Averaging the fairness of each model over all the protected features reveals that boosted trees, the best ensembled model in terms of fairness, scored an average ABROCA of 0.052. Comparing it to LR, the best performing model, which scored an average ABROCA of 0.049, enforces that simpler models can achieve better fairness with similar predictive performance.

5 Conclusion and Future Work

In this study, we investigated the impact of ensemble learning techniques such as bagging, boosting, voting, and stacking on both the predictive performance as well as the fairness in student dropout prediction. Our analysis was conducted on a real-world dataset of the Moroccan K-12 education system. We compared the ensemble learning methods to the base line machine learning models.

Our findings suggest that while ensemble learning improve the predictive performance, its effect on the fairness is inconsistent. Specifically seeing that methods such as the results varied from one protected group to another. For example we noticed that bagged trees demonstrated the lowest bias for students with disabilities, while boosted trees exhibited the least bias for boarding school availability. Moreover, not all ensemble learning methods improved fairness, techniques such as stacking and voting reinforced biases for students with disabilities. This non uniformity in results across attributes raises critical concerns regarding the ethical deployment of ensemble SDP models. Should institutions prioritize ensemble learning techniques if they do not consistently improve fairness? Our

results suggest that simply increasing model complexity does not guarantee equitable predictions.

As suggestions for future work, investigations on fairness-aware ensemble learning that incorporate fairness constraints into ensemble techniques through data pre-processing for example. A broader model selection for such studies should be taken into consideration. In addition, incorporating more fairness metrics such as demographic parity, equalized odds, or counterfactual fairness could provide a more comprehensive analysis. Finally, integrating these models into real-world interventions and assessing their impact on retention strategies could provide deeper insights into their practical utility.

References

1. Abdul Bujang, S.D., Selamat, A., Krejcar, O., Mohamed, F., Cheng, L.K., Chiu, P.C., Fujita, H.: Imbalanced classification methods for student grade prediction: a systematic literature review. IEEE Access **11**, 1970–1989 (2023)
2. Aboukacem, A., Berrada, I., Bergou, E.H., Iraqi, Y., Mekouar, L.: Investigating the predictive potential of large language models in student dropout prediction. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education, pp. 381–388. Springer, Cham (2024)
3. Anderson, H.J., Boodhwani, A., Baker, R.: Assessing the fairness of graduation predictions. In: Educational Data Mining (2019)
4. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
5. Campbell, C.: The socioeconomic consequences of dropping out of high school: evidence from an analysis of siblings. Soc. Sci. Res. **51**, 108–118 (2015)
6. De Witte, K., Cabus, S., Thyssen, G., Groot, W., van den Brink, H.M.: A critical review of the literature on school dropout. Educ. Res. Rev. **10**, 13–28 (2013)
7. Dutt, A., Ismail, M.A., Herawan, T.: A systematic review on educational data mining. IEEE Access **5**, 15991–16005 (2017)
8. Elbouknify, I., et al.: Student at-risk identification and classification through multitask learning: a case study on the Moroccan education system. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education, pp. 372–380. Springer, Cham (2024)
9. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
10. Gardner, J., Brooks, C., Baker, R.: Evaluating the fairness of predictive student models through slicing analysis. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge, LAK19, pp. 225–234. Association for Computing Machinery, New York (2019)
11. Hu, Q., Rangwala, H.: Towards fair educational data mining: a case study on detecting at-risk students. In: Educational Data Mining (2020)
12. Lee, H., Kizilcec, R.F.: Evaluation of fairness trade-offs in predicting student success (2020)
13. Loukina, A., Madnani, N., Zechner, K.: The many dimensions of algorithmic fairness in educational applications. In: Yannakoudakis, H., Kochmar, E., Leacock, C., Madnani, N., Pilán, I., Zesch, T. (eds.) Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 1–10. Association for Computational Linguistics, Florence (2019)

14. Nuankaew, W.S., Bussaman, S., Nuankaew, P.: Evolutionary feature weighting optimization and majority voting ensemble learning for curriculum recommendation in the higher education. In: Surinta, O., Kam Fung Yuen, K. (eds.) Multidisciplinary Trends in Artificial Intelligence. pp. 14–25. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20992-5_2
15. Plaia, A., Buscemi, S., Fürnkranz, J., Mencía, E.L.: Comparing boosting and bagging for decision trees of rankings. J. Classif. **39**(1), 78–99 (2022)
16. Rokach, L.: Ensemble-based classifiers. Artif. Intell. Rev. **33**(1–2), 1–39 (2010)
17. Romero, C., Ventura, S.: Educational data mining and learning analytics: an updated survey. WIREs Data Min. Knowl. Disc. **10**(3) (2020)
18. Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V.M., Gasevic, D., Chen, G.: Assessing algorithmic fairness in automatic classifiers of educational forum posts. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) Artificial Intelligence in Education, pp. 381–394. Springer, Cham (2021)
19. Sha, L., Raković, M., Das, A., Gašević, D., Chen, G.: Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. IEEE Trans. Learn. Technol. **15**(4), 481–492 (2022)
20. Shafiq, D.A., Marjani, M., Habeeb, R., Asirvatham, D.: Student retention using educational data mining and predictive analytics: a systematic literature review. IEEE Access **10**, 72480–72503 (2022)
21. Smirani, L.K., Yamani, H.A., Menzli, L.J., Boulahia, J.A.: Using ensemble learning algorithms to predict student failure and enabling customized educational paths. Sci. Program. **2022**, 1–15 (2022)
22. Tumer, K., Ghosh, J.: Error correlation and error reduction in ensemble classifiers. Connect. Sci. **8**(3–4), 385–404 (1996)
23. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**(2), 241–259 (1992)
24. Xiong, Z., et al.: A review of data mining in personalized education: current trends and future prospects. Front. Digit. Educ. **1**(1), 26–50 (2024)
25. Xiong, Z., et al.: A review of data mining in personalized education: current trends and future prospects. Front. Dig. Educ. **1**(1), 26–50 (2024)
26. Yu, R., Lee, H., Kizilcec, R.F.: Should college dropout prediction models include protected attributes? (2021)
27. Yu, R., Li, Q., Fischer, C., Doroudi, S., Xu, D.: Towards accurate and fair prediction of college success: evaluating different sources of student data. In: Educational Data Mining (2020)
28. An ensemble learning approach based on tabnet and machine learning models for cheating detection in educational tests. Educ. Psychol. Measur. **84**(4), 708–809 (2024)



Embedding Conversational Safety in AI for Education

Venkata Nagaraju Buddarapu(✉) and Collin Lynch

Department of Computer Science, North Carolina State University, Raleigh, USA
{vbuddar,cflynch}@ncsu.edu

Abstract. Large Language Models (LLMs) and generative AI hold significant promise in transforming education through personalized learning, instant feedback, and adaptive tutoring. However, deploying these models without robust conversational safety measures can amplify biases, produce harmful content, and overlook cultural nuances. This work emphasizes the need to embed conversational safety mechanisms directly into the design and development of LLMs for educational contexts. We examine the core challenges, including bias, emotional insensitivity, and cultural misalignment, and propose a framework to mitigate these risks. Although policies and regulations provide overarching guidelines, we argue that continuous, context-specific refinement of LLMs supported by quantitative metrics and human oversight offers a more proactive, learner-focused safeguard. By uniting these technical and procedural strategies, AI-driven educational tools can remain inclusive, trustworthy, and attuned to the evolving demands of modern classrooms.

Keywords: Large Language Models · Conversational Safety · Educational Generative AI · Bias · Fairness

1 Introduction

Advanced Large Language Models (LLMs) [25] have transformed education by enabling dynamic textbooks [5], real-time visualizations [9], and interactive dialogue agents [7]. Their adoption promises personalized and adaptive learning experiences at scale. However, adaptive agents can also harm learning by spreading misinformation [16] or offering emotionally detached, culturally insensitive responses [3]. Negative experiences risk exacerbating educational inequities.

Previous research [8] emphasizes that *conversational safety*, that is, maintaining respectful, inclusive, and bias-free dialogue, is important for student outcomes in both analog and online settings. LLMs, like human teachers [19], must navigate cultural and linguistic nuances [4] to avoid reinforcing disparities.

To address these risks, we argue that *conversational safeguards* should be incorporated into the design and implementation of LLM-based systems, and analogous mechanisms should be considered for other generative AI tools. Although conversational safety is a broad term, it is possible to decompose it

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 372–379, 2025.
https://doi.org/10.1007/978-3-031-98462-4_47

into key safety concerns and outline specific technical and procedural approaches to address them. We identify these concerns and outline open research areas. Taken together, we believe that this will enable the AI and Education community reap the benefits of LLMs while taking a more *proactive* approach to mitigating harms.

2 Key Conversational Safety Concerns

We divide conversational safety into five areas: bias, toxicity, emotional insensitivity, cultural insensitivity, and cognitive overload. Each concern is distinct and critical to building safer educational LLMs.

Bias describes the systematic provision of different responses to protected subgroups. Bias is a significant focus of study in education in general [17] and has been well studied for classification models [2]. However, there has not been a significant focus on educational dialogue or similar generative contexts. Consider that in a tutorial context, each utterance represents multiple value judgments. The agent must evaluate the state of the discourse, the state of knowledge, and other features. They must then select appropriate responses that may be more or less remedial or critical. Bias can also arise when a large model evaluates and classifies dialogue for other uses, e.g., in a group discussion setting. Bias then can arise when a dialogue or generative agent systematically "talks down" to one group of students or effectively excludes their comments from participation. Although individual responses can be corrected, systematic dialogue bias over time can affect learners.

Toxicity describes cases in which the LLM or generative model presents text or results that are actively offensive or otherwise unacceptable. This may include excessively negative language, unnecessary critiques, or even direct insults. Although some existing LLMs are often keyed to avoid prominent negative language, jailbreaks and workarounds alone demonstrate that they can generate them [27]. Research has extensively explored the issue of toxicity in LLMs, highlighting concerns about their ability to generate harmful content. Studies such as [21] introduce datasets designed to evaluate toxicity robustness, demonstrating that models can still produce offensive outputs even when guardrails are in place. Other work, such as [13], reveals that toxicity is not always explicit but can manifest subtly through biased or exclusionary language, making mitigation efforts more complex.

Emotional Insensitivity describes cases where the generated content fail to adapt to students' current emotional needs [22]. Students can become increasingly frustrated, confused, or apathetic as they work. In these local contexts, language that is critical or, worse yet, unnecessarily optimistic can risk increasing those feelings. Consequently, it is necessary to calibrate language to the student's emotional needs or rules globally (as we do with Toxicity) and at the moment. If we do not, we risk ending what may be a productive but otherwise frustrating interaction for no good reason.

Cultural Insensitivity refers to cases where models either miss appropriate cultural references or misuse them. Evaluation of open training datasets [15] shows most LLMs have a relatively strong bias towards internet data and American literature. LLMs often align with Western values [26], leading to outputs that misrepresent cultural nuances and reinforce stereotypes. While they are treated as generic, general-purpose models, they have a broad cultural attachment. This can leave them both unable to recognize specific cultural and unable to generate them. The impact of this disconnect can range from slight, producing agents that are marginally less engaging because they do not feel like authentic characters, to relatively severe, producing agents that cannot understand or express concepts to the target audience. Expert teachers, unlike LLMs, are integral members of their communities, constantly evolving as they observe and adapt to cultural differences. This plays a vital role in their ability to work with the students.

Finally, **Excess Cognitive Load**. In generating student-facing text, we must also be concerned about students' ability to process the text. Research on cognitive load has shown how educational materials are presented can inhibit student learning through unnecessary complexity [12]. The appropriate complexity level is dependent and, like well-defined reading levels, can be tuned based on students' language comprehension and performance. While reading levels can be assessed directly, it is helpful to consider the conceptual or structural complexity of generated text more broadly when adapting to students' needs.

Addressing these key conversational safety concerns can significantly enhance the quality and equity of AI-supported education as described in Table 1. Reducing bias promotes fairer student engagement and prevents the marginalization of diverse voices. Mitigating toxicity ensures that interactions remain respectful and supportive, building trust in AI tutors. Adapting to students' emotional needs fosters sustained engagement and stronger learning outcomes, while improving cultural sensitivity enables more authentic and inclusive interactions. Finally, managing cognitive load helps tailor content to students' comprehension abilities, making learning experiences more accessible and effective. Together, these efforts contribute to building educational LLMs that are not only safer, but also more responsive, inclusive, and pedagogically aligned.

3 Conversational Safety Framework

We propose a structured framework in Table 2 that addresses key conversational safety concerns in educational LLMs. The framework links risks such as bias, toxicity, emotional insensitivity, cultural insensitivity, and cognitive overload to evaluation methods, mitigation strategies, and iterative refinements. By systematically connecting concerns to actionable techniques and continuous feedback loops, the framework helps developers and designers prioritize conversational safety when building with third-party and open source LLMs. The goal is to create safer, fairer, and more inclusive learning environments.

We organize conversational safety methods into evaluations, filtering, mitigations, and iterative refinement, each aligned to different stages of LLM development. Evaluations identify bias, fairness, and safety risks before deployment but cannot adapt to real time interactions. Filtering blocks harmful content after generation but risks over restriction or circumvention. Iterative refinement updates models based on feedback, ensuring responsiveness to changing student needs and evolving safety expectations.

3.1 Evaluations

Ensuring conversational safety in educational LLMs requires fairness metrics and datasets tailored to learners' needs. General metrics such as Disparate Impact and Statistical Parity [2] are useful but miss linguistic, cultural, and educational nuances. Frameworks like AIF360 help evaluate bias but applying them to generative models requires secondary classifiers and post hoc aggregation, as single outputs do not reveal systematic bias.

Toxicity evaluation must address adversarial techniques like jailbreak prompts and simulation hacks [6]. Although LLMs are tuned to avoid explicit toxicity, subtle harms often persist, and perceptions of toxicity vary across student populations. Effective evaluation must combine broad benchmarks with context-specific assessments. Recent fairness work [4] expands bias detection but often overlooks domain-specific challenges. Educational applications require evaluation frameworks sensitive to reading levels, cultural references, and curriculum standards.

3.2 Mitigation

Mitigation strategies depend on model access. Open source LLMs allow input filtering, fairness-constrained training, and output adjustments, offering finer control. Third-party models, however, limit mitigation to input preparation and output moderation.

Emerging methods such as Direct Preference Optimization (DPO) [18] and Odds Ratio Preference Optimization (ORPO) [10] improve user alignment, but their effectiveness in addressing educational fairness and safety remains underexplored. Targeted domain-specific training [23] combining specialized data curation with parameter-efficient fine-tuning will be crucial for building safe and contextually aware educational LLMs.

3.3 Filtering

Real time content filtering remains essential even after mitigation. Systems such as Perspective API [14] and HateCheck [20] offer toxicity and hate speech detection but must be adapted to education-specific contexts. Sentiment analysis complements toxicity filtering by regulating emotional tone in AI-generated content. Education-focused filters must also ensure readability levels and curriculum alignment to maintain pedagogical appropriateness and student trust.

3.4 Iterative LLM Refinement

Conversational safety requires continuous monitoring and human oversight after deployment. Logging interactions and collecting feedback allow developers to curate incremental training datasets that address emerging safety gaps. Parameter-efficient fine-tuning methods such as LoRA [11] support ongoing updates without retraining from scratch. Human review is essential for annotating subtle biases, cultural references, and emotional sensitivity, ensuring LLMs maintain fairness and appropriateness over time.

3.5 Temporal Considerations

Conversational safety operates across stages: evaluations address pre deployment risks, filtering manages real time interactions, and iterative refinement sustains long term adaptation. Static safeguards are insufficient as cultural expectations and safety standards evolve. Continuous model updates guided by feedback and adaptive safeguards are necessary to maintain fairness, inclusivity, and effective learning outcomes. Table 2 summarizes how evaluations, filtering, and refinement interact across the development cycle to address safety challenges in educational LLMs.

4 Opportunities

Integrating LLM-driven educational tools offers unprecedented opportunities for personalized learning, but it also introduces challenges in bias mitigation, toxicity control, emotional and cultural sensitivity, and text complexity adaptation. Unlike rule-based intelligent tutoring systems [24], which provide structured responses with minimal risk, generative AI requires proactive safeguards to ensure equitable, inclusive, and developmentally appropriate interactions.

While regulatory frameworks such as the AI Bill of Rights establish broad ethical principles, technical solutions, including mitigation, filtering, and iterative refinement, offer promising approaches to enhancing user alignment to reduce conversational safety concerns. However, further research is needed to assess their effectiveness in educational settings.

Despite progress in AI safety, significant limitations remain, especially in the education domain. No universal fairness metric fully captures bias, toxicity, emotional insensitivity, cultural misalignment, and text complexity challenges. Additionally, perfect adaptability to individual learners remains an open challenge, and filtering mechanisms risk either over-restricting content or failing to detect subtle biases. Furthermore, static safeguards are insufficient, as AI systems must continuously evolve to reflect cultural and educational shifts. Addressing these challenges requires bias-aware conversational design, sentiment, and cultural sensitivity models, adaptive filtering, and continuous feedback-driven updates to create safer, more inclusive, and more responsive AI-powered learning environments.

Table 1. Key Safety Concerns and Potential Benefits of Addressing Them

Safety Concern	Potential Benefit
Bias	Reduced disparities, more equitable interactions, fairer student engagement, and inclusive AI interactions.
Toxicity	Reduced exposure to harmful content, improved moderation in sensitive discussions, safer AI mediated learning, and increased trust in AI tutors.
Emotional Insensitivity	Emotionally aware AI responses, better engagement strategies, enhanced student comfort, and stronger learning outcomes.
Cultural Insensitivity	Culturally inclusive responses, reduced misinterpretation of sensitive topics, more authentic student interactions, and improved engagement across diverse backgrounds.
Excess Cognitive Load	Text adapted to comprehension levels, simplified understanding, improved accessibility, and more effective learning experiences.

Table 2. Mapping Safety Concerns to Methods

Safety Concern	Evaluations	Mitigations/Filters	Iterative Refinement
Bias	Fairness metrics, Dataset audits, Bias testing	N/A (Bias is systemic)	User feedback loops, Targeted fine-tuning
Toxicity	Toxicity classifiers, Robust adversarial testing	Content moderation, Prompt filtering	User feedback loops, Adaptive moderation, Targeted fine-tuning
Emotional Sensitivity	Sentiment analysis benchmarks, Safety alignment tests	Tone moderation, Context-aware filtering	User feedback loops, Dynamic adaptation, Targeted fine-tuning
Cultural Insensitivity	Representation audits, Bias-aware NLP testing	Rule-based and phrase filtering	User feedback loops, Dataset expansion with diverse sources
Excess Cognitive Load	Readability assessment, Grade-level alignment	Adaptive text simplification models	User feedback loops, Continuous fine-tuning

Conversational safety principles must be embedded throughout the AI development lifecycle to build reliable and ethical AI-driven education. Future educational AI systems should uphold ethical standards and actively support diverse learning needs through engaging, respectful, and pedagogically sound interactions. However, safety research and implementation must extend beyond proprietary efforts, as seen with EdTech initiatives like Khanmigo [1]. Decentralizing and democratizing AI safety research will be critical to fostering transparent and scalable solutions.

While absolute conversational safety is unattainable, adopting a safety-aware approach in design and development ensures that educational AI remains accountable, adaptive, and aligned with evolving pedagogical needs. As AI adoption in education grows, the research community must collaborate to advance open, inclusive, and ethical AI for learning environments.

References

1. Khan Academy: Khanmigo (2023). https://www.khanmigo.ai/. Accessed 29 Apr 2025
2. Bellamy, R.K.E., et al.: AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias (2018). arXiv:1810.01943
3. Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., Zhao, X.: Bias of AI-generated content: an examination of news produced by large language models (2024). arXiv:2309.09825
4. Gallegos, I.O., et al.: Bias and fairness in large language models: a survey (2024). arXiv:2309.00770
5. Gunasekar, S., et al.: Textbooks are all you need (2023). arXiv:2306.11644
6. Gupta, A., Lin, H., Zhou, X.: Jailbreaking large language models: methods, impacts, and mitigations. arXiv preprint (2023). arXiv:2310.00892
7. Han, J., Yet al.: ChEDDAR: student-ChatGPT dialogue in EFL writing education (2024). arXiv:2309.13243
8. Han, S., Liu, D., Lv, Y.: The influence of psychological safety on students' creativity in project-based learning: the mediating role of psychological empowerment. Front. Psychol. **13**, 865123 (2022). https://doi.org/10.3389/fpsyg.2022.865123
9. Hong, J., Seto, C., Fan, A., Maciejewski, R.: Do LLMs have visualization literacy? An evaluation on modified visualizations to test generalization in data interpretation (2025). arXiv:2501.16277
10. Hong, J., Lee, N., Thorne, J.: ORPO: monolithic preference optimization without reference model (2024). arXiv:2403.07691
11. Hu, E.J., et al.: LoRa: low-rank adaptation of large language models (2021). arXiv:2106.09685
12. Huang, C.Y., Wei, J., Huang, T.H.K.: Generating educational materials with different levels of readability using LLMs (2024). arXiv:2406.12787
13. Lee, C., Park, S., Kumar, R.: Unveiling the implicit toxicity in large language models. arXiv preprint (2023). arXiv:2311.17391
14. Lees, A., et al.: A new generation of perspective API: efficient multilingual character-level transformers (2022). arXiv:2202.11176
15. Li, C., Chen, M., Wang, J., Sitaram, S., Xie, X.: CultureLLM: incorporating cultural differences into large language models (2024). arXiv:2402.10946
16. Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.Y., Wang, W.Y.: On the risk of misinformation pollution with large language models (2023). arXiv:2305.13661
17. Kirwan Institute for the Study of Race and Ethnicity: State of the science: implicit bias review 2018-2020 (2020). https://kirwaninstitute.osu.edu/research/state-science-implicit-bias-education-2018-2020. Accessed 17 Feb 2025
18. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: your language model is secretly a reward model (2024). arXiv:2305.18290
19. Rosales, R.R.: Unheard voices of Filipino immigrant teachers in the United States. Ph.D. thesis. University of Dayton (2025)
20. Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., Pierrehumbert, J.: HateCheck: functional tests for hate speech detection models. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 41–58. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.4. https://aclanthology.org/2021.acl-long.4/

21. Smith, J., Brown, K., Wang, T.: Realistic evaluation of toxicity in large language models. arXiv preprint (2024). arXiv:2405.10659
22. Sorin, V., et al.: Large language models and empathy: systematic review. J. Med. Internet Res. **26**, e52597 (2024). https://doi.org/10.2196/52597. https://www.jmir.org/2024/1/e52597
23. Merlyn Mind AI Team: Merlyn mind's education-domain language models (2023). https://www.merlyn.org/blog/merlyn-minds-education-specific-language-models
24. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. **46**(4), 197–221 (2011). https://doi.org/10.1080/00461520.2011.611369
25. Wang, S., et al.: Large language models for education: a survey and outlook (2024). arXiv:2403.18105
26. Wang, W., et al.: Cultural bias and cultural alignment in large language models. Proc. Natl. Acad. Sci. **3**(9), 346 (2023). https://academic.oup.com/pnasnexus/article/3/9/pgae346/7756548
27. Zhao, X., et al.: Weak-to-strong jailbreaking on large language models (2024). arXiv:2401.17256



Beyond Programming Assistance Towards the Integration of Generative AI into Computer Science Education: A Scoping Review

Sonia Proust-Androwkha[1], Valéry Psyché[2], Revekka Kyriakoglou[3], Anis Boubaker[4], Sébastien Lallé[5], Lamprini Chartofylaka[6], Guilherme Medeiros Machado[7], and Nour El Mawas[8(✉)]

[1] Université de Sherbrooke, Sherbrooke, Canada
sonia.proust-androwkha@usherbrooke.ca
[2] Université TÉLUQ, Québec, Canada
[3] Université Paris 8 Vincennes - Saint-Denis, Saint-Denis, France
[4] École de Technologie Supérieure, Montréal, Canada
[5] Sorbonne Université, CNRS, LIP6, 75005 Paris, France
[6] Université des Antilles, Guadeloupe, France
[7] LyRIDS, École centrale d'électronique, Paris, France
[8] CREM, Université de Lorraine, Metz, France
nour.el-mawas@univ-lorraine.fr

Abstract. This paper presents a scoping review of literature on using generative artificial intelligence (GenAI) in computer science (CS) education at the higher education (HE) level. This review's primary research question is: "How is GenAI utilized in the teaching and learning of CS in HE?" A total of 890 articles were initially identified following a protocol based on PRISMA guidelines. After the selection screening, 38 articles were included for quantitative and qualitative analysis. Students mainly use GenAI to assist with programming tasks. However, there is a lack of research for developing effective pedagogical strategies to fully integrate GenAI into CS education. While opinions on its effectiveness in HE are mixed, students' growing enthusiasm for GenAI stresses the critical need for research on pedagogical approaches to effectively handle its integration and facilitate the current educational paradigm shift. This shift calls for proactive pedagogical strategies and research to adapt teaching methods to an AI-driven context. Strategic GenAI use and regulation will equip new CS learners to thrive in the evolving tech landscape.

Keywords: Scoping Review · Computer Science in Higher Education (HE) · Generative Artificial Intelligence (GenAI) in Education (AIED)

1 Introduction

Every introduction of technology that affects learning comes with its share of fears and excitement, and the democratization of generative AI tools is no exception. Free access to tools like ChatGPT, Gemini, and Mistral that reach the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 380–389, 2025.
https://doi.org/10.1007/978-3-031-98462-4_48

general public led to a significant increase in GenAI use by students to complete their computer programming tasks. Such use raises questions on the impact of these models in learning and teaching coding in HE. Innovative approaches for GenAI are beginning to emerge in CS pedagogy. They could positively impact students' skill acquisition [40] and the pedagogical strategies to be implemented. However, using GenAI could lead students to focus more on mastering the formulation of effective prompts to obtain correct responses instead of mastering programming skills. One concern from the teaching perspective is how academic institutions and professors respond to these new practices. Milano et al. [21] described how universities generally adopt three responses: (i) banning the use of GenAI, (ii) using exclusively paper-based assessments, and (iii) proactively integrating GenAI into the learning experience. Each approach has its pros and cons, and may not be directly applicable in the field of CS. Kendon et al. [17] suggest that educators should reconsider the design of learning activities and assessment methods in light of recent developments, focusing on how students construct knowledge, apply skills, and develop processes. In this context, this paper's main research question is: ***"How is GenAI utilized in the teaching and learning of CS in HE?"*** To answer this question, we propose a scoping review of literature on the integration of GenAI in higher CS education.

2 Review Methodology

The literature review presented in this article is based on a scoping review, a method aimed at identifying and synthesizing the state of knowledge on a broad, underexplored, or complex question [41]. The protocol followed the PRISMA guidelines, a recognized method for structuring the stages of a review and minimizing bias or errors [22]. An advanced research was conducted within Web of Science, arXiv, and Érudit, using a search string combining three main components, each using keywords linked by the boolean operator "AND": GenAI: "generative artificial intelligence" OR "genAI" OR "generative AI" OR "large language model" OR "chatbot" OR "conversational agent" OR "chatGPT" OR "copilot" OR "gemini" OR "Bing" OR "Bard". Education: "computer science education" OR "computer education" OR "CS". Degree: "higher education" OR "college" OR "university" OR "undergrad" OR "graduate" OR "postgraduate" OR "adult education". After removal of duplicates, the references were imported in Covidence, a platform designed for conducting PRISMA-based systematic reviews. The inclusion criteria were as follows: (a) peer-reviewed scientific publications, (b) written in English, (c) from the past three years, and (d) a context specific to HE. The identification stage, completed on October 17, 2024, yielded 227 references. Each document was independently reviewed by two team members via Covidence. In case of disagreement, one or two additional members intervened. At the end of the selection process, 66 articles were retained (see Fig. 1). A thematic analysis was conducted to identify the main emerging categories and to structure the results according to the dominant orientations in the selected papers. This analysis revealed several research avenues. However, due to

editorial constraints, we chose to focus the presentation on the main axis in the corpus: *uses of GenAI and their impact on learning* (38 out of the 66 articles).

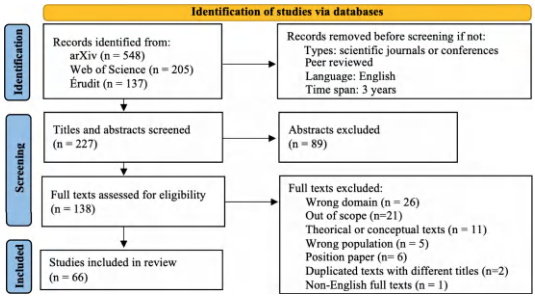


Fig. 1. PRISMA Flow Diagram

3 Results

3.1 Corpus Characteristics

Among the 38 papers of the corpus, 61% are from conferences, and 39% from journals. The number of publications remains low until the end of 2023, then steadily increases from January 2024. No articles precede March 2023. The types of research identified include experimental studies (21 articles) and case studies (20), 3 studies combining both. The most frequent data collection method is surveys, used in 30 studies. Digital trace analysis are also widespread (12 papers), and may include the collection of logs, student submissions, and eye-tracking data. Other methods, such as interviews (8 papers), analysis of academic records (5 papers) and observations (4 papers), are used to a lesser extent. 53% of the studies combine multiple methods. 68% of the studies focus on programming courses, 8% on software engineering, and 5% on algorithms/AI/data science (the rest being unreported). 92% of the courses are undergraduate. For students, the sample sizes range from 8 to 983 (mean of 181, median of 88, st. dev. of 246). Gender distribution is provided in 13 studies. On average, 72% of participants are men, 27% are women, and 1% have other gender identities. For the 7 studies that involve teachers, the sample sizes range from 1 to 57 (mean of 16, median of 8, std. dev. of 19). 95% of the studies involve participants from Western countries (50% North America, 29% Europe, 16% Australasia). Next, 16% focus on East Asia, 8% on the Middle East, one on South America, and none on Africa. 5 studies involve multiple countries.

3.2 Thematic Analysis on Uses of GenAI and Its Impact on Learning

In the corpus, 38 studies focus on students' perceptions of GenAI or its impact on their learning [1–16,18–20,23–30,32–40,42], while seven studies examine

teachers' perceptions [1,5,26,27,29,31,39]. They highlight several advantages and issues related to using GenAI in teaching and learning CS in HE.

•***Positive perceptions of GenAI use:*** GenAI applications are mainly categorized into code generation and debugging, searching for theoretical explanations or reformulations, and rapid solution validation. LLMs are used to create code skeletons, generate solution fragments, identify and correct errors, provide debugging suggestions, or speed up coding tasks. For example, in [14], students primarily use AI to verify or compare their own code, while in [30], some leverage AI to generate alternative solutions, broadening their perspectives and enhancing their problem-solving skills. One of the main appeals of AI for students is time efficiency. They appreciate how quickly AI provides answers, often more effectively than traditional search engines [2]. AI also serves as a virtual tutor, helping to clarify theoretical concepts, providing code examples, and generating ideas [2,6,14,25,30,37]. In [40], students note that AI simplifies repetitive technical aspects, freeing time for complex and creative tasks. The sometimes unexpected suggestions of AI [25] also provide new coding perspectives, enhancing the enjoyment of learning. Finally, AI usage appears to play a significant role in reducing student stress and increasing engagement [37,40]. AI supports self-paced learning beyond instructors' availability and offers a judgment-free space for questions [5,14,40]. This autonomy enhances their sense of control over their progress and helps alleviate academic pressure.

•***Barriers, challenges, and ambivalence in trusting AI use:*** Students exhibit ambivalence regarding their trust in AIED: On one hand, they find it useful, but on the other, they worry about inaccurate or incomplete responses. They note that AI can generate inconsistent, hard-to-validate, or even contradictory results depending on the context, reinforcing their skepticism. At times, they feel the need to consult a human expert to verify the information [30]. In another study [12], six out of 22 students feared that using ChatGPT might hinder the development of essential skills and jeopardize their success in the final exam. Similarly, the study [35] highlights that students' intention to use AI depends largely on their overall attitude: They weigh both the perceived usefulness of the tool and the risk of being penalized for unauthorized use by their institution. A study with nearly 150 students [24] indicates that while most students believed that ChatGPT improved productivity, more than half felt that it negatively affected intelligence and creativity.

Formulating adequate queries for AI remains a challenge. Students express dissatisfaction when their queries are too long or complex. Writing well-structured prompts to obtain accurate responses is challenging for many. The study [14] reveals that students unfamiliar with AI or with negative experiences of it, are much more reluctant to adopt it. Their apprehension stems mainly from the fear of receiving incorrect responses and being unable to distinguish facts from misinformation. In the same study, more experienced students appear more tolerant of AI, knowing that they must apply their knowledge to filter out inaccurate or irrelevant responses. Furthermore, some studies emphasize that while AI is considered useful for quickly generating ideas [6], it is also per-

ceived as potentially detrimental to deep learning. Some students even describe its use in academia as unethical, seeing it as a shortcut that induces guilt and stifles creativity. In study [1], where five instructors evaluated ChatGPT's performance, they confirmed the necessity of human intervention, especially for novice students.

•***Self-efficacy and the effects of GenAI use on learning***: Several studies explore GenAI's impact on learning, self-efficacy and skill development. For instance, study [40] shows that ChatGPT enhances computational thinking skills and self-efficacy by reducing purely syntactical writing blocks, enabling more creative and critical thinking. Likewise, [11] reports reduced cognitive load on repetitive tasks, helping students focus on complex skills. However, these benefits are not systematic. Study [20] finds that students with higher self-efficacy or better prior grades tend to use AI less, perceiving less need for it. In contrast, those who feel less competent rely more heavily on AI. Study [29] highlights a phenomenon of illusion of competence, particularly among novice programmers: Some students overestimate their abilities due to AI assistance. This phenomenon, also noted in [7], suggests a non-reflective use of AI that may lead to plagiarism.

Furthermore, other studies [16,29] observe lower performance in students during coding exams requiring deep understanding and application, compared to exercises focused on code reading and debugging. These findings suggest that AI reliance may hinder deep learning. On the other hand, study [18] finds that ChatGPT use had no statistically significant impact on student performance in both practical tasks and midterm exams. Conversely, research study [9] indicates that students using ChatGPT to create UML diagrams perform better than those who do not. By actively leveraging their skills to produce meaningful results, these students avoid total dependence on AI, which fosters learning.

4 Discussion and Conclusion

This scoping review outlines research on GenAI in teaching and learning CS in HE. We discuss here four key findings, perspectives and limitations of this work.

Finding 1 – The Evolution of Learners' and Teachers' Perceptions is Poorly Documented. Most research focuses on students' perceptions and self-reported uses of GenAI, highlighting their motivations and concerns regarding its use. These studies often rely on one-time data collection, conducted after a short exposure period to GenAI or without real pedagogical integration. Furthermore, no research investigates how these perceptions evolve as students integrate these tools more permanently. Teachers' perceptions in these processes are also under-explored, despite their key role in guiding and regulating these uses.

Finding 2 – Contrasting, Even Contradictory, Views on the Impact of GenAI on Learning. Twelve studies examine the impact of GenAI on learning based on empirical data. The conclusions are mixed: some studies suggest that GenAI facilitates knowledge acquisition and improves understanding of complex concepts, while others highlight risks of over-reliance, superficial thinking, and a weakening of fundamental skills.

Finding 3 – Lack of Theoretical Frameworks. A striking observation across the reviewed studies is the near-total absence of theoretical frameworks, particularly in research focused on perceptions of GenAI use. This gap limits the scope of findings and complicates the development of well-founded recommendations on how to integrate GenAI effectively in education.

Finding 4 – Need of Pedagogical Approaches to Structure Learning with GenAI. In programming courses, where the ability to apply knowledge is more crucial than mere theoretical acquisition, this duality raises a question: How can GenAI be leveraged to promote active learning and problem-solving rather than providing merely automated answers? This issue underscores the need to develop specific pedagogical approaches that frame the use of these tools to maximize their benefits without undermining students' skill development.

Perspectives. These findings raise several key issues that warrant further research. A first research avenue could explore how to prevent GenAI from replacing the cognitive effort necessary for developing programming skills. This would involve designing pedagogical strategies that enable students to use GenAI as a tool for reflection and problem-solving, which calls for the need to rethink learning and assessment. Another critical issue is analyzing how students' perceptions and practices evolve after prolonged exposure to GenAI to better understand the long-term effects of these tools on students' autonomy, engagement, and ability to develop critical and analytical CS skills. Additionally, it would be relevant to examine the impact of integrating GenAI on students' academic and professional trajectories. Research could explore whether early exposure to GenAI influences their adaptability to technological advancements, their relationship with knowledge, and their integration into a job market where AI is playing an increasingly significant role. Finally, teacher training appears to be a crucial lever to ensure an effective and ethical integration of GenAI. To guide students in using these tools, educators must be trained both in technical aspects and in pedagogical strategies that optimize benefits while mitigating potential drawbacks.

Limitations. This work has some limitations. First, we focused only on conference and journal articles, possibly omitting other promising initiatives. Second, thematic analyses inherently involve a degree of subjectivity. Although we sought to improve the reliability of our analysis through reflection and discussions, the identified themes reflect our interpretation of the available literature.

Conclusion. This paper presents a current panorama of the use of GenAI in CS education in HE. For this purpose, a systematic review was conducted, and we focused on 38 papers reporting on uses of GenAI and their impact on learning. This review highlights both the potential and challenges associated with integrating GenAI into the teaching and learning of CS in HE. It calls for in-depth reflection on pedagogical strategies, assessment methods, and usage frameworks in this context. More broadly, it raises the following question: Does the future of GenAI integration in CS education rely on a balanced approach that combines instructional support with critical reflection on its implications? This emerging

paradigm invites a rethinking of learning practices and the role of students in an environment where AI is becoming ubiquitous.

Acknowledgments. This study was funded by the COPCOT project (Project-ANR-22-CE38-0003) and the GENEPI project, funded by the Paris 8 University.

Funding Information. This research was supported in part by a grant from the Fonds de dveloppement acadmique du réseau (FODAR) of the Université du Qubec (UQ) and lÉcole de Technologie Supérieure (ÉTS).

References

1. Ahmed, Z., Shanto, S.S., Jony, A.I.: Potentiality of generative AI tools in higher education: evaluating ChatGPT's viability as a teaching assistant for introductory programming courses. steme **4**(3), 165–182 (2024). https://doi.org/10.3934/steme.2024011. http://www.aimspress.com/article/doi/10.3934/steme.2024011
2. Amoozadeh, M., et al.: Trust in Generative AI among Students: an exploratory study, pp. 67–73. Association for Computing Machinery (2024). https://doi.org/10.1145/3626252.3630842. https://dl.acm.org/doi/10.1145/3626252.3630842
3. Araujo, S.M., Cruz-Correia, R.: Incorporating ChatGPT in medical informatics education: mixed methods study on student perceptions and experiential integration proposals. JMIR Med. Educ. **10** (2024). https://doi.org/10.2196/51151
4. Bernstein, S., et al.: "Like a Nesting Doll": analyzing recursion analogies generated by CS students using large language models. In: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, Milan Italy, pp. 122–128. ACM (2024). https://doi.org/10.1145/3649217.3653533. https://dl.acm.org/doi/10.1145/3649217.3653533
5. Boguslawski, S., Deer, R., Dawson, M.G.: Programming education and learner motivation in the age of generative AI: student and educator perspectives. ILS (2024)
6. Budhiraja, R., Joshi, I., Challa, J.S., Akolekar, H.D., Kumar, D.: "It's not like Jarvis, but it's pretty close!" - examining ChatGPT's usage among undergraduate students in computer science, pp. 124–133. ACM (2024). https://doi.org/10.1145/3636243.3636257. https://dl.acm.org/doi/10.1145/3636243.3636257
7. Chen, B., Lewis, C., West, M., Zilles, C., Assoc Computing Machinery: Plagiarism in the age of generative AI: cheating method change and learning loss in an intro to CS course, pp. 75–85 (2024). https://doi.org/10.1145/3657604.3662046
8. Conde, M., Román-Gallego, J.Ã.: Do engineering students know how to use generative artificial intelligence? A case study, pp. 241–254. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-61691-4_16
9. Cámara, J., Troya, J., Montes-Torres, J., Jaime, F.J.: Generative AI in the software modeling classroom: an experience report With ChatGPT and unified modeling language. IEEE Softw. **41**(6), 73–81 (2024). https://doi.org/10.1109/MS.2024.3385309. https://ieeexplore.ieee.org/abstract/document/10494340
10. Denny, P., MacNeil, S., Savelka, J., Porter, L., Luxton-Reilly, A.: Desirable characteristics for AI teaching assistants in programming education. In: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, Milan Italy, pp. 408–414. ACM (2024). https://doi.org/10.1145/3649217.3653574. https://dl.acm.org/doi/10.1145/3649217.3653574

11. Gardella, N., Pettit, R., Riggs, S., ACM: performance, workload, emotion, and self-efficacy of novice programmers using AI code generation, pp. 290–296 (2024). https://doi.org/10.1145/3649217.3653615
12. Haindl, P., Weinberger, G.: Students' experiences of using ChatGPT in an undergraduate programming course. IEEE Access **12**, 43519–43529 (2024). https://doi.org/10.1109/ACCESS.2024.3380909. https://ieeexplore.ieee.org/document/10478015/
13. Hamme, S., Ottinger, S., Zonnchen, B., Hohendanner, M., Hobelsberger, M., Thurner, V.: ChatGPT in higher education: perceptions of computer science-related students — IEEE Conf. Publication — IEEE Xplore (2024). https://doi.org/10.1109/EDUCON60312.2024.10578742. https://ieeexplore.ieee.org/abstract/document/10578742
14. Hou, I., Mettille, S., Man, O., Li, Z., Zastudil, C., MacNeil, S.: The effects of generative AI on computing students' help-seeking preferences, pp. 39–48. Association for Computing Machinery (2024). https://doi.org/10.1145/3636243.3636248. https://dl.acm.org/doi/10.1145/3636243.3636248
15. Hutson, J., Jeevanjee, T.: Perceptions and aspirations of undergraduate computer science students towards generative AI: a qualitative inquiry. J. Biosensors Bioelectron. Res. **2**(3) (2024). https://doi.org/10.47363/JBBER/2024(2)120. https://digitalcommons.lindenwood.edu/faculty-research-papers/639
16. Jošt, G., Taneski, V., Karakatič, S.: The impact of large language models on programming education and student learning outcomes. Appl. Sci. **14**(10), 4115 (2024). https://doi.org/10.3390/app14104115. https://www.mdpi.com/2076-3417/14/10/4115
17. Kendon, T., Wu, L., Aycock, J.: AI-generated code not considered harmful. In: Proceedings of the 25th Western Canadian Conference on Computing Education, pp. 1–7 (2023)
18. Kosar, T., Ostojić, D., Liu, Y.D., Mernik, M.: Computer science education in ChatGPT era: experiences from an experiment in a programming course for novice programmers. Mathematics **12**(5), 629 (2024). https://doi.org/10.3390/math12050629. https://www.mdpi.com/2227-7390/12/5/629
19. Lauren, P., Watta, P.: Work-in-progress: integrating generative AI with evidence-based learning strategies in computer science and engineering education, pp. 1–5 (2023). https://doi.org/10.1109/FIE58773.2023.10342970. https://ieeexplore.ieee.org/abstract/document/10342970
20. Margulieux, L.E., et al.: Self-regulation, self-efficacy, and fear of failure interactions with how novices use LLMs to solve programming problems, pp. 276–282. ACM (2024). https://doi.org/10.1145/3649217.3653621. https://dl.acm.org/doi/10.1145/3649217.3653621
21. Milano, S., McGrane, J.A., Leonelli, S.: Large language models challenge the future of higher education. Nat. Mach. Intell. **5**(4), 333–334 (2023)
22. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group: Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. Ann. Internal Med. **151**(4), 264–269 (2009)
23. Nguyen, S., et al.: How beginning programmers and code LLMs (Mis)read each other. In: Proceedings of the CHI Conf. on Human Factors in Computing Systems, Honolulu, HI, USA, pp. 1–26. ACM (2024). https://doi.org/10.1145/3613904.3642706. https://dl.acm.org/doi/10.1145/3613904.3642706
24. Pabreja, K., Pabreja, N.: Understanding college students' satisfaction with ChatGPT: an exploratory and predictive machine learning approach using feature

engineering. MIER J. Educ. Stud. Trends Pract. (2024). https://doi.org/10.52634/mier/2024/v14/i1/2568. http://www.mierjs.in/index.php/mjestp/article/view/2568
25. Padiyath, A., et al.: Insights from social shaping theory: the appropriation of large language models in an undergraduate programming course, pp. 114–130. ACM (2024). https://doi.org/10.1145/3632620.3671098. https://dl.acm.org/doi/10.1145/3632620.3671098
26. Prakasha, G.S., Sanskriti, R., Ishani, B.: User experiences of ChatGPT among engineering students, teachers, and working professionals in India. J. Educ. Online **21**(2) (2024). https://www.proquest.com/scholarly-journals/user-experiences-chatgpt-among-engineering/docview/3075714557/se-2?accountid=12543
27. Prather, J., et al.: The robots are here: navigating the generative AI revolution in computing education. In: Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education, New York, NY, United States, pp. 108–159. Association for Computing Machinery (2023). https://doi.org/10.1145/3623762.3633499. https://dl.acm.org/doi/abs/10.1145/3623762.3633499
28. Prather, Jet al.: "It's Weird That it Knows What I Want": usability and interactions with copilot for novice programmers. ACM Trans. Comput.-Hum. Interact. **31**(1), 4:1–4:31 (2023). https://doi.org/10.1145/3617367. https://dl.acm.org/doi/10.1145/3617367
29. Prather, J., et al.: The widening gap: the benefits and harms of generative AI for novice programmers, vol. 1, pp. 469–486. Association for Computing Machinery (2024). https://doi.org/10.1145/3632620.3671116. https://dl.acm.org/doi/10.1145/3632620.3671116
30. Rajala, J., Hukkanen, J., Hartikainen, M., Niemelä, P.: "Call me Kiran" – ChatGPT as a tutoring chatbot in a computer science course, pp. 83–94. ACM (2023). https://doi.org/10.1145/3616961.3616974. https://dl.acm.org/doi/10.1145/3616961.3616974
31. Sheard, J., Denny, P., Hellas, A., Leinonen, J., Malmi, L., Simon: Instructor perceptions of AI code generation tools - a multi-institutional interview study. In: Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1, Portland OR USA, pp. 1223–1229. ACM (2024). https://doi.org/10.1145/3626252.3630880. https://dl.acm.org/doi/10.1145/3626252.3630880
32. Shoufan, A.: Exploring students' perceptions of ChatGPT: thematic analysis and follow-up survey. IEEE Access **11**, 38805–38818 (2023). https://doi.org/10.1109/ACCESS.2023.3268224. https://ieeexplore.ieee.org/document/10105236/
33. Silva, C., Ramos, F.N., De Moraes, R.V., Santos, E.: ChatGPT: challenges and benefits in software programming for higher education. Sustainability **16**(3), 1245 (2024). https://doi.org/10.3390/su16031245. https://www.mdpi.com/2071-1050/16/3/1245
34. Singh, H., Tayarani-Najaran, M.-H., Yaqoob, M.: Exploring computer science students' perception of ChatGPT in higher education: a descriptive and correlation study. Educ. Sci. **13**(9) (2023). https://doi.org/10.3390/educsci13090924. https://www.mdpi.com/2227-7102/13/9/924
35. Skripchuk, J., Bacher, J., Price, T.: An investigation of the drivers of novice programmers' intentions to use web search and GenAI, pp. 487–501. ACM (2024). https://doi.org/10.1145/3632620.3671112. https://dl.acm.org/doi/10.1145/3632620.3671112
36. Smith, C.Eet al.: Early adoption of generative artificial intelligence in computing education: emergent student use cases and perspectives in 2023, pp. 3–9.

ACM (2024). https://doi.org/10.1145/3649217.3653575. https://dl.acm.org/doi/10.1145/3649217.3653575
37. Styve, A., Virkki, O.T., Naeem, U.: Developing critical thinking practices interwoven with generative AI usage in an introductory programming course, pp. 01–08 (2024). https://doi.org/10.1109/EDUCON60312.2024.10578746. https://ieeexplore.ieee.org/abstract/document/10578746
38. Vadaparty, A., et al.: CS1-LLM: integrating LLMs into CS1 instruction. pp. 297–303. ACM (2024). https://doi.org/10.1145/3649217.3653584. https://dl.acm.org/doi/10.1145/3649217.3653584
39. Xue, Y., Chen, H., Bai, G.R., Tairas, R., Huang, Y.: Does ChatGPT help with introductory programming? An experiment of students using ChatGPT in CS1, pp. 331–341. ACM (2024). https://doi.org/10.1145/3639474.3640076. https://dl.acm.org/doi/10.1145/3639474.3640076
40. Yilmaz, R., Karaoglan Yilmaz, F.G.: The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. Comput. Educ. Artif. Intell. **4**, 100147 (2023). https://doi.org/10.1016/j.caeai.2023.100147
41. Zawacki-Richter, O., Kerres, M., Bedenlier, S., Bond, M., Buntins, K.: Systematic reviews in educational research: methodology, perspectives and application. Springer (2020)
42. Zhang, Z., Dong, Z., Shi, Y., Price, T., Matsuda, N., Xu, D.: Students' perceptions and preferences of generative artificial intelligence feedback for programming **38**, 23250–23258 (2024). https://doi.org/10.1609/aaai.v38i21.30372. https://ojs.aaai.org/index.php/AAAI/article/view/30372

Where Is the Environment in AI Ethics? Integrating Ecoliteracy Into AI Literacy

Daniela Inclezan(✉) and Luis I. Prádanos

Miami University, Oxford, OH 45056, USA
{inclezd,pradanli}@miamioh.edu

Abstract. With the rapid advancement of AI, particularly Generative AI, and the surrounding hype, concerns about its environmental impact have intensified. While some AI researchers have long warned about AI's resource consumption, mainstream attention has only recently increased. We attribute this to a broader lack of ecoliteracy among the public, compounded by varying levels of AI literacy. Ecoliteracy—the ability to understand natural systems and apply this knowledge to build sustainable communities—is essential as AI adoption expands across many realms, including education. We advocate for professional development (PD) programs that integrate AI literacy with ecoliteracy, equipping teachers and education leaders to make environmentally responsible decisions about AI-driven educational tools. Without such efforts, AI use in education will intensify resource demands, exacerbating ecosystem degradation and resource depletion. To address this, we propose a framework that integrates ecoliteracy into AI literacy for educator PD, highlighting activities that promote environmentally responsible AI adoption.

Keywords: Ethics · AI literacy · Ecoliteracy

1 Introduction

Since around 2010, AI has been reshaped by greater computing power and abundant online data, fueling the rapid advancement of Deep Learning (DL). Along with Generative AI, these systems now dominate public perceptions of AI; therefore, in this paper, we use the term AI to specifically refer to DL. Following the hype sparked by ChatGPT's release in November 2022, researchers and the public relatively quickly recognized AI's societal harms (e.g., bias, surveillance, misinformation, job displacement, and concerns on fairness, privacy, and transparency). However, awareness of its environmental impact has lagged behind. One of the earliest exceptions was the work of Bender et al. [1], which sounded the alarm about the environmental costs of AI before these systems were widely available. Their study linked these costs to environmental justice, emphasizing that the most disadvantaged and vulnerable communities disproportionately bear the burdens of ecological harm while lacking access to its benefits due to the digital divide. Over the past year, as AI adoption has expanded across industries, education, and the general public, its environmental effects—particularly the energy and water consumption of data centers, the resource demands of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 390–397, 2025.
https://doi.org/10.1007/978-3-031-98462-4_49

AI model training, and carbon emissions—have gained more media attention. However, significant environmental impacts remain largely overlooked, especially those related to AI's material footprint and full life-cycle assessment. These include resource extraction, electronic waste, the increasing use and turnover of electronic devices, and the growing infrastructure demands required to sustain AI technologies. It is extremely important for education professionals to be aware of AI's environmental costs to make informed decisions about the adoption of AI-powered tools in education, without jeopardizing the natural systems that make life (and thus education) possible.

We thus ask the questions: *Why have the public and 'AI in education' initiatives been slow to recognize AI's environmental impact? Why are certain aspects of AI's environmental footprint still not widely discussed? And how can education professionals be trained to make informed decisions about AI adoption in education while minimizing its environmental costs?*

2 Why Are AI's Environmental Impacts Overlooked?

2.1 Environmental Impacts of AI

The environmental impacts of AI that have received the most media attention include energy consumption, water usage, and carbon footprint. Training large language models, such as GPT-4 and other deep learning architectures, demands vast amounts of electricity, raising concerns about potential strain on power supplies for human communities [7]. In some cases, these energy demands have even contributed to the revival of previously decommissioned projects, such as the Three Mile Island nuclear plant in the U.S., which was shut down following a 1979 accident [24]. The energy demands of AI models result in substantial CO_2 emissions, with estimates suggesting that a single large-scale AI training run can produce as much carbon as multiple cars over their lifetime [13].

AI relies on vast amounts of data for training, which is stored on servers housed in massive data centers. These servers generate significant heat and risk overheating, needing extensive cooling systems. Water plays a crucial role in this cooling process, with data centers consuming large volumes, raising concerns about resource depletion, especially in water-scarce regions. In some cases, local communities have pushed back against new data center construction due to these environmental concerns, as seen with Google's proposed facility in Chile [3].

In contrast, other environmental issues have received far less attention, particularly those that emerge from a *life-cycle assessment of AI* and the *material infrastructure* it depends on, including resource extraction, e-waste, and expanding infrastructure demands. Understanding these impacts requires recognizing that AI is largely being justified by its perceived efficiency, making it appealing to businesses, education, and the public. However, these efficiency claims must be reconsidered in light of Jevons' paradox [10]: as a resource becomes cheaper or more efficient to use, its overall consumption tends to rise rather than decline. For example, as cars became more fuel-efficient, people drove more, leading to

a net increase in global fuel demand. In the case of AI, this paradox manifests as an ever-growing need for data centers, GPUs for training, personal AI-enabled devices, and the infrastructure required to sustain them (e.g., undersea internet cables). These trends are further exacerbated by a Western culture of overconsumption and planned obsolescence, where products are deliberately designed to become outdated or non-functional long before their actual lifespan ends, fueling an unsustainable cycle of technological turnover and the proliferation of e-waste. In educational contexts, for example, the adoption of AI-powered tools necessitates providing devices for every student. As AI providers continuously update their software, compatibility requirements may enforce planned obsolescence, accelerating the turnover of personal devices and laptops used in education.

The production of electronics essential for AI advancement relies on the extraction of rare earth elements, metals for servers and devices (including AI-enabled smart devices), and lithium, which is used in batteries for mobile devices, laptops, and data center backup power [8]. The extraction and manufacturing processes are highly polluting. Additionally, some of the minerals critical to electronics production are classified as *conflict minerals*, linked to geopolitical violence and human rights abuses. At the end of their life-cycle, the disposal of the electronics that power AI is equally problematic. Many electronic components cannot be effectively recycled or separated into distinct material types. Often, these discarded electronics are exported (sometimes illegally) to countries in Asia and Africa, particularly former colonies of Western powers, where they are dumped in impoverished areas. In these regions, vulnerable populations attempt to salvage valuable components under hazardous conditions, while toxic residues seep into water sources, posing severe environmental and health risks. This exploitative practice is known as *waste colonialism*. AI excels at optimizing systems and processes, making them more efficient and faster. However, within consumerist and growth-oriented societies, this acceleration often intensifies the extraction of non-renewable resources, fuels overconsumption, and generates more waste [16].

Uncritical adoption of AI in education will drive demand for more data centers to store student data and instructional materials, especially with video-based tools. It will also increase resource consumption for training and running AI models specifically tailored to education tasks, while requiring frequent upgrades to the latest electronic devices, contributing to higher resource use and e-waste, without necessarily advancing equity in education [9].

2.2 Potential Reasons

AI's significant environmental costs are often neglected due to several factors, all rooted in a broader lack of ecoliteracy and addressable through AI education:

1. **The Western worldview that positions humans as dominators of nature**, reinforcing a disconnect from environmental dependency. This mindset fosters flawed views of sustainability that treat the economy, society, and

environment as independent components (see Fig. 1(a)). In contrast, ecological economists emphasize that the economy exists within society, which in turn is embedded in and dependent on the environment (see Fig. 1(b)).

2. **The use of immaterial metaphors to describe AI in the media**, such as "the cloud," which obscure its materiality and its ecological footprint [2].
3. **The challenge of holistic or systemic thinking**, making it difficult to fully grasp and assess AI's material entanglements. Assessing AI's environmental impact demands systems thinking—analyzing interconnections and indirect effects across economic, social, and political domains [26]—as well as life-cycle assessment (LCA), a data-intensive method that tracks environmental costs from material extraction to disposal. For AI, LCA is especially challenging due to system complexity and limited corporate transparency.
4. **The prevalence of techno-solutionism and techno-optimism.** Techno-solutionism is the oversimplifying belief that technology alone can solve complex problems without addressing systemic issues. Techno-optimism is the assumption that technological advancements will drive continuous progress and well-being improvement, which is impossible on a finite planet.

3 AI Literacy

The term *AI literacy* was first introduced by Kandlhofer et al. in 2016, who defined it as the ability to understand "the techniques and concepts behind AI products and services" [17], building upon earlier notions of data, media, computer, and digital literacy [20]. In 2020, Long and Magerko [23] refined this definition as "a set of competencies that enable individuals to critically evaluate AI technologies, communicate and collaborate effectively with AI, and use AI as a tool online, at home, and in the workplace." Building on these definitions, a variety of AI literacy frameworks and assessment instruments have emerged. Table 1 provides a summary of the ethical concerns explicitly addressed in established AI literacy frameworks and assessment tools. Among the eight entries, only one references the Sustainable Development Goals (SDGs), which encompass a broad range of objectives beyond environmental issues, and none of the entries explicitly address *environmental concerns*. This highlights a significant gap in AI literacy, particularly in preparing educators to make informed decisions about

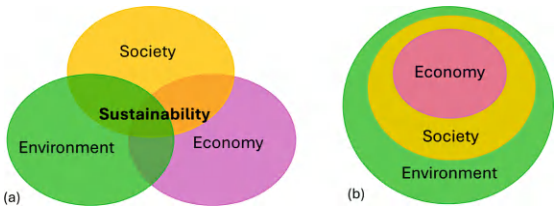


Fig. 1. Views on sustainability: (a) Mainstream view; (b) A more accurate view

AI in education and fostering public awareness of responsible AI use. While environmental considerations have started to appear in some AI curricula (e.g., [12]) and industry initiatives (e.g., [14]), they remain rare, peripheral, and have yet to be fully integrated into mainstream academic AI literacy frameworks.

4 Ecoliteracy

The omission of AI's environmental costs in the AI literacy frameworks we analyzed reflects a broader lack of ecoliteracy in the public consciousness. *Ecoliteracy* refers to an understanding of the principles that govern ecosystems, and the application of these principles to create sustainable human communities and societies [4,25]. In 1992, David Orr argued that the ecological crisis is, at its core, a crisis of education [29]. He pointed to Western education's compartmentalization and hyperspecialization as barriers to developing ecoliteracy, as they discourage holistic thinking and promote fragmented, abstract perspectives [30]. Orr called for breaking down disciplinary boundaries, fostering emotional connection to local ecosystems, and incorporating direct experiences in nature into education. In 2016, Timothy Goodwin [11] conducted interviews with prominent environmental figures to explore how education could be restructured to enhance

Table 1. Ethical concerns included in AI literacy frameworks and assessment tools.

Publication	Reference Describing AI Ethics
Kong et al. (2021) [21]	*[...] (1) the use of AI should not violate human autonomy; (2) the benefits brought by AI should outweigh the risks; and 3) the benefits and risks brought by AI should be distributed equally. These three principles, together with the* ***Sustainable Development Goals (SDGs)****, are the effective guidelines to follow.*
Ng et al. (2021) [28]	*Human-centered considerations (e.g., fairness, accountability, transparency, ethics, safety).*
Carolus et al. (2023) [5]	*Ethical and societal implications consists of the ability to understand the consequences of using AI for society*
Celik (2023) [6]	*personal data privacy, bias and discrimination of AI-based outcomes, and the proliferation of misinformation*
Laupichler et al. (2023) [22]	*the recognition of the importance of data privacy and data security in AI, ethical issues related to AI, and risks or weaknesses that may appear when applying AI technologies*
Kassorla et al. (2024) [18]	*data privacy, the potential for AI bias, and the responsible use of AI-generated content* *transparency and accountability in AI-assisted tasks* and *academic integrity, data privacy, and equitable access.*
Kong et al. (2024) [19]	*ethical principles and considerations around the usage of AI, with a focus on the beneficence of AI, fairness of AI, and human autonomy.*
Ng et al. (2024) [27]	*(1) reliability, (2) safety, (3) privacy, (4) responsibility, (5) transparency, (6) awareness and (7) social good that bring students' positive mindsets towards AI.* *reliability, safety, inclusiveness, transparency and accountability.*

ecoliteracy. Goodwin called for a paradigm shift in how individuals perceive their relationship with the world, moving away from a mindset of domination and ownership toward an understanding of being part of a larger, interconnected system, a perspective he termed *ecological thinking*.

Ecoliteracy's call to (1) reject the 'dominator of nature' view, (2) transcend disciplinary boundaries in favor of holistic thinking, and (3) emphasize embodied experiences in nature that highlight the materiality of technological advancements, directly addresses the root causes of why AI's environmental impacts are often dismissed (see Sect. 2.2). Ecoliteracy also challenges techno-solutionism by integrating insights from the humanities, and counters techno-optimism by emphasizing the physical constraints and limits of the environment [15,16].

5 Integrating Ecoliteracy Into AI Ethics

To help educators make informed decisions about AI tools, we propose PD programs that integrate ecoliteracy with AI literacy. We present a framework for embedding ecoliteracy into the ethical dimensions of AI literacy (e.g., [19,27]), organized around four key pillars (Table 2, 1^{st} column). The second column offers sample activities for integrating these concepts into educator PD programs.

Table 2. Activities integrating ecoliteracy into AI literacy for educator PD programs

Pillar	Sample Activities for PD Program: Participants ...
1. Understanding AI's Environmental Impact	– Get exposed to images and videos of data centers, servers, and their geographic locations to provide a sense of AI's massive physical infrastructure; or visit a data center for a firsthand experience of AI's material footprint. – Explore an interactive dataset on the energy and water consumption of an AI vs. a non-AI-based educational tool.
2. Systems Thinking and Interdisciplinary Integration	– Consider a widely used AI-powered educational tool (e.g., adaptive learning software, automated essay grading, AI tutoring systems) and evaluate whether it reinforces techno-solutionism (i.e., the assumption that AI alone can resolve educational challenges). – Explore how different AI-driven educational tools (text-based, audio-based, and video-based) vary in resource consumption, accessibility, and long-term sustainability.
3. Incorporating Humanities Perspectives	– Analyze historical and contemporary cases where the environmental costs of AI-powered educational tools, particularly e-waste, disproportionately impact marginalized communities. – Assess the design of AI-integrated educational technology, identifying features that promote or mitigate planned obsolescence (e.g., software updates forcing hardware upgrades).
4. Experiential and Embodied AI Learning	– Engage in a science-class-friendly hands-on activity using sensors and AI models to collect and analyze local environmental data for monitoring and prediction. Compare AI-generated insights with local knowledge to assess alignment and accuracy.

References

1. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. ACM, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445922
2. Brevini, B.: Is AI Good for the Planet? Digital Futures, Polity, Cambridge, UK (2021)
3. Cambero, F.A.: Google takes Chile data center plans back to square one on environmental concerns (2024). https://www.reuters.com/technology/google-takes-chile-data-center-plans-back-square-one-environmental-concerns-2024-09-17/
4. Capra, F.: The Web of Life: A New Scientific Understanding of Living Systems. Anchor Books, New York (1997)
5. Carolus, A., Koch, M.J., Straka, S., Latoschik, M.E., Wienrich, C.: MAILS - Meta AI literacy scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. Comput. Hum. Behav. Artif. Hum. **1**(2), 100014 (2023). https://doi.org/10.1016/j.chbah.2023.100014
6. Celik, I.: Exploring the determinants of artificial intelligence (AI) literacy: digital divide, computational thinking cognitive absorption. Telematics Inform. **83**, 102026 (2023). https://doi.org/10.1016/j.tele.2023.102026
7. CNBC: Data centers powering AI could use more electricity than entire cities (2024). https://www.cnbc.com/2024/11/23/data-centers-powering-ai-could-use-more-electricity-than-entire-cities.html
8. Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven, CT (2021)
9. Crooks, R.N.: Access Is Capture: How Edtech Reproduces Racial Inequality. University of California Press, Oakland, CA (2024)
10. Giampietro, M., Mayumi, K.: Unraveling the complexity of the Jevons paradox: the link between innovation, efficiency, and sustainability. Front. Energy Res. **6**, 26 (2018). https://doi.org/10.3389/fenrg.2018.00026
11. Goodwin, T.: Educating for ecological literacy. Am. Biol. Teacher **78**(4), 287–291 (2016). https://doi.org/10.1525/abt.2016.78.4.287
12. Green, B.P.: Teaching a Course on AI Ethics as Part of Engineering/CS Curricula. Markkula Center for Applied Ethics, Santa Clara University (2021). https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/embedding-ethics-into-computing-curricula-resources-and-suggestions/teaching-a-course-on-ai-ethics-as-part-of-engineeringcs-curricula/
13. Hao, K.: Training a single AI model can emit as much carbon as five cars in their lifetimes (2019). https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/
14. IBM: What is AI Ethics? https://www.ibm.com/think/topics/ai-ethics (2024). Accessed 11 Feb 2025
15. Inclezan, D., Prádanos, L.: Promoting ecoliteracy in an introductory database systems course: activities for the first week. In: Dougherty, J.D., Nagel, K., Decker, A., Eiselt, K. (eds.) The 45th ACM Technical Symposium on Computer Science Education, SIGCSE 2014, Atlanta, GA, USA, March 5-8, 2014, pp. 573–578. ACM (2014). https://doi.org/10.1145/2538862.2538909

16. Inclezan, D., Prádanos, L.I.: AI for sustainability: a dangerous fantasy or an unfulfilled potential. In: Torre, D.L., Appio, F.P., Masri, H., Lazzeri, F., Schiavone, F. (eds.) The Impact of Artificial Intelligence in Business and Society: Opportunities and Challenges, pp. 192–211. Routledge (2023). https://doi.org/10.4324/9781003304616-13
17. Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., Huber, P.: Artificial intelligence and computer science in education: From kindergarten to university . In: 2016 IEEE Frontiers in Education Conference (FIE), pp. 1–9. IEEE Computer Society, Los Alamitos, CA, USA (2016). https://doi.org/10.1109/FIE.2016.7757570
18. Kassorla, M., Georgieva, M., Papini, A.: AI literacy in teaching and learning: a durable framework for higher education. Tech. rep., EDUCAUSE (2024). https://www.educause.edu/content/2024/ai-literacy-in-teaching-and-learning/executive-summary
19. Kong, S.C., Cheung, M., Tsang, O.: Developing an artificial intelligence literacy framework: evaluation of a literacy course for senior secondary students using a project-based learning approach. Comput. Educ. Artif. Intell. **6**, 100214 (2024). https://doi.org/10.1016/j.caeai.2024.100214
20. Kong, S.C., Man-Yin Cheung, W., Zhang, G.: Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. Comput. Educ. Artif. Intell. **2**, 100026 (2021). https://doi.org/10.1016/j.caeai.2021.100026
21. Kong, S.C., Zhang, G.: A conceptual framework for designing artificial intelligence literacy programmes for educated citizens. In: Kong, S.C., Wang, Q., Huang, R., Li, Y., Hsu, T.C. (eds.) 25th Global Chinese Conference on Computers in Education (GCCCE 2021), pp. 11–15. Education University of Hong Kong (2021), https://www.eduhk.hk/gccce2021/proceedings.html
22. Laupichler, M.C., Aster, A., Haverkamp, N., Raupach, T.: Development of the "Scale for the assessment of non-experts' AI literacy" - an exploratory factor analysis. Comput. Hum. Behav. Rep. **12**, 100338 (2023). https://doi.org/10.1016/j.chbr.2023.100338
23. Long, D., Magerko, B.: What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–16. ACM, New York, NY, USA (2020). https://doi.org/10.1145/3313831.3376727
24. Mandler, C.: Three Mile Island nuclear plant will reopen to power Microsoft data centers (2024). https://www.npr.org/2024/09/20/nx-s1-5120581/three-mile-island-nuclear-power-plant-microsoft-ai
25. McBride, B.B., Brewer, C.A., Berkowitz, A.R., Borrie, W.T.: Environmental literacy, ecological literacy, ecoliteracy: what do we mean and how did we get here? Ecosphere **4**(5), 67 (2013). https://doi.org/10.1890/ES13-00075.1
26. Meadows, D.H.: Thinking in Systems: A Primer. Chelsea Green Publishing, White River Junction, VT (2008)
27. Ng, D., Wu, W., Leung, J., Chiu, T., Chu, S.: Design and validation of the AI literacy questionnaire: the affective, behavioural, cognitive and ethical approach. Br. J. Edu. Technol. **55**(3), 1082–1104 (2024). https://doi.org/10.1111/bjet.13411
28. Ng, D., Leung, J., Chu, S., Qiao, M.S.: Conceptualizing AI literacy: an exploratory review. Comput. Educ. Artif. Intell. **2**, 100041 (2021). https://doi.org/10.1016/j.caeai.2021.100041
29. Orr, D.W.: Ecological Literacy: Education and the Transition to a Postmodern World. SUNY Press, Albany, NY (1992)
30. Orr, D.W.: Earth in Mind: On Education, Environment, and the Human Prospect. Island Press, Washington, DC (1994)

Integrating Large Language Models and Machine Learning to Detect Struggle in Educational Games

Xiner Liu[1(✉)], Zhanlan Wei[1], Ryan S. Baker[1], Shari J. Metcalf[3], Jiayi Zhang[1], Amanda Barany[1], Stefan Slater[1], Luke Swanson[2], and David J. Gagnon[2]

[1] University of Pennsylvania, Philadelphia, PA 19102, USA
xiner@upenn.edu
[2] University of Wisconsin, Madison, WI 53711, USA
[3] Harvard University, Cambridge, MA 02138, USA

Abstract. This study explores the integration of Large Language Models (LLMs), specifically GPT-4o, with machine learning (ML) to automatically detect struggle behaviors in a science exploration game. We evaluate GPT-4o's ability to analyze text replays, which convert raw gameplay logs into human-readable sequences of player actions, against existing machine-learned struggle detectors. The LLM and ML approaches achieve comparable performance but have complementary strengths. GPT-4o is effective at identifying struggle in instances where the human-engineered features used in traditional ML models fail to capture meaningful behavioral patterns, while ML detectors perform better at identifying struggle in tasks that involve numerical reasoning (e.g., frequent pauses) and in more complex tasks where structured features provide clearer signals.

Keywords: Behavior prediction · Large language models · GPT · Text replay

1 Introduction and Related Work

Educational games have been shown to support learning by immersing students in interactive environments (e.g., [7]). They are well-suited for teaching skills that benefit from practical application (e.g., solving mathematical problems [18], understanding scientific phenomena [1], or exploring historical events [14]). Scholars have long recognized the importance of understanding students' emotions and behaviors during gameplay, which influence how students interact with the game [22] and shape their learning outcomes [19, 23]. Automated detectors have been developed to identify emotions and behaviors from real-time gameplay (e.g., [6, 25]), sometimes using physical sensors (e.g., [10]). However, these detectors have been expensive and time-consuming to develop and rarely scale across systems (though outside of game context, see [17]).

Recently, the emergence of large language models (LLMs) like GPT have opened new possibilities for analyzing textual data at scale in educational contexts (e.g., [2]). Prior research suggested that LLMs can reliably process text for various tasks such as

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 398–405, 2025.
https://doi.org/10.1007/978-3-031-98462-4_50

identifying thematic patterns [5], automating assignments grading [15], and providing feedback on written responses [24]. However, LLMs cannot currently directly ingest the types of interaction data that detectors have worked from. To use this new generation of methods, we must find a way to translate interaction data into a form that LLMs can easily understand and corresponds to something in their training corpuses.

One promising approach draws on a method previously used by researchers to label interaction data for training detectors. Humans, like LLMs, cannot easily read interaction data. Therefore, Baker et al. [4] proposed re-representing interaction data in an easy-to-read format, text replays—readable, sequential narratives of students' actions derived from interaction logs. Text replay has been used by teams of coders to generate ground truth labels for training ML models to detect specific behaviors (e.g., [3, 8, 11, 16, 21]). They allow rapidly identification of the behavior of interest (2–6 times faster than classroom or video observation) while achieving inter-rater agreement only slightly lower than classroom observation [4]. The formatted narratives help researchers recognize patterns that would be difficult to detect from raw data alone.

Text replay, at its core, is a textual representation of interaction data, and it is possible that LLMs may be able to read text replays just as humans can. In this study, we examine whether GPT-4o can detect moments of struggle in the science exploration game *Wake* using text replays. Our goal is to assess the feasibility of this approach and to compare its performance with more traditionally machine-learned detectors.

2 Methodologies and Results

Wake: Tales from the Aqualab is a science education game designed for middle school students in grades 6–9. In the game, players take on the role of a young marine biologist named Olivia to learn more about marine biomes and their respective ecological systems in the game. Each ecosystem includes multiple research sites where players take on "jobs." A "job" represents a scientific assignment or mission within the game, such as studying the impact of an invasive species or monitoring biodiversity in a specific habitat. *Wake* uses the Open Game Data infrastructure to track and log player activities [9]. The opengamedata-unity package connects the game to a cloud-based server, which records player actions throughout each session. The game organizes telemetry data into three categories: *Player Actions* (general navigations, e.g., entering data into scientific tools). *System Events* (system messages, e.g., narrative elements that direct the player's actions). *Progression Events* (in-game milestones, e.g., player completing tasks). The telemetry data records 33 distinct player actions, 12 system events, and 6 progression events. Each event includes metadata that specifies the timing, sequence, player identification, and game state at the time of the event. These events are transmitted to the Open Game Data server in real time creating a chronological log of gameplay.

The data used in the study were sampled from data collected from January 2024 to May 2024. This dataset includes 19,186 players across 42,889 sessions. We sampled data in three rounds. In the first round, we selected 200 cases (40 per month) to evaluate how well GPT and ML could detect struggle for *Wake*. Based on what we learned, we refined the instructions given to GPT. The second round included 100 new cases (20 per month) to check if the updated prompt caused overfitting and if GPT and ML could

be integrated for more accurate struggle detection. The third round added another 100 cases to test whether the hybrid GPT-ML model could generalize to unseen data. We only included jobs that lasted at least five minutes and had over 100 logged actions to make sure there was enough gameplay data for analysis.

2.1 Creating and Coding Text Replay Clips

We then converted log data into text replays to facilitate the examination and coding of student struggle behaviors. Each replay included all actions from when a player accepted a job until they moved to a different job. Log events unrelated to gameplay or job completion were removed. Using timestamps from the logging system, we calculated the time elapsed since the start of each gameplay session (defined as a continuous period between entering and exiting the game). Elapsed time was rounded to the nearest second and listed cumulatively. If multiple events had the same timestamp, they were grouped under the same time to reflect co-occurrence. Session numbers were recorded to separate distinct play periods, and elapsed time was reset at the start of each session. Each text replay began with the job name, which allowed coders to cross-reference it with the *teacher guide* (sites.google.com/wisc.edu/waketeacherguide) that details each job's objectives and expected progression as context for identifying student struggling.

Evaluating the performance of GPT requires reliable ground truth labels. After generating the text replays, two graduate students with expertise in qualitative coding independently labeled each clip as "Struggle" or "Not struggle," based on all actions within each clip. Struggle was defined as any indicator where the player might be facing difficulties or in need of help. The definition was intentionally simplistic and broad, as the goal was not to exhaustively list all possible manifestations of struggle. Instead, both coders were encouraged to evaluate each case individually while considering the specific context in which the struggle emerged. Differences in the coders' game experience led to low initial inter-rater reliability ($\kappa = 0.45$) after round 1. Discrepancies between coders were resolved through moderated discussions to reach a consensus. After discussions, their agreement improved to $\kappa = 0.76$ in round 2 and 0.78 in round 3.

2.2 Re-Construction the Machine Learned Struggle Detection

As part of efforts to improve student engagement and provide timely in-game scaffolding, a previous study developed a machine-learned detector to automatically identify instances of struggle within the game [12]. This serves as an ideal benchmark for evaluating the effectiveness of GPT in detecting struggle. To accommodate updates in the game's data logging structure, we re-created 58 features from that previous study. The features were aggregated, selected, and validated following the same approach in [12].

2.3 Automated Coding of Struggle with GPT

We coded the 200 text replay clips extracted from round 1 using GPT via OpenAI's application programming interface (API). For this study, we used the GPT-4o-2024–11-20 model with all hyperparameters set to default values, except for the temperature, which

was set to 0 to increase output consistency. Each prompting approach was replicated three times on a 10% subset (20 clips) of the round 1 data, with no more than two cases showing disagreement across the three runs. We then applied each prompt to code all 200 clips and calculated agreement between GPT's predicted struggle labels and the human-coded ground truth (see Table 1).

1. **Partially Knowledge Engineered Prompt.** We first assess GPT's ability to analyze student interactions and identify struggle based on the same broad definition outlined above. This setup incorporates a partially knowledge-engineered framework, rather than relying solely on the model's reasoning abilities, to reduce the likelihood that GPT would produce inconsistent or irrelevant interpretations of gameplay behavior.
2. **Incorporating Contextual Information.** Human coders rely on implicit knowledge to identify struggles in text replays. GPT, however, does not have access to this contextual knowledge by default. Without such information, GPT might misinterpret behaviors that would otherwise clearly indicate a struggle. Therefore, contextual details (specifically, job objectives, task difficulty ratings, and the expected progression of actions) were added to the text replay clips as a second prompting method.
3. **Annotated Examples with Human Explanations.** Earlier strategies relied solely on GPT's interpretation, this approach, however, introduced four annotated examples with human-provided explanations. These examples were drawn from a separate dataset and provided explicit instances of struggle with rationales. The examples illustrated situations where students 1) paused extensively during an experiment 2) frequently returned to previously visited locations, 3) repeatedly performed actions that deviated from the hints and 4) cycled through the same choices in experiments.

When annotated examples were included in the prompt, GPT performed comparably to the XGBoost model in detecting struggle. This prompting approach was then selected for the next stage of research. However, incorporating contextual information led to a decline in the model's performance, which aligns with previous research suggesting that excessive information can overwhelm GPT during decision-making [13].

Each job in the game is rated by the designers for difficulty across three core dimensions, *Experimentation*, *Argumentation*, and *Modeling*, on a scale from 1 to 5 (see teacher's guide). We calculated the average difficulty of each job to examine whether model accuracy varied with difficulty level. GPT ($\kappa = 0.76$) showed greater reliability on jobs with lower average difficulty (< 1.5, $n = 22$) than XGBoost ($\kappa = 0.69$). One possible explanation is that simpler jobs involve fewer player actions, which produce less variation across engineered features. This reduces the usefulness of engineered patterns that XGBoost relies on to make predictions. A similar reason also explains why.

GPT ($\kappa = 0.63$) is less accurate than XGBoost ($\kappa = 0.74$) on high-difficulty jobs (avg. Difficulty > 4.5, $n = 15$), where more complex player behavior generates patterns that are better captured by the engineered features used by XGBoost.

GPT and XGBoost show contrasting performance for two specific jobs: *Missing Whale* and *Final*. GPT matched human labels perfectly ($\kappa = 1$, $n = 6$), whereas XGBoost shows substantial difficulty ($\kappa < 0$). This may be due to the unusual structure of these jobs. *Missing Whale* is the only job that players can accept without the required upgrade, which often force them to quit and switch to another job before returning. *Final* consists primarily of scripted dialogue, so the player normally completes the job with minimal

Table 1. Performances metrics for each prediction method

Model	Method	Kappa (κ)	Precision	Recall	F1
GPT	Knowledge Engineered	0.68	0.82	0.98	0.89
	Contextual Information	0.45	0.64	0.97	0.77
	Examples	0.70	0.84	0.97	0.90
XGBoost	Feature Engineering [12]	0.72	0.88	0.91	0.90

actions. As a result, the gameplay data for these jobs contains fewer measurable patterns that reduces the predictive power of XGBoost.

Further analysis showed that GPT failed to identify cases where long pauses between actions served as struggle indicators, possibly due to the use of cumulative timestamps in the text replays, which required interval calculations difficult for GPT [20]. We also noted that GPT was more accurate in identifying non-struggle cases (precision = 0.95) than struggle. This may stem from explicit definition of struggle, which leads GPT to label a clip as non-struggle only if it detects a complete absence of relevant indicators.

Iterative Prompt Engineering Through Review of Miscoded Cases. GPT's ability to interpret narrative elements can sometimes cause misclassifications. For example, it consistently labeled all *Eat Seaweed* job clips as struggle due to a scripted line where the character Olivia expresses uncertainty to the guiding character V1ct0r, despite the player making steady progress. To correct this, we updated the prompt to clarify the distinction between story context and player actions. The update slightly improved GPT's performance on the original 200 clips (κ: 0.70 $\longrightarrow$ 0.72) and maintained stable accuracy on 100 unseen clips from round 2 ($\kappa = 0.71$), with consistent output across replications (no more than two differences across three runs on 20 clips).

2.4 Integrating GPT and XGBoost to Detect Struggle

Leveraging these findings, we created a hybrid model that combines the strengths of GPT and XGBoost for predicting struggle (see Fig. 1). If the student is working on a uniquely structured job or one with low difficulty averaged across the three dimensions, we use GPT's prediction. For jobs with high average difficulty, we follow XGBoost's prediction. In other cases where GPT predicts "Not struggle," we accept it if the clip shows no frequent pauses. If frequent pauses are present, we refer to XGBoost and assign "Not struggle" only if its predicted probability exceeds 0.75. If both models predict struggle, we label the case as such. In all remaining cases where the models disagree, we label the clip as "Not struggle" if XGBoost's probability exceeds 0.75, and "Struggle" otherwise. This threshold was selected based on the AUC-ROC curve constructed from the validation set (clips from round 2). The hybrid model achieved κ of 0.92 when validated using human-coded labels for the 200 clips from round 1. On 100 unseen clips from round 3, κ was 0.77, with precision 0.92 and recall 0.86. Although this drop suggests some overfitting, the hybrid model still outperformed the individual models ($\kappa_{XGboost} = 0.69$; $\kappa_{GPT(examples)} = 0.69$).

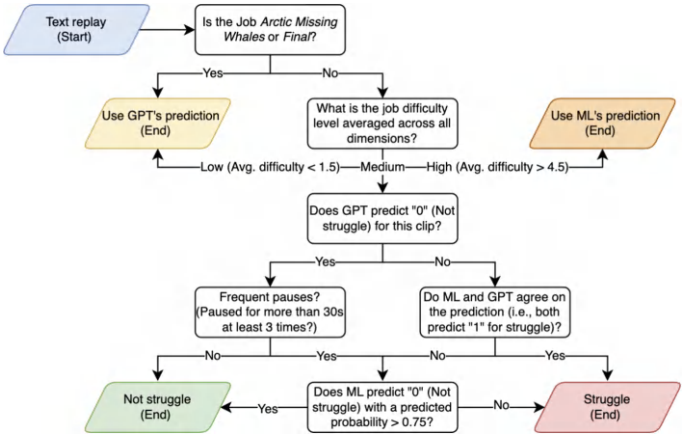


Fig. 1. Hybrid framework for predicting struggle using both GPT and XGBoost detector

3 Discussion and Conclusions

This paper examines the potential of GPT-4o to automatically detect student struggle behaviors in a science exploration game by analyzing text-based replays of log data. We find that GPT-4o performs comparably to a machine-learned detector trained for the same task but has different strengths. GPT-4o was effective in identifying struggle in simpler jobs or jobs with structures not well-represented in the training data and could use information from narrative elements or system messages. This allowed it to identify struggle that manifests in ways that are hard to quantify. GPT-4o also offers practical advantages. Building a machine-learned model requires selecting and tuning features, which can be time-consuming and technically demanding. In contrast, GPT-4o analyzes log data directly once it is converted to text. While some preprocessing is still needed, the overall setup is more accessible to researchers or educators without machine learning expertise. In contrast, XGBoost performed better when struggle followed measurable patterns, such as frequent or prolonged pauses between actions, that were well captured in the numerical features used to train the model. These differences suggest that the two approaches are complementary. Our hybrid model that combines GPT-4o and XGBoost better detected struggle in unseen data than either method alone.

This study has several limitations. First, we could not be certain that the clips labeled as instances of struggle by humans actually reflected the student's experience of struggle, a general problem in using external expert judgment to assess internal experiences. Second, GPT-4o's performance may depend on how struggle manifests in a specific game, and generalizability remains uncertain. Third, it is not entirely clear how GPT-4o arrives at its predictions, making it hard to be certain whether its judgments are based on meaningful factors or spurious correlations. Despite these limitations, GPT-4o shows promise for making behavior detection more accessible and scalable. Its ability to work directly from text replays opens new possibilities for detecting student behaviors across a range of digital learning environments.

Acknowledgments. We would like to thank NSF #DRL-1907437 for support of this project.

References

1. Anderson, J.L., Barnett, M.: Learning physics with digital game simulations in middle school science. J. Sci. Educ. Technol. **22**, 914–926 (2013)
2. Baidoo-Anu, D., Ansah, L.O.: Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. J. AI **7**(1), 52–62 (2023)
3. Baker, R.S., de Carvalho, A.: Labeling student behavior faster and more precisely with text replays. In: Educational Data Mining (2008)
4. Baker, R.S., Corbett, A.T., Wagner, A.Z.: Human classification of low-fidelity replays of student actions. In: Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems, pp. 29–36 (2006)
5. Barany, A., et al.: ChatGPT for education research: exploring the potential of large language models for qualitative codebook development. In: International Conference on Artificial Intelligence in Education, pp. 134–149 (2024)
6. Carpenter, D., et al.: Detecting off-task behavior from student dialogue in game-based collaborative learning. In Artificial Intelligence in Education: 21st International Conference, pp. 55–66 (2020)
7. De Freitas, S.: Are games effective learning tools? A review of educational games. J. Educ. Technol. Soc. **21**(2), 74–84 (2018)
8. Dicerbo, K., Kidwai, K.: Detecting player goals from game log files. In: Educational Data Mining (2013)
9. Gagnon, D.J., Swanson, L.: Open game data: a technical infrastructure for open science with educational games. In: Joint International Conference on Serious Games, pp. 3–19 (2023)
10. Ghali, R., Frasson, C., Ouellet, S.: Towards real time detection of learners' need of help in serious games. In: The Twenty-Ninth International Flairs Conference (2016)
11. Gobert, J.D., Baker, R.S., Wixon, M.B.: Operationalizing and detecting disengagement within online science microworlds. Educ. Psychol. **50**(1), 43–57 (2015)
12. Liu, X., et al.: Struggling to detect struggle in students playing a science exploration game. In: Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play, pp. 83–88 (2023)
13. Liu, X., Zhang, J., Barany, A., Pankiewicz, M., Baker, R.S.: Assessing the potential and limits of large language models in qualitative coding. In: International Conference on Quantitative Ethnography, pp. 89–103 (2024)
14. McCall, J.: Teaching history with digital historical games: An introduction to the field and best practices. Simul. Gaming **47**(4), 517–542 (2016)
15. Nilsson, F., Tuvstedt, J.: GPT-4 as an automatic grader: the accuracy of grades set by GPT-4 on introductory programming assignments (2023)
16. Owen, V.E., et al.: Detecting wheel-spinning and productive persistence in educational games. Int. Educ. Data Min. Soc. (2019)
17. Paquette, L., Baker, R.S.: Comparing machine learning to knowledge engineering for student behavior modelling: a case study in gaming the system. Interact. Learn. Environ., pp. 585–597 (2019)
18. Polycarpou, I., Krausea, J., Rader, C., Kembel, C., Poupore, C., Chiu, E.: Math-City: an educational game for K-12 mathematics. Procedia-Soc. Behav. Sci. **9**, 845–850 (2010)

19. Richey, J.E., et al.: Gaming and confrustion explain learning advantages for a math digital learning game. In: International Conference on Artificial Intelligence in Education, pp. 342–355 (2021)
20. Satpute, A., et al.: Can LLMs master math? Investigating large language models on math stack exchange. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2316–2320 (2024)
21. Sharma, P., Stewart, A.E., Li, Q., Ravichander, K., Walker, E.: Building learner activity models from log data using sequence mapping and hidden Markov models (2024)
22. Shute, V.J., et al.: Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. Comput. Educ. **86**, 224–235 (2015)
23. Taub, M., Sawyer, R., Lester, J., Azevedo, R.: The impact of contextualized emotions on self-regulated learning and scientific reasoning during learning with a game-based learning environment. Int. J. Artif. Intell. Educ. **30**, 97–120 (2020)
24. Wan, T., Chen, Z.: Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. Phys. Rev. Phys. Educ. Res. **20**(1), 010152 (2024)
25. Yildirim, S., Narayanan, S., Potamianos, A.: Detecting emotional state of a child in a conversational computer game. Comput. Speech Lang. **25**(1), 29–44 (2011)



The Impact of Oversampling Techniques on the Detection of Cognitive Presence

Vitor Rolim[1], Cleon Xavier[2,5(✉)], Luiz Rodrigues[3,5], Newarney Costa[2,5], Rafael Lins[1], Dragan Gašević[6], and Rafael Mello[4,5]

[1] Federal University of Pernambuco, Recife, Brazil
[2] Instituto Federal Goiano, Iporá, Brazil
cleon.junior@ifgoiano.edu.br
[3] Technological Federal University of Paraná, Apucarana, Brazil
[4] Federal Rural University of Pernambuco, Recife, Brazil
rafael.mello@ufrpe.br
[5] CESAR School, Recife, Brazil
[6] Monash University, Melbourne, Australia

Abstract. Detecting cognitive presence in online learning environments helps understand and enhance educational interactions in online discussions. However, the lack of annotated data and class imbalance challenges building accurate machine-learning models for this task. This study explores the impact of oversampling techniques using GPT-4 and BERT for the automatic classification of cognitive presence in online discussion, aiming to address data scarcity and improve model generalization. We found that oversampling enhances classification accuracy and Cohen's Kappa, reaching up to 0.59 and 0.40, respectively, for a small sample of the original dataset. These results highlight the potential of recent oversampling methods in improving the detection of cognitive presence, informing future research and practical applications in online education.

Keywords: Community of Inquiry · Natural Language Processing · Online Discussions · LLMs.

1 Introduction

Online discussions forums have become an essential resource for interaction and knowledge construction between students and instructors in online learning [18]. The effectiveness of these online discussions significantly depends on the ability to monitor and analyze the interactions in order to enhance educational support and foster a collaborative and engaging learning environment [24].

The Community of Inquiry (CoI) model [8] is a well-known theoretical framework designed to help instructors understand and enhance interactions in online environments. The CoI describes three dimensions that could support the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 406–413, 2025.
https://doi.org/10.1007/978-3-031-98462-4_51

progress of an online discussion: **social presence**, which refers to students' ability to present themselves as real individuals and interact with others in an online environment; **teaching presence**, which involves the role of instructors in facilitating and directing educational interactions; and **cognitive presence**, which focuses on students' ability to construct and apply theoretical concepts through reflection and dialogue [8]. Analysis of online learning interactions through the lens of this model provides a broad view of the learning process, engagement, and the quality of educational experience. However, conventional analysis methodologies face scalability and real-time application challenges due to the time-consuming nature of manual coding of the content of online discussions [16].

The growing need for rapid and adaptive analysis highlights the importance of automated approaches that can efficiently process large volumes of data. Automation accelerates the analysis process and can provide faster and more informed educational responses, potentially transforming the learning dynamics in online discussions [11]. Additionally, the use of machine learning techniques to analyze the content of online discussions can reveal complex interaction and knowledge patterns that are otherwise challenging to detect manually [7]. In the context of CoI, previous work had applied different machine learning algorithms to automate the analysis of the cognitive presence [1,14]. However, the success of these techniques relies heavily on the availability of high-quality annotated datasets to train machine learning models, which are often scarce and expensive to produce, particularly in specialized domains [4].

To mitigate the issue of limited annotated data, active learning has proven to be an effective strategy for undersampling by selectively identifying the most informative data points for annotation [19]. This reduces the effort of labeling large datasets while enhancing model performance by focusing on critical instances. On the other hand, large language models (LLMs) can facilitate oversampling by generating synthetic data to supplement training datasets [6]. In the context of CoI analysis, while active learning has demonstrated promising outcomes [20], the potential of using LLMs to generate synthetic data for cognitive presence analysis remains unexplored, warranting further investigation.

Therefore, this study aimed to address these limitations by applying oversampling techniques to mitigate the scarcity of annotated data and reduce dependency on large volumes of labeled data. Specifically, we leveraged GPT-4 and BERT models for oversampling in a dataset to extract cognitive presence. The goal of exploring these techniques was to enhance the generalization and applicability of the models, facilitating the automatic and effective analysis of cognitive presence in online environments. Initial results indicate a modest improvement in model performance compared to models without oversampling.

2 Background

In the CoI model, Cognitive presence is crucial for fostering critical thinking, problem-solving, and knowledge construction, being operationalized through a four-phase model [8]. The process begins with a *Triggering Event*, where a

problem is identified to spark curiosity and engagement, followed by *Exploration*, where students exchange ideas and gather information to understand the problem. In *Integration*, they synthesize and critically analyze this information to form cohesive solutions, culminating in *Resolution*, where they apply their knowledge to solve the problem, reinforcing practical skills and confidence [8].

The automated analysis of cognitive presence in online discussions has been predominantly based on a foundational dataset of 1,747 discussion messages originally examined by Kovanović et al. [17]. Their benchmark study achieved a classification accuracy of 58.4% with Cohen's Kappa = 0.41 using text mining techniques within the CoI framework. Subsequent research has largely employed this same dataset to facilitate direct methodological comparisons. Farrow et al. [4] conducted a comprehensive evaluation of linguistic features and class imbalance mitigation, reporting accuracy = 62% with Cohen's Kappa of 0.43 when implementing SMOTE alongside Random Forest classification. Their subsequent work [5] achieved Cohen's Kappa of 0.42 while demonstrating that threaded discourse structure outperformed temporal sequencing in predictive capability.

Hu et al. [14] introduced deep learning architectures, with CNN models reaching accuracy = 63% ($\chi = 0.48$) while incorporating explainable AI (XAI) techniques for model interpretability. Rolim et al. [21] addressed annotation efficiency through active learning paradigms, maintaining classification performance at accuracy = 59% ($\chi = 0.38$) while reducing required annotations by 80%. The sole exception to this dataset-centric paradigm was Ba et al. [1], who incorporated multi-course data to develop a BERT-based model achieving Cohen's Kappa of 0.76.

In that context, this research explores how adopting data augmentation techniques can improve the automatic classification of cognitive presence categories within the CoI model, especially in scenarios of data scarcity. Prior studies emphasize the importance of automating cognitive presence analysis in online discussions [4]; however, more efficient models typically require large volumes of annotated data [21]. A promising solution involves leveraging data augmentation with models like GPT-4 and BERT to mitigate class imbalances and enhance model performance [6]. Therefore, the Research Question (RQ) addressed in this study is: **To what extent can data augmentation techniques improve the automatic classification of cognitive presence categories?**

3 Method

The dataset used in this study originally included six iterations of a fully online, research-intensive master's level course in software engineering, delivered through the Moodle LMS at a Canadian public university between 2008 and 2011. The course was structured into six modules covering 14 different topics in software engineering, such as software requirements, design, and maintenance. As part of the assessment, students were required to record a video presentation on a research paper related to one of the course topics and post a URL to an online discussion, where peers would debate the presentation. Participation in

this activity contributed 15% to the final grade. Over the six course iterations, 81 students produced 1,747 discussion messages [10].

The discussion messages were coded for cognitive presence by two expert annotators using the framework defined in Garrison et al. [9]. If a message did not demonstrate indicators of any cognitive presence phase, it was categorized as "Other." The inter-rater reliability was outstanding, with a percent agreement of 98.1% and a Cohen's Kappa of 0.974. Only 32 disagreements occurred, all of which were resolved through discussion.

Farrow et al. [4] proposed a best practice for evaluating model effectiveness by reserving the most recent course offering for the testing set, which she adopted as the last run of the course. We adopted this approach for comparison reasons. However, unlike Farrow et al. [4] and other similar studies, which used all previous course runs for training, we opted to use only the first run. This decision aligns with the primary objective of our study, which focuses on generating new syntactic instances to address the challenge of limited data for analysis. Table 1 shows the distribution of cognitive presence for training and test sets.

Table 1. Distribution of cognitive presence for training and test sets

	Training set (n = 212)		Test set (n = 237)	
Phase	Messages	%	Messages	%
Other	27	12.75%	28	11.81%
Triggering event	34	16.03%	28	11.81%
Exploration	90	42.45%	76	32.07%
Integration	47	22.16%	83	35.02%
Resolution	14	6.61%	22	9.28%

Then, to address the issue of imbalanced training data, we employed oversampling techniques with the aid of two state-of-the-art models: BERT and GPT. BERT's contextual embeddings allowed us to generate meaningful variations of existing data by capturing diverse semantic formats. At the same time, GPT's generative capabilities enabled the creation of synthetic samples that closely resembled the original dataset. By leveraging both models, we enriched the training set, improving the model's ability to generalize and perform effectively on tasks despite limited initial data availability.

To implement BERT for data augmentation, we utilized the `nlpaug`[1] library. In this context, BERT understands the context of words bidirectionally and generates deep contextual embeddings, improving the representation of textual data through word substitution in text. For each cognitive presence category, additional instances were generated to equalize the number of samples with the dominant category, "Exploration". The augmentation process was carefully designed to preserve linguistic integrity by ensuring that stopwords remained unchanged while a maximum of 100 randomly selected words per instance were modified. This approach allowed us to expand the dataset without compromising the core semantic structure of the original text.

[1] https://nlpaug.readthedocs.io/en/latest/index.html.

GPT, an autoregressive language model, generates synthetic text that maintains the context and stylistic features of the original data [15], addressing issues of class imbalance and scarce annotated data. In this study, GPT-4 was employed for data augmentation to balance cognitive presence categories in the dataset. Three prompt strategies were used: (1) zero-shot without context (generating posts using only category labels); (2) zero-shot with context (including a CoI model explanation); and (3) few-shot with context (adding specific examples of messages from the training set to guide GPT-4 in generating new instances).

Most previous work utilized traditional natural language processing models to automatically extract the cognitive presence [4,5,14,17,21]. However, these approaches heavily rely on the use of linguistic features, which may not always be readily accessible or easy to implement in practical applications. Furthermore, recent research on the topic has highlighted the potential of leveraging BERT to effectively operationalize the extraction of cognitive presence [1], demonstrating its promise in automating this process. Therefore, this study adopted BERT [3] as a classifier for the proposed task.

BERT, a contextual language model built upon a deep neural network with bidirectional processing capabilities, generates embeddings that adapt to the specific textual context of each lexicon occurrence, enabling it to capture variations in meaning [3]. Its architecture allows for fine-tuning by adding a single output layer, optimizing the model for a wide range of NLP tasks with minimal adjustments. This work uses **DeBERTa** [13], an advanced language model that enhances BERT with disentangled attention and relative position encoding to better capture deep contextual and semantic information. Recent literature suggests that when utilizing *embeddings* from models based on the Transformer architecture, such as DeBERTa, there is no need to incorporate traditional textual features, such as word frequencies or cohesion metrics, as the *embeddings* sufficiently encapsulate the information relevant to the classification task [23].

As a baseline, we adopted the pipeline proposed by Farrow et al. [4]. It integrates traditional text mining features, such as named entity counting and message-level metrics (including similarity with previous and subsequent messages, reply counts, and initial message identification, among others), with linguistic tools like Linguistic Inquiry and Word Count (LIWC) [22] and Coh-Metrix [12] to extract indicators of cognitive presence from textual contributions. As a classifier, Farrow et al. [4] applied the Random Forest algorithm, an ensemble method that builds multiple decision trees during training and combines them to produce a more robust and accurate model. Random Forest was used in several experiments due to its effectiveness in handling noisy and imbalanced data, as well as its strong performance in classifying the cognitive presence categories in the CoI model. Similarly, we adopted the Random Forest model (i.e., referred to as *no rebalancing* in Farrow et al. [4]) as our baseline.

To evaluate the effectiveness of the proposed models in classifying cognitive presence, we used two metrics: accuracy and Cohen's Kappa coefficient. Accuracy indicates the proportion of correct classifications made by the model out of the total cases tested. While it provides a general understanding of the model's abil-

ity to classify instances, it can be misleading in imbalanced datasets, where some classes may be overrepresented. To provide a more robust evaluation, we also used Cohen's Kappa coefficient, which measures the level of agreement between two raters when classifying items into mutually exclusive categories, accounting for the possibility of agreement occurring by chance [2].

4 Results and Discussion

Table 2 presents the results of the cognitive presence classification models. The baseline Random Forest model achieved an accuracy of 0.32 and a Cohen's Kappa of 0.05, which differs from [4] due to the dataset split. The DeBERTa model with no data augmentation performed slightly worse, an accuracy of 0.19 and a Cohen's Kappa of 0.04. With the sampling approaches, the DeBERTa model demonstrated significant improvements when combined with oversampling techniques using BERT and GPT-4. The models reached similar performance, with up to 0.59 and 0.40 of accuracy and Cohen's Kappa, respectively. The augmentation with BERT also generated similar outcomes. These results suggest that oversampling with language models can mitigate class imbalance.

Table 2. Results using the proposed training and test sets. Oversampling = OS.

Model	Acc.	Kappa	Model	Acc.	Kappa
RandomForest	0.32	0.05	DeBERTa (OS with Prompt 1)	0.59	0.37
DeBERTa	0.19	0.04	DeBERTa (OS with Prompt 2)	0.59	0.40
DeBERTa (OS with BERT)	0.58	0.40	DeBERTa (OS with Prompt 3)	0.59	0.39

Answering our RQ, our results show that employing oversampling with GPT-4 and BERT outperformed traditional models like Random Forest [4,5,14,17]. Notably, active learning has been applied in earlier work to address the challenge of limited annotated data in online discussions [21]. While [21] focused on selectively identifying the most informative instances for annotation using active learning, our study adopts a different approach by leveraging oversampling techniques with GPT-4 and BERT to generate synthetic data and balance class distributions. Although active learning demonstrates potential, it still relies on the availability of a high-quality dataset to identify relevant instances for training the model. In contrast, oversampling, particularly with GPT-4, requires only a few examples to create synthetic data through few-shot prompts. This makes it a more flexible solution when working with smaller or less comprehensive datasets.

Despite the clear improvements by oversampling the dataset, it is essential to acknowledge the trade-offs involved. The computational and financial cost of training and using large language models may outweigh the performance gains in specific contexts, especially when resources are constrained, and real-time application is required [15]. Additionally, to assess the generalizability of the

developed models, it is essential to test these approaches on multiple datasets from diverse educational contexts. Another direction for future work is testing the models in active virtual learning environments, which will offer valuable insights into their performance in authentic settings. Furthermore, we call for investigations on whether fine-tuning, optimization techniques, and training on larger datasets lead to improvements in classification accuracy.

Acknowledgments. This study was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (310888/2021-2) and Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Finance Code - 001).

References

1. Ba, S., Hu, X., Stein, D., Liu, Q.: Assessing cognitive presence in online inquiry-based discussion through text classification and epistemic network analysis. Br. J. Edu. Technol. **54**(1), 247–266 (2023)
2. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Farrow, E., Moore, J., Gašević, D.: Analysing discussion forum data: a replication study avoiding data contamination. In: LAK' 2019, pp. 170–179. LAK19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3303772.3303779
5. Farrow, E., Moore, J., Gašević, D.: Dialogue attributes that inform depth and quality of participation in course discussion forums. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, pp. 129–134 (2020)
6. Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E.: A survey of data augmentation approaches for NLP. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 968–988. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.findings-acl.84, https://aclanthology.org/2021.findings-acl.84
7. Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., Romero, C.: Text mining in education. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **9**(6), e1332 (2019). https://doi.org/10.1002/widm.1332
8. Garrison, D.R., Anderson, T., Archer, W.: Critical inquiry in a text-based environment: Computer conferencing in higher education. Internet Higher Educ. **2**(2–3), 87–105 (1999). https://doi.org/10.1016/S1096-7516(00)00016-6, http://www.sciencedirect.com/science/article/pii/S1096751600000166
9. Garrison, D.R., Anderson, T., Archer, W.: Critical thinking, cognitive presence, and computer conferencing in distance education. Am. J. Distance Educ. **15**(1), 7–23 (2001). https://doi.org/10.1080/08923640109527071
10. Gašević, D., Adesope, O., Joksimović, S., Kovanović, V.: Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. Internet Higher Educ. **24**, 53–65 (2015)

11. Gašević, D., Kovanović, V., Joksimović, S.: Piecing the learning analytics puzzle: a consolidated model of a field of research and practice. Learn. Res. Pract. **3**(1), 63–78 (2017). https://doi.org/10.1080/23735082.2017.1286142
12. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: analysis of text on cohesion and language. Behav. Res. Methods Instrum. Comput. **36**(2), 193–202 (2004)
13. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: decoding-enhanced BERT with disentangled attention. Int. Conf. Learn. Representations (2021)
14. Hu, Y., Mello, R.F., Gašević, D.: Automatic analysis of cognitive presence in online discussions: an approach using deep learning and explainable artificial intelligence. Comput. Educ. Artif. Intell. **2**, 100037 (2021)
15. Kalyan, K.S.: A survey of GPT-3 family large language models including ChatGPT and GPT-4. Nat. Lang. Process. J., 100048 (2023)
16. Kovanovic, V., Joksimovic, S., Gasevic, D., Hatala, M.: What is the source of social capital? The association between social network position and social presence in communities of inquiry (2014)
17. Kovanović, V., Joksimović, S., Gašević, D., Hatala, M.: Automated cognitive presence detection in online discussion transcripts. In: Proceedings of the Workshops at the LAK 2014 Conference co-located with 4th International Conference on Learning Analytics and Knowledge (LAK'14). Indianapolis, IN (2014). http://ceur-ws.org/Vol-1137/
18. Milman, N.B.: Distance education (2015)
19. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM Comput. Surv. (CSUR) **54**(9), 1–40 (2021)
20. Rolim, V., Mello, R.F., Nascimento, A., Lins, R.D., Gašević, D.: Reducing the size of training datasets in the classification of online discussions. In: 2021 International Conference on Advanced Learning Technologies (ICALT), pp. 179–183. IEEE (2021)
21. Rolim, V., Mello, R.F., Nascimento, A., Lins, R.D., Gašević, D.: Reducing the size of training datasets in the classification of online discussions. In: 2021 International Conference on Advanced Learning Technologies (ICALT), pp. 179–183 (2021). https://doi.org/10.1109/ICALT52272.2021.00061
22. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010). https://doi.org/10.1177/0261927X09351676
23. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)
24. Xia, C., Fielder, J., Siragusa, L.: Achieving better peer interaction in online discussion forums: a reflective practitioner case study. Issues Educ. Res. **23**(1), 97–113 (2013)

From Recall to Reasoning: Automated Question Generation for Deeper Math Learning Through Large Language Models

Yongan Yu(✉), Alexandre Krantz(✉), and Nikki G. Lobczowski(✉)

McGill University, Quebec, Canada
{yongan.yu,alexandre.krantz}@mail.mcgill.ca, nikki.lobczowski@mcgill.ca

Abstract. Educators have started to turn to Generative AI (GenAI) to help create new course content, but little is known about how they should do so. In this project, we investigated the first steps for optimizing content creation for advanced math. In particular, we looked at the ability of GenAI to produce high-quality practice problems that are relevant to the course content. We conducted two studies to: (1) explore the capabilities of current versions of publicly available GenAI and (2) develop an improved framework to address the limitations we found. Our results showed that GenAI can create math problems at various levels of quality with minimal support, but that providing examples and relevant content results in better quality outputs. This research can help educators decide the ideal way to adopt GenAI in their workflows, to create more effective educational experiences for students.

Keywords: Large Language Models · Educational Question Generation · Bloom's Taxonomy · Webb's Depth of Knowledge · Retrieval Augmented Generation

1 Introduction

With the rapid advancements in LLMs, AI has significantly impacted educational development [2,26], transforming how academic content is created and delivered. Recent studies demonstrate LLMs can enhance learning experiences across various scenarios [6,17]. Particularly, recent studies have explored AI applications, including personalized suggestions [32], and learning behavior analysis [3].

While over 60% of educators have experimented with ChatGPT, less than 20% feel adequately prepared to integrate it effectively [13]. One application of LLMs that could significantly support teachers is Question Generation (QG). A high-quality LLM QG tool could significantly reduce the workload of educators [4], as it would free up the time spent on creating problem sets and answers. It could also result in more practice problems for students, enhancing their learning

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 414–422, 2025.
https://doi.org/10.1007/978-3-031-98462-4_52

experience. From a technical standpoint, QG is an existing sub-field of natural language processing (NLP), focused on enabling the automated creation of educational content directly from reference material, such as textbooks [7].

Despite the transformative potential of AI in education, its integration remains underutilized, especially in automated question generation. A need-finding study by Wang et al. (2023) [25] reveals that educators often express reservations about adopting AI tools, citing concerns about the relevance and quality of AI-generated content. Moreover, existing automatic QG tools are not widely used in classrooms due to their limited range in types and difficulty levels [15]. Most systems primarily produce simple recall questions, failing to sufficiently challenge students or promote deeper cognitive processing.

Problem Statement. Our study investigates the gap between AI capabilities and effective educational implementation by answering: **RQ1**: *What impact does increased contextual information have on the quality and cognitive depth of LLM-generated questions?* and **RQ2**: *How can LLM-based question generation systems be designed to produce questions across varying levels of cognitive depth?* We hypothesize that developing a context-aware AI framework integrating established educational taxonomies [14] will enable the generation of high-quality, diverse, and cognitively appropriate questions that align closely with specific educational objectives and content. Our goal is to bridge the gap between AI capabilities and educational needs, potentially increasing educator confidence and the adoption of AI tools in classrooms for improved educational outcomes.

2 Related-Work

The rapid rise of GenAI, exemplified by ChatGPT's release in late 2022, has significantly impacted the educational landscape [19], with the potential to automate 20–40% of teachers' administrative tasks [6]. However, current GenAI models show limitations in math education. For instance, while effective for retrieving theorems, they struggle as conversational tutors and make errors even with elementary problems [9]. These issues stem from a lack of logical reasoning [20], highlighting the need for specialized AI systems in math education.

QG systems have evolved from simple recall questions [28] to more sophisticated models like "QG-Net" [29], which uses recurrent neural networks to generate quiz questions from educational content. However, math-specific QG faces unique challenges, including limited benchmark datasets [31] that can lead to overfitting and question redundancy [12]. Our study aims to address these limitations by creating higher-quality math questions while building AI systems that can truly adapt to students' learning journeys.

3 *Study 1:* Exploring Current Capabilities of GenAI

We explored how GenAI could enhance educational practices, particularly in intelligent tutoring systems and automated quizzing [24]. Our goal was to integrate GenAI seamlessly into student and teacher workflows, ensuring reliability

and context awareness. To achieve this, we investigated whether GenAI could generate comprehension questions relevant to course content and analyzed their cognitive depth using Bloom's Taxonomy [1].

Study Design. We tested three "level of context" scenarios with progressively more information provided to the GenAI. Using a mathematical logic course covering topics such as satisfiability and the compactness theorem, we designed prompts instructing the model to generate five comprehension questions with answers. We employed Gemini-1.5-Pro [22] with a temperature of 0 [33] to minimize output variations. In Scenario 1 (minimal), the instructor provided only the course syllabus and a brief topic summary; in Scenario 2 (moderate), they added their notes for a specific class session; and in Scenario 3 (comprehensive), they included the syllabus, class notes, and references covering course material.

Evaluation Metrics. A math student with expertise in logic evaluated outputs via three metrics: relevance (binary score indicating whether the question was within lesson scope), depth (score from 0-6 corresponding to Bloom's Taxonomy levels), and correctness (binary score assessing answer accuracy).

Table 1. Performance comparisons for three context scenarios. Note: Values are presented as mean ± standard deviation.

Context Scenario	Relevance	Depth	Correctness
Minimal	**1.00 ± 0.00**	2.60 ± 1.14	0.60 ± 0.55
Moderate	**1.00 ± 0.00**	2.60 ± 1.82	0.80 ± 0.45
Comprehensive	0.80 ± 0.45	**2.40 ± 2.07**	**1.00 ± 0.00**

Findings from Initial GenAI Testing. In this preliminary experimentation, our analysis revealed that GenAI indeed can create content that is relevant and high-quality with input support. Results from testing were surprising in the level of relevance because there was a declining trend in the relevance level of the output as we added additional context. However, we found the depth of the questions remained relatively constant, and the correctness of the answers generated improved as more context was provided, reaching perfect accuracy with comprehensive context.

Advancing to the Next Stage: From Bloom's Taxonomy to DOK. Our initial investigation revealed that additional context improved question correctness but not relevance, as the GenAI sometimes extrapolated and went beyond the bounds of what was taught in the course (i.e., creating "hallucinations"). Since instructors can generally verify answer correctness, we prioritized improving relevance in our next phase. We considered approaches like retrieval-augmented generation (RAG) [16] to better ground outputs in provided materials, which could potentially resolve this issue by forcing the AI to source its generation from a "chunk" of the materials provided. Additionally, while Bloom provided a starting framework, we found Webb's depth of knowledge (DOK) [30] framework to be better suited for math education. DOK emphasizes task

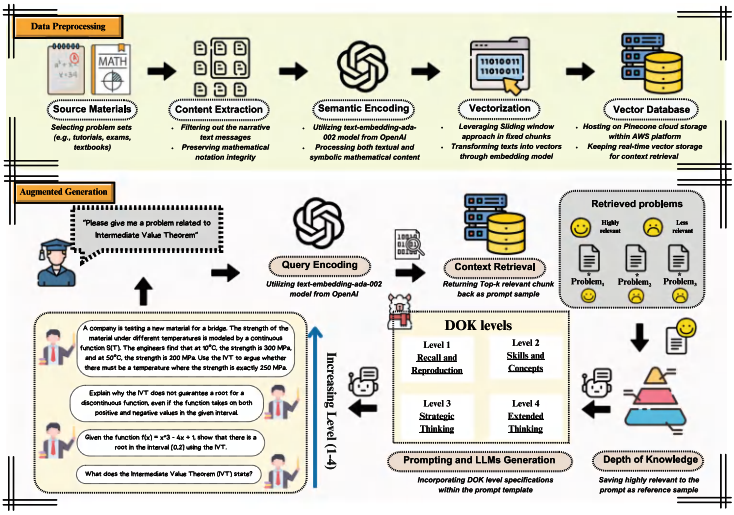


Fig. 1. Overview of the proposed framework with two core components

complexity and contextual knowledge application rather than just cognitive processes, aligning more effectively with mathematical problem-solving and curriculum standards. This framework enables a more precise mapping of question difficulty to the cognitive processes involved in mathematical reasoning, from basic recall to complex problem-solving.

4 *Study 2*: Developing an Improved Framework

Building upon prior findings, we introduced QG-DOK, a question generation framework integrating RAG with Webb's Depth of Knowledge to generate context-aware questions with varying cognitive depth levels, addressing limitations identified in our initial phase. Thus, our second research objective explored how GenAI can generate questions of varying difficulty, making them more adaptable to teachers in math education. Our system, illustrated in Fig. 1, comprises two core components.

RAG Framework. QG-RAG was implemented using a naive RAG framework [10] and integrated with DOK:

- **Data preprocessing**: We gathered and refined mathematical content from textbooks, tutorials, and practice problems, then transformed it into vector representations through an embedding model. These vectors were stored in a database to facilitate efficient retrieval of semantically relevant materials.
- **Augmented generation**: When a user submits a query, the system retrieves relevant content from the vector database to provide contextual grounding. To adjust difficulty and cognitive depth, we incorporated four DOK levels

into our question generation process, as shown in Fig. 2 Ⓐ: *Recall and Reproduction (level-1)*: retrieving basic facts, definitions, and formulas with minimal cognitive effort; *Skills and Concepts (level-2)* selecting appropriate methods and organizing information to solve routine problems; *Strategic Thinking (level-3)*: reasoning, planning, and applying concepts in non-routine scenarios; *Extended Thinking (level-4)*: making connections across concepts and solving complex, multi-step problems.

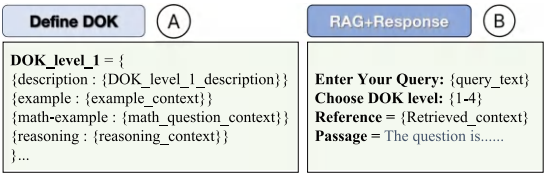


Fig. 2. Level-1 prompt template example Ⓑ User input interface

Implementation Details. As shown in Fig. 1, we embedded a corpus of mathematical content using the TEXT-BEDDING-ADA-002 model. To enhance semantic relevance, documents were segmented into fixed-size chunks using a sliding-window approach [27]. UI functionality is shown in Fig. 2 Ⓑ, users were prompted to input two key pieces of information: the mathematical concept they wish to explore and the desired DOK level. In second study, we evaluated three off-the-shelf LLMs for question generation with default temperature settings, including GPT-4o [21], DEEPSEEK-V3 [5] and GEMINI-1.5-PRO [11].

Evaluation Metrics. To assess our QG-DOK framework, we compared two implementations: (1) DOK, providing only DOK level definitions, and (2) DOK+ RAG, which retrieved relevant examples from a vector database. We evaluated using G-Eval [18] to measure relevance, DOK alignment, and appropriateness. Additionally, we incorporated the Paraphrase n-gram Change (PINC) score [23] to quantify lexical diversity.

Findings from Improved Work. Our evaluation demonstrated that DOK+ RAG consistently outperformed DOK-only across all tested LLMs , particularly for higher-order thinking skills (DOK Levels 3 & 4). DOK+RAG improved both relevance and appropriateness scores, with GPT-4o showing the most significant gains in appropriateness. Although DOK alignment showed mixed results, DOK+RAG generally improved depth accuracy at higher cognitive levels. Also, high PINC scores (average 0.92) indicated strong lexical diversity in question rephrasing. Despite these improvements, challenges remained. The depth alignment at Level 2 was inconsistent across the models, suggesting that LLMs struggle with categorizing mid-level cognitive complexity. We also identified persistent issues in mathematical notation handling that would benefit from LaTeX-based representation in future implementations.

Table 2. Performance comparison of LLMs across three evaluation metrics in question generation. **Bold** and underline indicate the highest and second-highest scores

		Relevance		DOK alignment		Appropriateness		PINC
		DOK	DOK+RAG	DOK	DOK+RAG	DOK	DOK+RAG	
GPT-4o	Level 1	**0.85**	0.79	0.81	**0.78**	0.80	0.91	**0.94**
	Level 2	0.81	**0.85**	0.30	0.73	0.72	0.90	0.93
	Level 3	0.79	0.81	0.52	0.71	0.74	**0.95**	0.92
	Level 4	0.78	0.80	0.29	0.61	0.70	0.82	0.90
	Average	0.81	0.81	0.48	0.71	0.74	0.90	0.92
Deepseek-V3	Level 1	0.84	0.82	**0.80**	0.68	**0.82**	0.89	**0.94**
	Level 2	0.80	0.84	0.75	0.72	0.74	0.88	0.92
	Level 3	0.77	0.80	0.63	0.78	0.72	0.85	0.90
	Level 4	0.66	0.71	0.39	0.59	0.71	0.80	0.92
	Average	0.77	0.79	0.64	0.69	0.75	0.86	0.92
Gemini-1.5	Level 1	0.80	0.75	0.75	0.77	0.76	0.85	0.92
	Level 2	0.75	0.78	0.49	0.68	0.68	0.82	0.90
	Level 3	0.72	0.74	0.52	0.75	0.69	0.88	0.89
	Level 4	0.63	0.73	0.46	0.58	0.65	0.75	0.92
	Average	0.73	0.75	0.56	0.70	0.70	0.82	0.91

5 Conclusion and Discussion

Through two interconnected studies, our research provides insights into GenAI's potential in math education. *Study 1* revealed that while GenAI can produce relevant questions using Bloom's Taxonomy, it struggles with higher cognitive levels and tends to generate plausible yet incorrect information when given more context. Building on these findings, *Study 2* introduced the QG-DOK framework, integrating Webb's DOK levels with RAG. By leveraging resources that educators are already familiar with, both quality and depth of generated questions improved. Our findings support earlier research suggesting that AI can effectively generate educational content but requires careful design to ensure cognitive depth and relevance [6]. The improvements in *Study 2* address limitations identified in *Study 1*, particularly in generating deeper thinking questions.

6 Limitations

Although our results highlight the potential of GenAI in educational content creation, a few limitations remain. Bloom's taxonomy, though often used to categorize cognitive understanding, has been critiqued for oversimplifying the interconnected nature of learning [8], reflecting concerns in our study. Switching to DOK to inform the AI for question generation helps alleviate this to a degree to meet our exploratory goals, but it is still possible that simply describing each level is not enough instruction for the AI to create an appropriate question. Future work could explore how educators design problems to derive a step-by-step framework for AI-driven question generation.

We also acknowledge the constraints of our data input and evaluation methods. We used a single reference content as context for the AI in each study, and a single human evaluator. Given the exploratory nature of this study, though, it was important to limit the parameters to best explain variations in the quality. Moving forward, researchers can test and evaluate the output more systematically, given that our findings have highlighted the capabilities (and limitations) of current genAI. As such, our work serves as an important preliminary step in advancing question generation through AI for advanced math.

Acknowledgements. This work was supported by an exploratory Interdisciplinary Research Development award from McCAIS at McGill University.

References

1. Anderson, L.W., Krathwohl, D.R.: A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, complete Inc, Addison Wesley Longman (2001)
2. Antu, S.A., Chen, H., Richards, C.K.: Using LLM (large language model) to improve efficiency in literature review for undergraduate research. LLM@ AIED, pp. 8–16 (2023)
3. Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H.Y., Hussain, A.: Educational data mining to predict students' academic performance: a survey study. Educ. Inf. Technol. **28**(1), 905–971 (2023)
4. De Kuthy, K., Kannan, M., Ponnusamy, H.S., Meurers, D.: Towards automatically generating questions under discussion to link information and discourse structure. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 5786–5798 (2020)
5. DeepSeek: DeepSeek-V3 technical report (2024). https://arxiv.org/abs/2412.19437
6. Denny, P., et al.: Computing education in the era of generative AI. Commun. ACM **67**(2), 56–67 (2024)
7. Elkins, S., Kochmar, E., Cheung, J.C., Serban, I.: How teachers can use large language models and bloom's taxonomy to create educational quizzes. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 23084–23091 (2024)
8. Fadul, J.A.: Collective learning: applying distributed cognition for collective intelligence. Int. J. Learn. **16**(4) (2009)

9. Frieder, S., et al.: Mathematical capabilities of ChatGPT. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
10. Gao, Y., et al.: Retrieval-augmented generation for large language models: a survey. arXiv preprint arXiv:2312.10997 (2023)
11. Gemini: Gemini 1.5: unlocking multimodal understanding across millions of tokens of context (2024). https://arxiv.org/abs/2403.05530
12. Guo, S., Liao, L., Li, C., Chua, T.S.: A survey on neural question generation: methods, applications, and prospects. arXiv preprint arXiv:2402.18267 (2024)
13. Hojeij, Z., Kuhail, M.A., ElSayary, A.: Investigating in-service teachers' views on ChatGPT integration. Interact. Technol. Smart Educ. (2024)
14. Irvine, J.: Taxonomies in education: overview, comparison, and future directions. J. Educ. Dev. **5**(2), 1 (2021)
15. Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. **30**, 121–204 (2020)
16. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)
17. Li, H., et al.: Bringing generative AI to adaptive learning in education. arXiv preprint arXiv:2402.14601 (2024)
18. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-Eval: NLG evaluation using GPT-4 with better human alignment. arXiv preprint arXiv:2303.16634 (2023)
19. Lo, C.K.: What is the impact of ChatGPT on education? A rapid review of the literature. Educ. Sci. **13**(4), 410 (2023)
20. McCarthy, J.: Artificial intelligence, logic, and formalising common sense. Machine Learning and the City: Applications in Architecture and Urban Design, pp. 69–90 (2022)
21. OpenAI: GPT-4o system card (2024). https://arxiv.org/abs/2410.21276
22. Reid, M., et al.: Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
23. Scaria, N., Dharani Chenna, S., Subramani, D.: Automated educational question generation at different bloom's skill levels using large language models: strategies and evaluation. In: International Conference on Artificial Intelligence in Education, pp. 165–179. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-64299-9_12
24. Shute, V.J., Psotka, J.: 19. intelligent tutoring systems: past, present, and future (1994)
25. Wang, D., Tao, Y., Chen, G.: Artificial intelligence in classroom discourse: a systematic review of the past decade. Int. J. Educ. Res. **123**, 102275 (2024). https://doi.org/10.1016/j.ijer.2023.102275, https://www.sciencedirect.com/science/article/pii/S0883035523001386
26. Wang, N., Johnson, M.: AI education for K-12: connecting AI concepts to high school math curriculum. In: Workshop on Education in Artificial Intelligence K-12, 28th International Joint Conference on Artificial Intelligence. https://par.nsf.gov/biblio/10440200
27. Wang, X., et al.: Searching for best practices in retrieval-augmented generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 17716–17736 (2024)
28. Wang, X., Fan, S., Houghton, J., Wang, L.: Towards process-oriented, modular, and versatile question generation that meets educational needs. arXiv preprint arXiv:2205.00355 (2022)

29. Wang, Z., Lan, A.S., Nie, W., Waters, A.E., Grimaldi, P.J., Baraniuk, R.G.: QG-Net: a data-driven question generation model for educational content. In: Proceedings of the Fifth Annual ACM Conference on Learning at Scale, pp. 1–10 (2018)
30. Webb, N.L.: Depth-of-knowledge levels for four content areas. Lang. Arts **28**(March), 1–9 (2002)
31. Wu, H., Hui, W., Chen, Y., Wu, W., Tu, K., Zhou, Y.: Conic10K: a challenging math problem understanding and reasoning dataset. arXiv preprint arXiv:2311.05113 (2023)
32. Xiong, Z., et al.: A review of data mining in personalized education: current trends and future prospects. Front. Digit. Educ. **1**(1), 26–50 (2024)
33. Zhu, Y., Li, J., Li, G., Zhao, Y., Jin, Z., Mei, H.: Hot or cold? Adaptive temperature sampling for code generation with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 437–445 (2024)

Enhanced Interpretable Knowledge Tracing *for* Students' Performance Prediction with Human-understandable Feature Space

Sein Minn[1,2(✉)] and Roger Nkambou[1]

[1] AI Research Center, Université du Québec à Montréal, Montreal, Canada
{sein.minn,nkambou.roger}@uqam.ca
[2] Department of Information and Communication Technologies (ICT), Asian Institute of Technology, Khlong Nueng, Thailand

Abstract. Knowledge Tracing (KT) plays a central role in assessing students' skill mastery and predicting their future performance. While deep learning-based KT models achieve superior predictive accuracy compared to traditional methods, their complexity and opacity hinder their ability to provide psychologically meaningful explanations. This disconnect between model parameters and cognitive theory poses challenges for understanding and enhancing the learning process, limiting their trustworthiness in educational applications. To address these challenges, we enhance interpretable KT models by exploring human-understandable features derived from students' interaction data. By incorporating additional features, particularly those reflecting students' learning abilities, our enhanced approach improves predictive accuracy while maintaining alignment with cognitive theory. Our contributions aim to balance predictive power with interpretability, advancing the utility of adaptive learning systems. The source code and data for replicating our experiments are available at https://github.com/Simon-tan/IKT.git.

Keywords: Knowledge tracing · student modeling · Bayesian networks · causality · feature engineering

1 Introduction

Adaptive learning systems have revolutionized modern education by dynamically adjusting to each learner's requirements in real time [5,8]. However, implementing such personalized learning at scale presents significant challenges, where intelligent tutoring systems must effectively assess and adapt to the learning trajectories of tens or even hundreds of thousands of students [15].

KT models are the backbone of intelligent tutoring systems, driving personalized content delivery and feedback. Artificial Intelligence (AI) plays a critical role in this process, with KT models using machine learning techniques to infer

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 423–430, 2025.
https://doi.org/10.1007/978-3-031-98462-4_53

a student's conceptual and procedural knowledge based on their task performance [1,9]. Among KT methods, Deep Knowledge Tracing (DKT) [19] stands out for its use of Neural Networks (NNs), which model student learning over time and deliver superior predictive performance, highlighting the transformative impact of deep learning in education [7,10,19].

Bayesian Knowledge Tracing (BKT) [1], an earlier and interpretable KT model, utilizes a Hidden Markov Model to track a student's knowledge state. In contrast, DKT overcomes this limitation by using NNs to model cross-skill dynamics, improving predictive accuracy. Despite its improved performance, DKT's reliance on complex hidden layers and large numbers of parameters reduces its interpretability, making it less transparent and trustworthy for educators. The lack of interpretability in models like DKT highlights the need for methods that combine accurate predictions with clear, actionable insights. Toward a transparent, psychologically meaningful model, we aim to extend IKT [15], focusing on simplicity while maintaining comparable predictive performance.

2 Exploring Human-Understandable Feature Space

Drawing inspiration from the Interpretable Knowledge Tracing (IKT) model [15], we propose an Enhanced Interpretable Knowledge Tracing (EIKT) model that introduces a richer and more interpretable feature space through advanced feature extraction. The IKT model predicts a student's future responses using three core latent features: individual skill ID, skill mastery, and problem difficulty. Beyond this foundation, EIKT incorporates enhanced ability profiling to explore the interplay between item difficulty and skill correctness.

Skill Mastery. In a typical BKT-based learning environment [1], a student's skill mastery is dynamically updated with each item's response. BKT operates using four key probabilities: **Initial Knowledge** ($P(L_0)$): The probability that a student already knows the skill before any practice; **Learning** ($P(T)$): The probability that the student will transition from not knowing the skill to knowing it after an opportunity to practice; **Guess** ($P(G)$): The probability that the student correctly answers an item despite not having mastered the skill; **Slip** ($P(S)$): The probability that the student incorrectly answers an item despite having mastered the skill.

The probability that a student knows the skill at any given time t is continuously updated based on their observed outcomes, recorded as either successful (Obs = 1) or unsuccessful (Obs = 0). This iterative update mechanism allows BKT to provide a dynamic and data-driven assessment of a student's evolving knowledge state.

$$P(L_t|1) = \frac{P(L_t)(1 - P(S))}{P(L_t)(1 - P(S)) + (1 - P(L_t))P(G)} \tag{1}$$

$$P(L_t|0) = \frac{P(L_t)P(S)}{P(L_t)P(S) + (1 - P(L_t))(1 - P(G))} \tag{2}$$

$$P(L_{t+1}) = P(L_t|obs) + (1 - P(L_t|obs))P(T) \tag{3}$$

$$\textbf{skill mastery}(s_t) = \delta(P(L_t), s_t) \tag{4}$$

In this context, $\delta(P(L_t), s_t)$ represents a function that maps the mastery level of a specific skill s_t at the current timestamp, based on the student's cumulative interaction history.

Problem Difficulty serves as a distinct feature for predicting student performance in previous studies [10,15,17]. The difficulty of a problem p_j is determined on a scale of 1 to 10. Problem difficulty $\mathbf{Pd}(P_j)$ is calculated as:

$$\mathbf{Pd}(p_j) = \begin{cases} \delta(p_j) & \text{if } |N_j| \geq 4 \\ 5 & \text{otherwise} \end{cases} \tag{5}$$

$$\delta(p_j) = \left\lfloor \frac{\sum_i^{|N_j|} O_i(p_j)}{|N_j|} \cdot 10 \right\rfloor \tag{6}$$

and where p_j represents the j^{th} problem; N_j is the set of students who attempted problem p_j; $O_i(p_j)$ the outcome of the first attempt from student i to problem p_j, 1 if successful, 0 otherwise. $\delta(p_j)$ is a function that maps the *average success rate* of problem p_j onto 10 levels.

Ability Profile: Learners Trajectory. Approaches like DKT-DSC and IKT [15,16] introduced the use of cluster IDs as an ability profile, first applied in deep learning-based models and later extended to Tree-Augmented Naive Bayes (TAN). While clustering offers a cost-efficient solution for deep learning-based models, it may sacrifice some predictive accuracy.

To address these challenges, we reformulate the *ability profile* by directly introducing two types based on problem difficulty and skill correctness. This enhancement improves the model's ability to track and adapt to student learning over time.

- **Skill Correctness-based Ability Profile** is calculated to represent a student's cumulative ability in mastering the skill associated with the current problem. This measure provides valuable insights into a student's skill-specific strengths and weaknesses, enabling personalized learning interventions.

$$sr_profile = R(x_j)_{1:t} = \sum_{t=1}^{z} \frac{(x_{jt})}{|N_{jt}|}, \tag{7}$$

  where x_{jt} is the total correct outcome of attempt of skill x_j of the problem being correctly answered at timestamp t; $|N_{jt}|$ is the total number of practice attempts of skill x_j up to timestamp t;
- **Item Difficulty-based Ability Profile** is calculated to represent a student's cumulative ability in handling problems of varying difficulty levels.

This metric helps in understanding how well a student performs across different difficulty tiers and provides a basis for adaptive instructional strategies tailored to their proficiency.

$$df_profile = R(\mathbf{Pd}(p_j))_{1:t} = \sum_{t=1}^{t} \frac{(\mathbf{Pd}(p_{jt}))}{|N_{jt}|}, \tag{8}$$

where $\mathbf{Pd}(p_{jt})$ is the total correct outcome of the attempts to problem difficulty level like current problem (p_j) being correctly answered until timestamp t; $|N_{jt}|$ is the total number of practice attempts of problem difficulty level like current problem $\mathbf{Pd}(p_j)$ up to timestamp t.

Instead of utilizing a cluster ID as an ability profile, as done in DKT-DSC and IKT, we employed a more refined ability profile derived from two key dimensions: item difficulty and skill correctness. This approach captures the cumulative ability of a student by leveraging their performance on problems categorized by difficulty levels and specific skills.

3 Predicting Student Performance

Inspired by Interpretable Knowledge Tracing (IKT), we employ the Tree-Augmented Naive Bayes (TAN) method [3,13]. TAN extends the Naive Bayes network by designating the class node representing problem correctness as the root, causally connected to evidence nodes such as skill ID, skill mastery, ability profile, and problem difficulty. As shown in Fig. 1, this approach maintains the directed acyclic graph (DAG) structure [4,11,12], ensuring efficient reasoning while retaining interpretability.

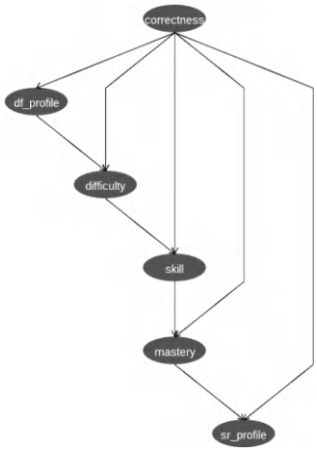


Fig. 1. Tree Structure of TAN for EIKT.

EIKT performs inference by using five meaningful features f_t: skill ID (skill), skill mastery (mastery), skill correctness-based ability profile (sr_profile), item Difficulty-based ability profile (df_profile), and problem difficulty (difficulty) as evidenced at the current timestamp t (for brevity we omit indexing all terms by student i, s_t, P_j in Fig. 1 and Eq. 9).

$$P(correctness_t = y|f_t) = \frac{P(y)P(f_t|y)}{\sum_{y'} P(y')P(f_t|y')} \tag{9}$$

$$\begin{aligned}\text{where } P(f_t|y) =& P(\text{df_profile}|y)P(\text{diffiuclty}|y, \text{df_profile})P(\text{skill}|y, \text{diffiuclty}) \\ & P(\text{mastery}|y, \text{skill})P(\text{sr_profile}|y, \text{mastery})\end{aligned}$$

The class node (correctness) represents the predicted probability that the student would correctly answer the problem with the associated skill as described in Fig. 1. We can achieve interpretation via the conditional probability tables of each node with their causal links [14].

The primary computational cost arises from extracting skill mastery using Bayesian Knowledge Tracing (BKT), while training the TAN model takes only a few seconds. Consequently, the overall training cost of our approach remains effectively equivalent to that of BKT, with the TAN modeling cost being negligible. Unlike IKT, our approach eliminates profile clustering, thereby reducing the computational expenses associated with clustering. This allows EIKT to significantly lower model complexity, training time, and computational costs compared to deep learning models while providing causal explanations for greater interpretability.

4 Experiments

We compare the next problem student performance prediction of our model with well-known KT models mentioned above: BIRT [20], BKT [1], PFA [18], DKT [19], DKT-DSC [16], DKVMN [21], AKT-R [6] and IKT [15]. To validate the proposed model, we conducted experiments using three public datasets from two distinct computer-based tutoring scenarios in educational settings: ASSISTments: 1) ASSISTments 2009–2010 (Skill Builder) 2) ASSISTments 2012–2013 [2][1] and Cognitive Tutor Algebra 2005–2006 (KDD Cup 2010) dataset on PSLC DataShop[2]. For all datasets, only the first correct attempts to original problems are considered in our experiment. We remove data with missing values for skills and problems with duplicate records. We used five-fold cross-validation to evaluate all datasets, splitting each fold into 80% training and 20% testing students. To the best of our knowledge, these are among the most well known publicly available knowledge tracing datasets.

[1] https://sites.google.com/site/assistmentsdata/.
[2] https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp.

Results. The EIKT structure results in a student model with better explanation with causal relations and higher predictive performance. The results in Tables 1 demonstrate that EIKT outperforms significantly the well-known KT models in all tested datasets. When we compare EIKT with our second-best performer IKT and AKT-R (scores are underlined in Table 1). EIKT shows slightly better performance than IKT but has superior performance compared to other models tested in our experiments. The improvement is ranging from 0.12% to 0.78% in terms of AUC. When we compare in terms of RMSE, it still keeps almost the same performance as IKT but shows better performance than other models. So EIKT shows better performance than any other method in both AUC and RMSE (except AKT-R on the ASS-12 dataset).

Table 1. Average AUC and RMSE result for all tested models across datasets.

Type	Datasets	BIRT	PFA	BKT	DKT	DKT-DSC	DKV-MN	AKT-R	IKT	**EIKT**
AUC	ASS-09	0.750	0.701	0.651	0.721	0.735	0.710	0.767	0.797	**0.801**
	ASS-12	0.744	0.672	0.623	0.713	0.721	0.707	**0.777**	0.767	0.773
	Algebra	0.812	0.754	0.642	0.784	0.792	0.780	0.845	0.845	**0.848**
	Average	0.768	0.709	0.638	0.739	0.749	0.732	0.796	0.802	**0.806**
RMSE	ASS-09	0.440	0.454	0.471	0.450	0.434	0.451	0.423	0.412	**0.411**
	ASS-12	0.441	0.440	0.510	0.430	0.427	0.430	**0.409**	0.412	0.412
	Algebra	0.374	0.391	0.440	0.380	0.373	0.380	0.357	0.358	**0.356**
	Average	0.418	0.428	0.473	0.420	0.411	0.420	0.396	0.395	**0.394**

5 Conclusion

Unlike the cluster-ID used as an ability profile in IKT, we introduce two enhanced ability profiles that improve student performance prediction while preserving model interpretability. By incorporating these features into the Tree-Augmented Naïve Bayes (TAN) structure with explicit causal relationships, we develop a Knowledge Tracing (KT) model that achieves superior predictive accuracy without relying on extensive parameters or complex architectures.

We propose a causal probabilistic student model, EIKT, which integrates five key features: skill ID, student skill mastery (probability of mastering a skill), two ability profiles (based on problem difficulty and skill correctness), and problem difficulty. Unlike deep learning-based KT models that rely on hidden states, numerous parameters, and complex architectures, EIKT leverages explicit, interpretable features derived from skills, problems, and students, enhancing predictive accuracy and causal interpretability. By eliminating profile clustering, EIKT further reduces the computational expenses required in IKT, making it even more efficient. Experiments on three public datasets demonstrate that EIKT outperforms established KT models, including deep learning

approaches, while maintaining significantly lower computational costs. The integration of carefully engineered features captures greater data variance, while the TAN framework provides causal insights. By incorporating skill mastery, ability profiles, and problem difficulty, EIKT effectively accounts for student variability, enabling more personalized and accurate predictions.

While our primary focus is on prediction accuracy, we also highlight the importance of understanding the causal effects of explainable features to inform adaptive instruction in personalized learning systems. Future work will explore knowledge acquisition and learning behaviors to further optimize human learning in evolving educational environments.

References

1. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User Model. User-Adap. Inter. **4**(4), 253–278 (1994)
2. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User Model. User-Adap. Inter. **19**(3), 243–266 (2009)
3. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Mach. Learn. **29**(2), 131–163 (1997)
4. Fu, S., Minn, S., Desmarais, M.C.: Towards the efficient recovery of general multi-dimensional Bayesian network classifier. In: Perner, P. (ed.) MLDM 2014. LNCS (LNAI), vol. 8556, pp. 16–30. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08979-9_2
5. Fu, S., Zhang, Y., et al.: On the recommender system for university library. International Association for Development of the Information Society (2013)
6. Ghosh, A., Heffernan, N., Lan, A.S.: Context-aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2330–2339 (2020)
7. Minn, S.: BKT-LSTM: efficient student modeling for knowledge tracing and student performance prediction. arXiv preprint arXiv:2012.12218 (2020)
8. Minn, S.: AI-assisted knowledge assessment techniques for adaptive learning environments. Comput. Educ. Artif. Intell. **3**, 100050 (2022)
9. Minn, S., Desmarais, M.C., Fu, S.K.: Refinement of a Q-matrix with an ensemble technique based on multi-label classification algorithms. In: Verbert, K., Sharples, M., Klobučar, T. (eds.) Adaptive and Adaptable Learning: 11th European Conference on Technology Enhanced Learning, EC-TEL 2016, Lyon, France, September 13-16, 2016, Proceedings, pp. 165–178. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-45153-4_13
10. Minn, S., Desmarais, M.C., Zhu, F., Xiao, J., Wang, J.: Dynamic student classiffication on memory networks for knowledge tracing. In: Yang, Q., Zhou, Z.-H., Gong, Z., Zhang, M.-L., Huang, S.-J. (eds.) Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part II, pp. 163–174. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-16145-3_13
11. Minn, S., Fu, S., Lv, T.: Algorithm for exact recovery of Bayesian network for classification. Appl. Res. Comput. **33**(05), 1327–1334 (2016)
12. Minn, S., Fu, S.K.: Accelerating structure learning of Bayesian network. Comput. Sci. (2016)

13. Minn, S., Fu, S., Desmarais, M.C.: Efficient learning of general Bayesian network classifier by local and adaptive search. In: 2014 International Conference on Data Science and Advanced Analytics (DSAA), pp. 385–391. IEEE (2014)
14. Minn, S., Shunkai, F.: Fast & efficient learning of Bayesian networks from data: knowledge discovery and causality. In: 2023 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 966–975. IEEE (2023)
15. Minn, S., Vie, J.J., Takeuchi, K., Kashima, H., Zhu, F.: Interpretable knowledge tracing: simple and efficient student modeling with causal relations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 12810–12818 (2022)
16. Minn, S., Yu, Y., Desmarais, M.C., Zhu, F., Vie, J.J.: Deep knowledge tracing and dynamic student classification for knowledge tracing. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 1182–1187. IEEE (2018)
17. Minn, S., Zhu, F., Desmarais, M.C.: Improving knowledge tracing model by integrating problem difficulty. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1505–1506. IEEE (2018)
18. Pavlik, P.I., Cen, H., Koedinger, K.R.: Performance factors analysis–a new alternative to knowledge tracing. In: 14th International Conference on Artificial Intelligence in Education (2009)
19. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, pp. 505–513 (2015)
20. Wilson, K.H., Karklin, Y., Han, B., Ekanadham, C.: Back to the basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation. arXiv preprint arXiv:1604.02336 (2016)
21. Zhang, J., Shi, X., King, I., Yeung, D.Y.: Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th International Conference on World Wide Web, pp. 765–774 (2017)

So What? Unpacking the Complexities in Collaborative Problem Solving with AI-Augmented Sense-Making

Seiyon M. Lee[1(✉)], Hongming Li[1], Shan Zhang[1], Zirui Zhong[2], Ji-Eun Lee[2], and Anthony F. Botelho[1]

[1] University of Florida, Gainesville, FL 32601, USA
{leeseiyon,hli3,zhangshan,botelho}@ufl.edu
[2] Worcester Polytechnic Institute, Worcester, MA 01609, USA
{zzhong4,jlee13}@wpi.edu

Abstract. Visual representations of data in dashboards have the potential to translate learning process data into actionable insights for educational stakeholders. This is particularly useful for teachers when their capacity to monitor students is further strained in orchestrating complex activities, such as paired collaborative problem-solving. Despite great interest in dashboards over the past decade, their effectiveness has been under scrutiny with concerns about the cognitive overload for users with limited data literacy, thus questioning their practical utility in supporting decision-making and reflective practices. In this paper, we present a functional prototype for K-12 teachers and demonstrate the use of generative AI to unpack rich details of collaborative problem-solving processes in mathematics. Through a case study, we share our human-centered approach to design, ensuring that AI-augmented insights are not only interpretable and actionable, but centered around authentic needs. Based on the insights and experiences in the co-design session, our tool includes features, such as a contextualized description of students' breakthrough and struggle moments. This work contributes to ongoing efforts to support users' sense-making process of data visualizations in dashboards by leveraging generative AI.

Keywords: Learning Analytics · Dashboard · CSCL · Co-design · Human-in-the-loop · Generative Artificial Intelligence · Large Language Models

1 Introduction

Over the past decade, dashboards have become an instrumental part of educational technology, helping translate learning process data into actionable insights for educators, learners, and other stakeholders [7,15,17,18]. A key element of dashboard design is visual representations of data in a way that highlights meaningful patterns across students' learning processes [16]. Dashboards and data visualization thereof are particularly useful for teachers in facilitating and orchestrating complex activities and the use of digital tools, where it is often difficult to monitor classrooms that are likely working on different parts at any given time

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 431–438, 2025.
https://doi.org/10.1007/978-3-031-98462-4_54

[6,17]. Dashboards can not only be used in-class for greater awareness and in situ decision-making, but they can also capture and document fine-grained traces of student progress [9,12].

Despite the potential, the effectiveness of dashboards remains questionable [7]. One of the known limitations is the cognitive load on teachers' ability to understand visually encoded information about students' learning processes to be able to use dashboards to effectively orchestrate a classroom [13]. To better support teacher's real-time sense-making, recent work has started to explore AI in dashboard designs [5,20]. One of the recent examples included integrating AI chatbots into dashboards, allowing users to ask for contextualized explanations of data visualization [20]. However, few studies have considered a human-centered approach for AI-augmented dashboards, raising greater concerns for K–12 education contexts where there is greater need for support in orchestrating computer-supported collaborative leaning activities.

In this paper, we present a functional prototype, specifically for K-12 teachers, and demonstrate the use of generative AI to unpack rich details of collaborative problem-solving processes in mathematics. Through a case study, we share our human-centered approach to design and development to ensure that AI-augmented insights are not only interpretable and actionable but centered around authentic needs. Specifically, we report on how we extracted key insights by grounding ourselves in the teacher's perspectives directly and indirectly in Sect. 1. Then, we discuss how generative AI was leveraged to provide detailed accounts of paired collaborative learning processes in math Sect. 3. In the last part Sect. 4, we illustrate the key features of our tool.

2 Grounding in the Teacher Perspective

In dashboard design, taking a human-centered design approach has been promising to distill the most relevant information from big process data [1,2,12]. In our study, teacher input guided our early stages of the design, revealing insights for greater support in data sense-making than we previously imagined. Specifically, we worked with a math teacher at a research development school in the southeastern U.S. who volunteered out of genuine interest in providing new learning opportunities for her students[1]. Her willingness to integrate it into her classroom distinguishes our work from studies that prioritize technological proficiency [14].

Two researchers visited the school for a week to design and lead onboarding sessions for GraspableMath (GM), a math notation tool and classroom-based platform where mathematical objects are reified as physical objects that students can manipulate and transform on screen. GM was chosen because its developers had recently released a collaborative whiteboard prototype that allows students to interact with the same math content simultaneously like Google Docs (see [19] for details). Each researcher then led a collaborative problem-solving activity using GM in the classroom for Grade 6 ($n = 43$), who had consented to

[1] This study was conducted under IRB-approved protocol #IRB202301185 at the University of Florida.

the study. The teacher was present in both settings. This structure allowed the teacher to focus on different aspects of collaborative problem-solving in the classroom while the researchers facilitated the activities.

After a week of collaborative activity, two researchers and the teacher engaged in an open reflection on collaborative activities as facilitators and observers, we asked: "*What are some superpowers that you would find useful specifically during a collaborative math activity in which students are interacting with GM?*" While this question may be effective in eliciting needs from participants [6], in our case, the question was less effective because teachers generally reported positive observations from the collaborative activity, with greater focus on the novelty of collaborative activities in math classrooms, as most platforms have students working in isolation (e.g., MathSpace, FluidityMath, MathDragon, DudaMath). To further guide the discussion, we asked: *"Which aspects of collaboration do you find empowering for you or engaging for your students? How would you rank them in order of importance, and why?"* This question revealed interesting insights about teacher's priorities in collaborative activities, which focused on (1) combined work from both partners, followed by (2) higher-level thinking, (3) engagement, and (4) feedback.

We then shared a data visualization created from demo data. To gauge the interpretability and actionability of visualization, we asked "*What did you learn from this visualization, and what would you do?*" Although the visualization provided a general summary of collaborative processes (see [10] for more details), it offered only limited insight into understanding problem-solving processes. For example, the teacher noticed that one student made more contributions on the whiteboard than the other, but it was not clear how individual actions were relevant, whether students were building on each other's work or working independently. While there are examples of on-demand AI interpretation using chatbot interfaces (e.g., VizChat [20]), our interaction with the teacher, as well as our experiences in facilitating collaborative activities, revealed that it may be often unrealistic and resource-intensive to come up with questions to ask. With concrete insights on what teachers need and would likely benefit from, we developed an AI-augmented dashboard that not only visualizes the complex collaborative math problem-solving process but also provides detailed summaries of collaborative learning, including mathematical concepts that students engaged with or struggled with.

3 Building AI-Augmented Sense-Making Tool

Our AI-augmented analytics pipeline is focused on translating collaborative problem-solving log data into meaningful information that supports teachers in better understanding student learning and decision-making. Moreover, language models suffer from contextual windows while handling large-scale interaction logs, which can influence their capacity to analyze complete session data. As illustrated in Fig. reffig:system-architecture, our system solves these issues by

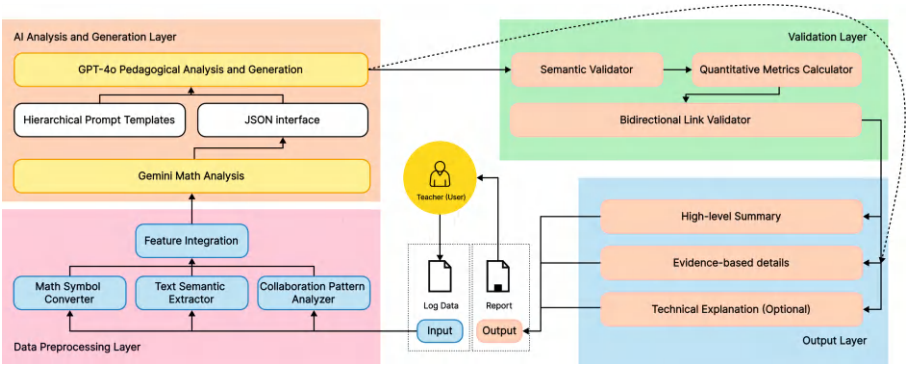


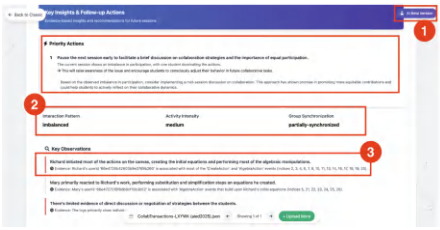
Fig. 1. A system architecture of our AI-augmented analysis, involving three key layers: (1) Data Preprocessing for feature extraction and integration, (2) AI Analysis and Generation with dual-provider architecture, and (3) Validation and Output for ensuring quality and generating multi-level explanations.

means of a well-designed three-phase architecture combining targeted data preprocessing, context-aware artificial intelligence analyses, and pedagogically based insights generating (Fig. 1).

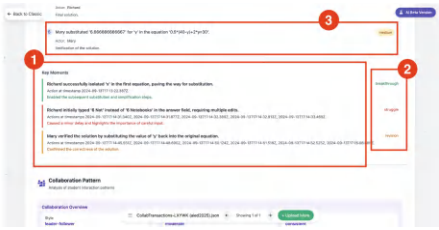
Knowledge Pre-processing and Feature Extraction. The pre-processing stage involved a novel feature extraction approach, which is well-suited for collaborative learning in math. Using a domain-specific transformation pipeline, the approach can greatly reduce data complexity while maintaining pedagogically meaningful patterns. For instance, both the symbolic process and the underlying reasoning processes are preserved when handling the mathematical solution sequence "$2x + 3 = 7 \rightarrow 2x = 4 \rightarrow x = 2$." Usually reducing the data by 70–80%, this selective retention method helps retain important information, such as mathematical reasoning strategies.

Dual-Provider AI Analysis Pipeline. The dual-provider design focuses on the complementary strengths of multiple AI models through a sequential-parallel processing pipeline [3]. Gemini-2.0-flash-exp first analyzes mathematical content to identify key patterns, such as inference processes, misconceptions, and strategies. Along with contextual metadata, this is passed to GPT-4o, which is optimized for generating pedagogically relevant narratives, or detailed explanations. Examples include how a student interacted with other students for collaborative problem-solving, and how students used a specific strategy (e.g. distributive property), indicating a breakthrough moment.

We also have **Multi-stage Validation and Output Generation**, a pipeline designed to acknowledge ambiguity and present multiple interpretations rather than making definitive claims when encountering edge cases, making it transparent for the users to consider uncertainties in AI-augmented analysis for certain unique cases.



(a) AI-Generated Teaching Recommendations: Insights into student collaboration patterns. The numbered highlights indicate key observations.



(b) Problem-Solving Analysis: Breakdown of key problem-solving moments. Red boxes highlight critical points for teacher intervention.

Fig. 2. AI-augmented Dashboard Insights: Visualizing student collaboration and problem-solving dynamics.

4 Case Study of Functional Prototype

We present a case study based on real-world collaborative problem-solving data from college students, demonstrating how our AI-augmented feature on the current collaborative learning dashboard[2] helps teachers make sense of student data in a meaningful way, such as group dynamics, and key problem-solving strategies. Although the reports are created for each group, we examine two groups of students who engaged in the same activity, and compare similarities and differences in their collaboration. All names displayed in images and discussions are pseudonyms, and the data have been anonymized to protect student privacy.

AI-Augmented Summary of Collaboration. Two groups of students were tasked with solving a system of linear equations in a collaborative whiteboard. In one group, the AI identified an imbalanced interaction pattern where *Richard initiated most of the actions on the canvas, creating the initial equations and performing most of the algebraic manipulations,"* while his partner *primarily reacted to Richard's work, performing substitution and simplification steps on equations he created."* This pattern resulted in limited reciprocal discussion, with the AI detecting *"no explicit communication events recorded."*

The second group exhibited a more distributed workload, though with occasional shifts in participation dominance. The AI noted that *"initial equation setup was more evenly divided, with both students contributing independently before engaging in substitution and simplification collaboratively."* However, the report also indicated that *"one student took a leading role in verification steps, while the other followed rather than questioning or contributing alternative approaches."*

Highlights in Collaborative Problem-Solving. The AI system flagged key problem-solving moments where teacher intervention could be beneficial. As seen

[2] The dashboard is now publicly available at https://gmdashboard.viablelab.org/ai/, featuring demo with anonymized student data.

in Fig. 2b, the system identified crucial points such as a **breakthrough** when *"Richard successfully isolated 'x' in the first equation, paving the way for substitution."* Additionally, areas of **struggle** were highlighted, including multiple failed text edits when Richard initially mistyped a variable name, causing minor delays. Finally, the AI flagged **revision** steps where Mary validated Richard's solution by substituting values back into the original equation, confirming the correctness. The highlights in Fig. 2a indicate quick summaries of group synchronization, activity intensity, and key observations to help teachers quickly assess collaboration quality.

5 Discussion and Conclusion

In this paper, we present an AI-augmented dashboard prototype that is inspired by a need for better support for the sense-making of students' collaborative problem-solving processes beyond data visualization. Our case study showed the potential of leveraging AI-augmented explanatory summaries from log data, which enabled the exploration of the fine-grained, action level log data without video or audio data.

In the design of AI-powered tools, human-in-the-loop is not only useful for validation and improving accuracy but also to spark significant conversations about trust and transparency. For example, existing research-practice partnerships, such as our own context, allow us to discuss with teachers fundamental details, such as how student performance should be measured and represented in the visual displays. This can help demystify a black-box nature of the complex algorithms that hide how summaries or analyses are undertaken. Establishing a shared understanding that data is less neutral, objective measurement of learning but one of many representations [8] of what teachers might already know from their experience would help teachers to consider AI components in dashboards as possible supporting evidence to drive their decisions when monitoring complex tasks such as orchestration of collaborative activities.

This work has several limitations that future work should address to further validate and enhance the practical utility of the tool. First, our work is limited to exploring relatively only a portion of the entire collaborative problem-solving process or key takeaways from data. Future work can explore more narrative, story-telling approaches to data communication [4,11] to account for users' varying goals and needs. Also, any efforts to integrate generative AI for analysis raise questions of accuracy and require external validation, which can introduce similarly resource-intensive processes in the design. Future research needs to assess the accuracy, reliability, and informativeness through practical use cases based on research-practice partnerships. Addressing these areas is important to fulfill the untapped potential of dashboards to go beyond instrument, or by-product but rather a bridge iterating between research and practice.

Acknowledgments. We would like to thank the National Science Foundation (#2331379), Bill and Melinda Gates Foundation, Schmidt Futures, OpenAI, and Google DeepMind.

References

1. Aleven, V., Xhakaj, F., Holstein, K., McLaren, B.M.: Developing a teacher dashboard for use with intelligent tutoring systems. Technology **34**, 44 (2010)
2. Casamayor, A., Amandi, A., Campo, M.: Intelligent assistance for teachers in collaborative e-learning environments. Comput. Educ. **53**(4), 1147–1154 (2009)
3. Deroy, A., Maity, S.: Exploring the mathematical reasoning capabilities of Gemini. Authorea Preprints (2024)
4. Fernandez Nieto, G.M., Kitto, K., Buckingham Shum, S., Martinez-Maldonado, R.: Beyond the learning analytics dashboard: alternative ways to communicate student data insights combining visualisation, narrative and storytelling. In: LAK22: 12th International Learning Analytics and Knowledge Conference, pp. 219–229 (2022)
5. Gobert, J.D., Dickler, R., Adair, A.: Using an AI-based dashboard to help teachers support students' learning progressions for science practices. In: Handbook of Research on Science Learning Progressions, pp. 488–498. Routledge (2024)
6. Holstein, K., Hong, G., Tegene, M., McLaren, B.M., Aleven, V.: The classroom as a dashboard: co-designing wearable cognitive augmentation for k-12 teachers. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge, pp. 79–88 (2018)
7. Kaliisa, R., Misiejuk, K., López-Pernas, S., Khalil, M., Saqr, M.: Have learning analytics dashboards lived up to the hype? A systematic review of impact on students' achievement, motivation, participation and attitude. In: Proceedings of the 14th Learning Analytics and Knowledge Conference, pp. 295–304 (2024)
8. Kitto, S., Chiang, H.M., Ng, O., Cleland, J.: More, better feedback please: are learning analytics dashboards (lad) the solution to a wicked problem? Adv. Health Sci. Educ. 1–17 (2024)
9. van Leeuwen, A., Knoop-van Campen, C.A., Molenaar, I., Rummel, N.: How teacher characteristics relate to how teachers use dashboards: results from two case studies in k-12. J. Learn. Analy. **8**(2), 6–21 (2021)
10. Li, H., et al.: Math in motion: analyzing real-time student collaboration in computer-supported learning environments. In: Proceedings of the 17th International Conference on Educational Data Mining, pp. 533–541 (2024)
11. Martinez-Maldonado, R., Echeverria, V., Fernandez Nieto, G., Buckingham Shum, S.: From data to insights: a layered storytelling approach for multimodal learning analytics. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–15 (2020)
12. Martinez Maldonado, R., Kay, J., Yacef, K., Schwendimann, B.: An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In: Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, 14–18 June 2012. Proceedings 11, pp. 482–492. Springer, Cham (2012)
13. Prieto, L.P., Dimitriadis, Y., Asensio-Pérez, J.I., Looi, C.K.: Orchestration in learning technology research: evaluation of a conceptual framework. Res. Learn. Technol. **23** (2015)
14. Rodriguez Triana, M.J., Martínez-Monés, A., Asensio-Pérez, J.I., Dimitriadis, Y.: Monitoring-aware learning design process: pilot studies in authentic CSCL scenarios (2015)
15. Schwarz, B.B., De Groot, R., Mavrikis, M., Dragon, T.: Learning to learn together with CSCL tools. Int. J. Comput.-Support. Collab. Learn. **10**, 239–271 (2015)

16. Verbert, K., et al.: Learning dashboards: an overview and future research opportunities. Pers. Ubiquit. Comput. **18**, 1499–1514 (2014)
17. Verbert, K., Ochoa, X., De Croon, R., Dourado, R.A., De Laet, T.: Learning analytics dashboards: the past, the present and the future. In: Proceedings of the Tenth International Conference on Learning Analytics and Knowledge, pp. 35–40 (2020)
18. Voyiatzaki, E., Avouris, N.: Support for the teacher in technology-enhanced collaborative classroom. Educ. Inf. Technol. **19**, 129–154 (2014)
19. Weitnauer, E., Landy, D., Ottmar, E.: Graspable math: towards dynamic algebra notations that support learners better than paper. In: 2016 Future Technologies Conference (FTC), pp. 406–414. IEEE (2016)
20. Yan, L., et al.: Vizchat: enhancing learning analytics dashboards with contextualised explanations using multimodal generative AI chatbots. In: International Conference on Artificial Intelligence in Education, pp. 180–193. Springer, Cham (2024)

Towards A Student-Facing Dashboard to Support Learning of Inquiry Competencies in Interactive Simulations

Eman Ganaiem[1(✉)], Tanja Käser[2], and Ido Roll[1,3,4]

[1] Faculty of Education in Science and Technology, Technion, Haifa, Israel
Ganaiem.Eman@gmail.com, Roll@technion.ac.il
[2] School of Computer and Communication Sciences (IC), EPFL, Lausanne, Switzerland
Tanja.Kaeser@epfl.ch
[3] Faculty of Data and Decision Sciences, Technion, Haifa, Israel
[4] ETS, Princeton, NJ, USA

Abstract. Inquiry-based learning with interactive simulations is widely used in STEM education, yet its open-ended nature presents challenges for learners. To support students in reflecting on their inquiry processes, Learning Analytics Dashboards (LADs) offer a promising solution. While learning analytics has been used to analyze inquiry patterns, no prior studies, to our knowledge, have developed learner-friendly dashboards to help middle school students visualize their inquiry processes. Addressing this gap, this study presents the design of two mockups for dashboard aimed at supporting inquiry in interactive simulations. Using a participatory design approach, study 1 involved collaboration between middle school students (n = 18), teachers (n = 5), and a designer to co-design inquiry dashboards, investigating students' preferences for visual design elements that effectively represent inquiry process. This resulted in two mockups: the Island Design, which integrates a story-like metaphor for engagement, and the Venn Diagram Design, which prioritizes simplicity. Study 2 evaluated interpretation of the designs with 119 students. Results showed that while the Island Design strongly conveyed most inquiry competencies, students preferred the Venn Diagram for its simplicity, highlighting the need to balance engagement with clarity. We present key design preferences and how these are applied in the designs, as well as trade-offs between design elements. This paper marks a step toward developing inquiry dashboards to support inquiry learning in simulations.

Keywords: Learning Analytics Dashboards · Interactive Simulations · Inquiry Competencies · Participatory Design

1 Introduction

Inquiry-based learning is an active, self-directed process of scientific discovery [1] that is essential in science education for developing critical thinking and conceptual understanding [2]. Interactive simulations, a form of Exploratory Learning Environments (ELE), have shown promise in enhancing inquiry learning by having students learn science

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 439–445, 2025.
https://doi.org/10.1007/978-3-031-98462-4_55

by experimenting [3]. However, learning with simulations is cognitively demanding and requires additional support [4]. One approach for support is to present learners with visualizations of their inquiry to make scientific thinking visible and allow them to reflect on their processes. Learning Analytics Dashboards (LADs) serve this purpose by providing powerful visualizations which offer insights into learners' inquiry [5, 6], and improve students' awareness and reflection [7, 8]. Yet, without involving end-users, dashboards may fail to meet learners' needs [9].

This paper presents the design and evaluation of an inquiry dashboard mockup for middle school students using a participatory approach, where students actively contributed to the design. However, designing such a dashboard is challenging, as it requires balancing expert input with students' limited knowledge and inquiry skills. Additional challenges are that the dashboard must be flexible enough to represent diverse inquiry processes, avoid simulation-specific features, focus on descriptive (not performance) metrics, and communicate effectively with students [10]. Beyond these design considerations, the dashboard must visualize four key inquiry competencies: (i) Experimenting with *all* variables to investigate their effects [11], (ii) Applying *Control-of-Variables Strategy (CVS)* to design experiments that control all variables except the one being tested [12], (iii) Sampling a *wide range* of data points to improve the reliability of results [13], and (iv) Conducting *multiple measurements* for more accurate results [13].

With the two studies, we address the following research questions:

1. What visual design elements do stakeholders (middle-school students and teachers) prefer to include or exclude in the inquiry dashboard?
2. How do the final mockup dashboard designs align with the needs and preferences of the stakeholders?
3. How effective are the final designs in communicating their goals? That is, do middle school students understand the conveyed information?

The first study used a participatory design approach in which middle school students helped create inquiry dashboard mockups aligned with their needs. Together with teachers and a designer, they discussed preferences and visual features in recorded design sessions. The goal was to identify elements that support student understanding and engagement (RQ1 & RQ2). The second study evaluated the two resulting designs to determine their clarity, effectiveness, and any needed improvements (RQ3).

2 Study 1: Designing an Inquiry Dashboard Mockup

We used a participatory design approach to develop mockup dashboards that addressed the key challenges mentioned earlier.

Participants included 18 ninth-grade students, 5 STEM teachers, and 1 designer. The study spanned three phases across seven 90 min sessions.

Phase 1: Teachers engaged in three sessions to explore simulation-based inquiry, and generate mockup ideas using sketching activities. These mockups served as "idea cards" for the next phase. Teachers were also prepared to mentor students in phase 2.

Phase 2: Middle school students participated in three sessions, exploring inquiry through simulations and collaborating with teachers to design new mockups based on their preferences. Teachers mentored up to four students each, using Phase 1 designs to inspire ideas and clarify the study's goals.
Phase 3: Teachers, the researcher, and the designer worked together to refine the student-generated mockups to better align with the study goals.

2.1 Study 1 Results: Design Preferences and Final Mockups

Transcripts from sessions revealed five key preferences for dashboard visuals:

Timeline for Progress: Both students and teachers valued a timeline to track when and how variables were changed.
Dashboard Legend: A clear legend was considered essential for interpreting visual elements without needing extra explanation.
Range of Variable Values: Students wanted a scale to show measurement values clearly, suggesting min-to-max markers.
Consistency in Representation: Students found designs with unrelated visual metaphors confusing and hard to interpret. Most reported needing extra time to understand them and preferred clearer, and a more consistent representation.
Simplified Design and Interaction: While both teachers and students appreciated some level of interactivity, they agreed that excessive complexity could be overwhelming. They preferred a simple, intuitive interface that allowed for easy data interpretation without requiring too much interaction.

At the end of study 1, two designs- "Island" (Fig. 1) and "Venn Diagram" (Fig. 2)- were selected for best aligning with the study goals. Table 1 summarizes how each design represents key features, including the four targeted inquiry competencies.

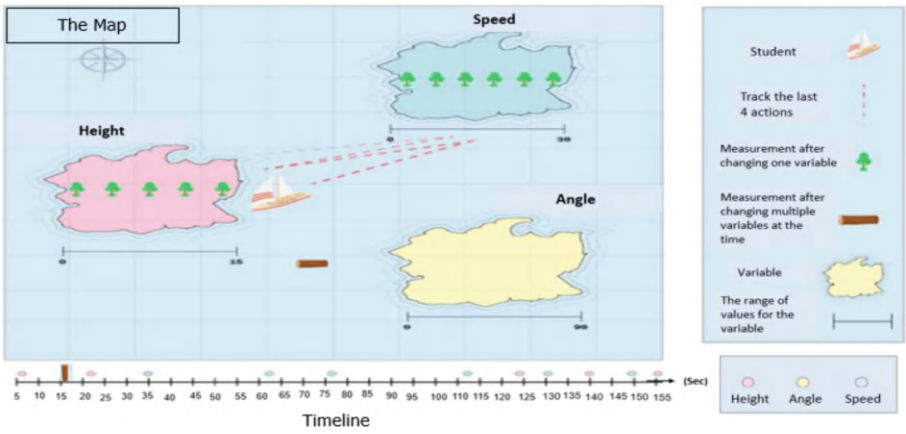


Fig. 1. The Island Design: Information about measurements (Map) and the process (Timeline).

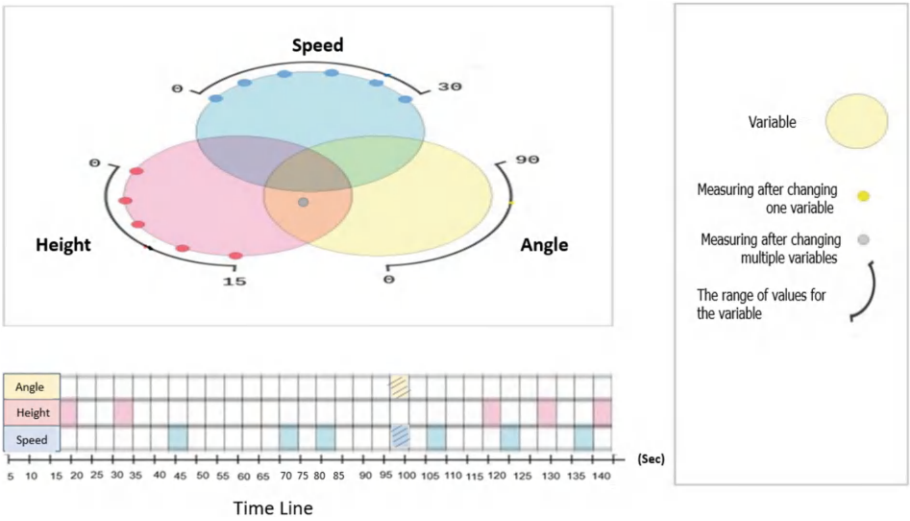


Fig. 2. The Venn Design: The circles describe the measurements, with a timeline below.

Table 1. Visual features and their representation in the Island and Venn Diagram designs.

Feature	Island Design	Venn Diagram Design
Independent variables	Are represented as islands	Are represented by circles
CVS actions	A "tree" is planted on the corresponding island, indicating that CVS was maintained	A dot appears on the circumference of the corresponding circle
Non-CVS actions	A "tree trunk" icon appears in the sea, indicating that not all variables were isolated	A gray dot appears in the overlapping area when multiple variables are changed before measurement
Range of data	Each island includes an axis displaying the variable's range. The tree is placed at the corresponding measurement value	Each circle has a scale displaying the variable's range, and the dot is placed on the circumference at the corresponding value
Number of measurements	The number of trees planted on each island	The number of dots on the circumference of each circle
Timeline	Provides a chronological view of all measurements, distinguishing between CVS (tree) and non-CVS (tree trunk) measurements	CVS measurements are shown as colored boxes, and non-CVS measurements as colored boxes overlaid with gray lines

(*continued*)

Table 1. (*continued*)

Feature	Island Design	Venn Diagram Design
Interactivity	A boat moving in real time shows the last four actions taken in the simulation	N/A

3 Study 2: Evaluating the Mockup Designs

A total of 119 seventh- and eighth-grade students from four classes participated in a single 90 min session. Students were randomly assigned by class to one of two sequences: two classes (61 students) viewed the Island design first, followed by the Venn design, while the other two classes (58 students) viewed the Venn design first.

Each session began with a 10 min introduction to the study's goals, followed by a 20 min demonstration of the projectile motion simulation to introduce the targeted inquiry competencies. Students were then shown the first dashboard (Island or Venn) with an explanation of its visual elements, followed by the second dashboard (30 min total). Finally, students responded individually to 10 questions presented one at a time on a board, recording their answers on answer sheets (30 min).

To evaluate students' understanding of the designs, we used 10 multiple-choice questions across three categories. The first category assessed their ability to interpret inquiry competencies, with four questions per design (eight in total). Question 9 asked which map design (Island or Venn) was clearer, and Question 10 addressed timeline clarity. Figure 3 shows the Island version of the interpretation questions. An equivalent set was used for the Venn design.

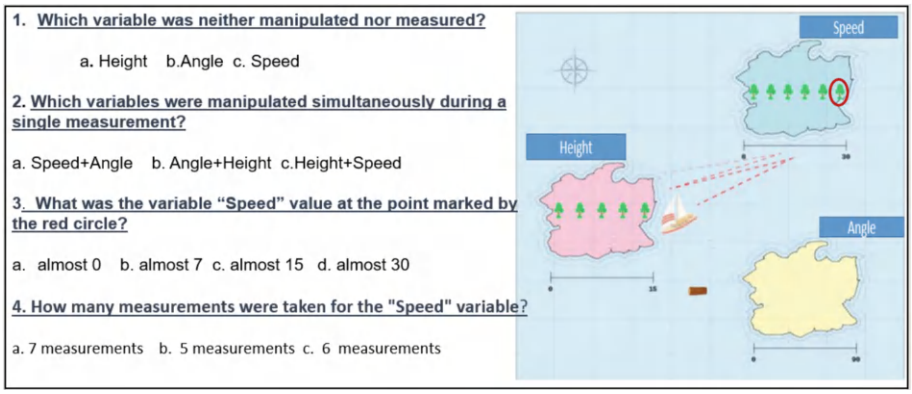


Fig. 3. Set of questions given to evaluate interpretation of visual design elements in the island design

3.1 Study 2 Results – Evaluating the Mockup Designs

The results of the first eight questions, designed to evaluate how effectively the dashboard designs communicated inquiry competencies. Most questions in the Venn design had an accuracy of 83–90%, except for the CVS question, which scored 56%. In the Island design, most questions showed very high accuracy (92% +), while the CVS question scored only 37%. Table 2 summarizes the percentage of correct responses for each question in both designs.

Table 2. Students Correctly Answering Questions on Inquiry Competencies by Dashboard Design. Reported data is % correct (# of students correct out of overall n = 119)

Topic:	Manipulating variables	CVS	Measurements	Range
Island	92% (109)	37% (44)	93% (111)	92% (110)
Venn Diagram	83% (49)	56% (67)	87% (103)	90% (107)

Questions 9 and 10 assessed which of the two 'Map' and 'Timeline' visualizations students found easier to understand. For the 'Map' visualization, 73% preferred the version in the Venn design, while 7% found both equally clear. For the 'Timeline', 69% favored the one in the Venn design, with 5% perceiving no difference between the two.

4 General Discussion and Summary

Study 1 aimed to design a mockup for an inquiry dashboard and explore stakeholder preferences regarding its features and alignment with their needs. The two mockups, the "Island" and the "Venn diagram". The Island Design includes a moderate narrative element, providing context without overwhelming complexity, while the Venn Diagram is more abstract, focusing entirely on data representation without any narrative elements. They both effectively met the four key design criteria. Both designs visualized the four inquiry competencies; emphasized process over performance, focused on inquiry actions rather than content-specific elements. Key design preferences as suggested by students were incorporated in both designs.

Study 2 evaluated how well the two mockups communicated inquiry processes. The Island Design had high success rates (92% +), though only one-third of students correctly interpreted the non-CVS symbol. The Venn Diagram showed similar success (83–90%), but the gray dots used for non-CVS were also confusing. Despite these challenges, students found the Venn Diagram clearer and preferred it for both map and timeline views. Notably, it was originally designed by students, suggesting the value of their input even without prior expertise.

Together, the studies highlight the benefits of participatory design. Involving students led to a preferred, intuitive dashboard, balancing clarity and engagement. While the focus was on communication rather than learning outcomes, the results offer insights into designing student-centered dashboards for inquiry learning.

Acknowledgments. This work was supported by the Neubauer Foundation and by ISF Grant No. 1573/21.

Disclosure of Interest. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Pedaste, M., et al.: Phases of inquiry-based learning: definitions and the inquiry cycle (2015). https://doi.org/10.1016/j.edurev.2015.02.003
2. Harrison, A.M., Schunn, C.D.: The transfer of logically general scientific reasoning skills. In: Forbus, K., Gentner, D., Regier, T., Mahwah, N., Erlbaum, J. (eds.) Proceedings of the 26th Annual Conference of the Cognitive Science Society, pp. 541–546 (2004)
3. Roll, I., et al.: Understanding the impact of guiding inquiry: the relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning. Instr. Sci. **46**(1), 77–104 (2018). https://doi.org/10.1007/s11251-017-9437-x
4. Eslinger, E., White, B., Frederiksen, J., Brobst, J.: Supporting inquiry processes with an interactive learning environment: inquiry island. J. Sci. Educ. Technol. **17**, 610–617 (2008)
5. Perez, S., et al.: Identifying productive inquiry in virtual labs using sequence mining. In: Artificial Intelligence in Education: 18th International Conference, AIED 2017, Wuhan, China, pp. 18287–18298. Springer (2017)
6. Wang, K.D., Cock, J.M., Käser, T., Bumbacher, E.: A systematic review of empirical studies using log data from open-ended learning environments to measure science and engineering practices. Br. J. Educ. Technol. **54**(1), 192–221 (2023). https://doi.org/10.1111/bjet.13289
7. Gobert, J.D., et al.: Real-time scaffolding of students' online data interpretation during inquiry with Inq-ITS using educational data mining Cyber-Phys. Lab. Eng. Sci. Educ., 191–217 (2018). https://doi.org/10.1007/978-3-319-76935-6_8
8. Molenaar, I., Horvers, A., Dijkstra, R., Baker, R.S.: Personalized visualizations to promote young learners' SRL: the learning path app. ACM Int. Conf. Proc. Ser., 330–339 (2020). https://doi.org/10.1145/3375462.3375465
9. Prieto-Alvarez, C.G., Martinez-Maldonado, R., Anderson, T.D.: Co-designing learning analytics tools with learners. Learn. Anal. Classr., 93–110 (2019). https://doi.org/10.4324/9781351113038-7
10. Roll, I., Aleven, V., McLaren, B.M., Koedinger, K.R.: Designing for metacognition-applying cognitive tutor principles to the tutoring of help seeking. Metacogn. Learn. **2**(2–3), 125–140 (2007). https://doi.org/10.1007/s11409-007-9010-0
11. Johnson, B., Christensen, L.B.: Educational Research: Quantitative, Qualitative, and Mixed Approaches. SAGE
12. Stender, A., Schwichow, M., Zimmerman, C., Härtig, H.: Making inquiry-based science learning visible: the influence of CVS and cognitive skills on content knowledge learning in guided inquiry. Int. J. Sci. Educ. **40**(15), 1812–1831 (2018). https://doi.org/10.1080/09500693.2018.1504346
13. Arnold, J.C., Kremer, K., Mayer, J.: Understanding Students' Experiments-What kind of support do they need in inquiry tasks? Int. J. Sci. Educ. **36**(16), 2719–2749 (2014). https://doi.org/10.1080/09500693.2014.930209

HuTCH: Human Teachable Concept Highlighter for Post-hoc Visual Explanations

Erfan Mirhaji[1(✉)], Nikhil Krishnaswamy[1], Jill Zarestky[2], Lisa Mason[3], Sarath Sreedharan[1], and Nathaniel Blanchard[1]

[1] Department of Computer Science, Colorado State University, Fort Collins, USA
{erfan.mirhaji,nikhil.krishnaswamy,sarath.sreedharan,
nathaniel.blanchard}@colostate.edu
[2] School of Education, Colorado State University, Fort Collins, USA
jill.zarestky@colostate.edu
[3] Colorado State University Extension, Fort Collins, USA
lisa.mason@colostate.edu

Abstract. In recent years, deep learning (DL) models have surpassed human experts in a variety of tasks. However, when it comes to educational contexts, human experts possess something these high-performing models currently lack—the ability to provide clear and approachable explanations tailored to human learners. Efforts to develop explainable methods for these models have typically been designed for DL-experts rather than learners. In this study, we propose a novel approach that combines high-performance models with the teachability of human expert explanations. Specifically, we focus on generating post-hoc, human-understandable explanations for key expert-defined concepts on a novel task: training citizen scientists to identify pollinators from images. Our method, HuTCH, transforms the representational space and highlights relevant segments of an image for learners based on essential concepts—for example, automatically highlighting hair in an image as a key concept for bee identification. We evaluate HuTCH's performance by comparing against traditional saliency maps and expert annotations, and show that HuTCH concepts better align with expert annotations. The proposed framework bridges the gap between models' accuracy and human-teachable features, contributing to the advancement of explainable AI for use in pedagogy.

Keywords: Explainable AI · Concept-Based Teaching · Concept Highlighting · Feature Localization

1 Introduction

In recent years, deep learning has made rapid advances. We currently have deep learning systems that can match or even surpass human-level performance in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 446–453, 2025.
https://doi.org/10.1007/978-3-031-98462-4_56

many tasks. Concurrently, there have been various efforts within the Explainable AI (XAI) field to generate explanations for why the system made specific decisions [5,9,15]. Unfortunately, these systems still do not provide a source of lessons for teaching non-expert learners about the target task. This is in no small part because most existing explanations are designed to help AI or domain experts debug and verify these systems. Nonetheless, if we are able to capitalize on these models to generate easy-to-understand lessons about tasks these systems excel at, this could open novel opportunities to automatically generate lessons for many learning problems. Towards this end, assuming access to domain experts, we develop an entirely new XAI paradigm that combines expert-specified rules with internal representations used by AI systems to automatically generate task-specific lesson content.

Here, we explicitly focus on a visual learning task, namely identifying different pollinator groups from their pictures. Convolutional Neural Networks (CNNs) can be trained to solve this task with a near-expert level accuracy. We introduce a novel method called HuTCH, which extends existing XAI methods, to generate lessons as to why the current image belongs to a specific group, which could be easily followed by learners with no familiarity with AI concepts or expertise in entomology. Specifically, HuTCH leverages concepts and rules used by domain experts to provide a conceptual explanation as to why the insect in the picture corresponds to a specific pollinator group. HuTCH will make use of internal representations learned by the CNN to determine what rules and concepts are applicable for the given image. We then make it further accessible to non-domain experts by illustrating visually each concept used in the rule to make the determination.

2 Related Work

Feature-attribution methods, for instance, gradient-based methods, have been popular in generating explanations for vision models [9]. Saliency maps are one example of such methods [20]. Other methods like LIME [18] and Shapley Additive Explanations (SHAP) [15] have been introduced to generate explanations. Even though many methods besides LIME and SHAP have been established, they can generate misleading explanations on the true reason for a prediction by a model [2,3,7,26]. These methods are designed for domain experts who can associate meaningful concepts with visual features highlighted in the explanation.

Concept-based explanations are another type of post-hoc explanations. In a concept-based explanation, explanations go beyond features of each image and identify higher-level human-understandable concepts that are true for the entire dataset [8]. Concept Activation Vectors (CAVs) are another method used in concept-based explanations to provide clarification about a neural net's internal state in terms of human-friendly concepts [13]. Because of the human-explainability focus of these works, they are easier to understand by humans. TCAV and its explainability implications have been widely used in medical applications [6,16,22]. Some other methods suggest configuring the architecture of

the model to achieve Post-hoc Concept Bottleneck models (PCBMs) [25]. These explanability methods also assume the familiarity of the user with the defined concepts. If a layperson is presented with these explanations, they might not understand them, as a familiarity with said concepts is implicitly expected by concept-based explanations. Our Human Teachable Concept Highlighter (HuTCH) method produces explanations based on human-understandable concepts with the aid of a field expert, and further highlights the region containing the concept. We separate the decision-making process of the model and the explanations that are generated by our model, which are highly teachable and interpretable to learners.

3 Methodology

In this paper, we highlight a region of an input image containing a teachable concept to learners based on information from a human expert using the HuTCH framework. HuTCH's highlights are more similar to that of human expert's, indicating its applicability as a tutor for teaching visual tasks.

3.1 Concepts, Datasets and Model

A visual identification task pertinent to teaching citizen scientists is the classification of bees versus wasps. Entomologists look for specific characteristics when encountering an insect that is either a bee or a wasp; for example, hair on the body, or the narrowness of the abdomen and thorax connection, commonly known as a "pinch-waist". These two are some of the most important concepts differentiating bees and wasps. These concepts are suitable for citizen scientists to aid them in this specific visual identification task. As a result, we tuned our explainability method to these two features.

The data used in this study is gathered from the iNaturalist website [1]. We chose bee species that display hair on their body, and wasp species that have the pinch-waist feature. In total we selected 23 bee species and 17 wasp species, and 226,892 images of bees and 229,921 images of wasps were collected. To highlight regions containing a certain concept using our HuTCH method, we defined a concept dataset with 200 bee and 200 wasp images. A human expert selected one region in each image that contains the corresponding concept, and one region where the concept is absent, resulting in two datasets: one for each concept, containing 200 positive and 200 negative examples. Next, we augmented our dataset, resulting in 9,600 positive and 9,600 negative samples.

We used 60 bee images and 60 wasp images for our comparisons. First, the original images are resized to 224 by 224 and then segmented using the Mask R-CNN ResNet-50 object instance segmentation method [11] with zero threshold to find the largest segment. The results are then given to the expert, the saliency highlighter, and our two HuTCH highlighter methods, each outputting their highlights. For the CNN model, we chose the ResNet-152 architecture [12], and the model was re-trained as a whole, accounting for mean and standard deviation

values of the training dataset. The trained ResNet-152 CNN model was tested on 34,022 bee and 34,481 wasp images, achieving a combined F_1 score of 96%.

3.2 Concept Activation Vectors

CNNs transform images to higher representational spaces and extract many complex features. In the case of ResNet-152, we use the output of the second convolutional layer of the third bottleneck layer of the fourth residual stage, which contains 512 features each with a spatial dimension of 7×7, our layer of interest. We then pass each of our concept datasets to the model and train a linear classifier on the flattened activations of our layer of interest. The vector containing weights of this linear classifier is the Concept Activation Vector. Later on, using the dot product of the flattened activations of a new image and the CAV of a concept, we can calculate the alignment of that image with that concept.

3.3 Comparison with Human Expert and CNN Highlights

Using the selected 120 test images, we ask the expert, the saliency map method, and our HuTCH method to highlight what each considers the most important region. The human expert highlights the relevant concept region in each image; in the case of a bee, the hair region, and in the case of a wasp, the pinch-waist region. The expert highlights only one continuous region. The CNN model, using the normalized gradients with a threshold of 0.05, calculates the saliency map of each image. The acquired saliency map of each image is then further partitioned into rectangles. Each map is partitioned into rectangles of widths and heights of 56, 74 and 112, and are saved separately to create new images referred to as sub-images. The average gradient value of all sub-images is then measured and sorted. The sub-images corresponding to the top K highest average gradient values are overlapped and combined, which is the highlighted region of the saliency map.

3.4 HuTCH Highlights

Our HuTCH method highlights regions in which a certain concept is most present. First, each image is passed to the CNN model to make a prediction of whether it is a bee or a wasp. Then, the concept to look for is set to hair or pinch-waist respectively; this is done to combine the accuracy of CNNs with the teachability of expert-defined concepts. We then further segment the image with a 0.05 threshold to filter out the background from the insect. Next, using two different methods, HuTCH further partitions the image; once using rectangles, called HuTCH Rectangle, and another time using all the segments generated by the Mask R-CNN segmentation method [11], called HuTCH Segmented.

In the HuTCH Rectangle, similar to the saliency partitioning, the image is partitioned into all possible rectangles with sides of 56, 74, and 112 pixels, referred to as sub-images. This dissects every single image into two datasets, one dataset of sub-images of HuTCH Rectangle, one dataset of sub-images of

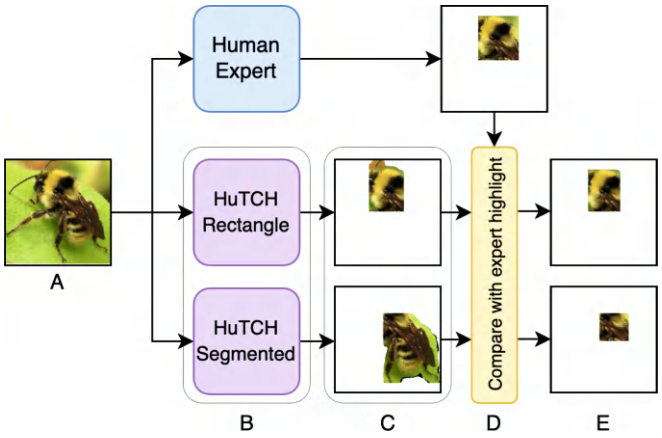


Fig. 1. The workflow of highlighting images by each method. **A** Each input image is filtered to the biggest object segment. **B** The two HuTCH methods partition the image into sub-images via rectangles and masks respectively, and calculate the sub-images with the highest CAV scores. **C** The combined region of top K sub-images is highlighted by each method. **D** The highlights are compared to the expert's highlighted region via IoU and Dice metrics. **E** The overlapping region is shown.

HuTCH Segmented. Finally, each sub-image of each HuTCH dataset is passed through the model to acquire the activation from the layer of interest. The dot products of the flattened activations and the CAV of the respective concept are then measured and sorted [13]. The sub-images corresponding to the K highest dot products are merged together, producing the highlighted region of each HuTCH method.

Table 1. The average and standard deviation of IoU scores between highlighted regions of each method and expert highlights.

Method	Top 1	Top 2	Top 3
Saliency Rectangle Highlight	0.155 ± 0.166	0.189 ± 0.164	0.209 ± 0.164
HuTCH Segmented Highlight	0.237 ± 0.165	0.237 ± 0.124	0.238 ± 0.112
HuTCH Rectangle Highlight	0.275 ± 0.211	0.309 ± 0.190	0.307 ± 0.161

Table 2. The values for Dice coefficient between highlighted regions of each method and expert highlights.

Method	Top 1	Top 2	Top 3
Saliency Rectangle Highlight	0.237 ± 0.288	0.288 ± 0.220	0.317 ± 0.216
HuTCH Segmented Highlight	0.356 ± 0.212	0.368 ± 0.159	0.372 ± 0.140
HuTCH Rectangle Highlight	0.388 ± 0.263	0.439 ± 0.230	0.446 ± 0.198

4 Results and Discussion

In order to compare the highlighted region from each method (Saliency Rectangle, HuTCH Rectangle and HuTCH Segmented) with the highlighted region by the expert, we used two metrics; Intersection over Union (IoU) and Dice coefficient. The HuTCH Rectangle and HuTCH Segmented methods are compared to the region that the expert has marked as relevant for learners to correctly identify the insect. The overall workflow can be seen in Fig. 1. The results of the average and standard deviations of comparisons can be viewed in Tables 1 and 2 on the aggregated top K regions respectively.

The concept that the expert highlights is something that is understandable by humans and is used to teach about different taxa in entomology. Compared with salience maps, HuTCH methods more accurately align with expert highlights of key visual concepts for learners. By combining the teachability of our methods and the high accuracy of CNNs, we can teach learners with high certainty and high explainability. In this work, the HuTCH framework bridges the gap between AI models' accuracy and human-teachable features, contributing to the incorporation of explainable AI in teaching. We have demonstrated that concept-based highlighting achieves more alignment with expert highlighting and thus improves the teachability of the explanations. Our method demonstrates how AI models' high performance can be coupled with human-understandable interpretability. This explainability approach can be used as an enhancement for educational settings, complementing educators or fully taking the expert's place for visual teaching tasks.

5 Limitation and Future Work

In our segmentation method, we used the mask R-CNN ResNet-50 model, which is trained on Microsoft's COCO dataset; sometimes unrelated background patches found their way into our sub-images, which resulted in erroneous CAV scores. More fine-tuned and specialized segmentation methods that can better segment bees and wasps into different sections would help us work with finer concepts used by human experts. Another limitation was the quality of the dataset as they were captured by amateurs and enthusiasts, sometimes lacking focus or visibility of insects. A cleaner dataset would improve the accuracy of both classification and concept highlights. Currently, we have only used two concepts, body hair and pinch-waist, for training the CAVs and subsequently the linear classifier. We also only trained the CNN on a few bee and wasp species exhibiting these features. For future work, we can expand the concepts and the training dataset to include more features and more species showing those features.

Acknowledgments. This study was funded by NSF grant 2303019.

References

1. inaturalist. https://www.inaturalist.org. Accessed Jan 2025
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018)
3. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: quantifying interpretability of deep visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6541–6549 (2017)
4. Blanchard, E.G., Mohammed, P.: On cultural intelligence in LLM-based chatbots: implications for artificial intelligence in education. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education, pp. 439–453. Springer, Cham (2024)
5. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**(8) (2019). https://doi.org/10.3390/electronics8080832, https://www.mdpi.com/2079-9292/8/8/832
6. Clough, J.R., Oksuz, I., Puyol-Antón, E., Ruijsink, B., King, A.P., Schnabel, J.A.: Global and local interpretability for cardiac MRI classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 656–664. Springer (2019)
7. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 3681–3688 (2019). https://doi.org/10.1609/aaai.v33i01.33013681, https://ojs.aaai.org/index.php/AAAI/article/view/4252
8. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019)
9. Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., Giannotti, F.: A survey of methods for explaining black box models. ACM Comput. Surv. **51** (2018). https://doi.org/10.1145/3236009
10. Gütl, C., García-Barrios, V.M.: The application of concepts for learning and teaching. In: Proceedings of 8th International Conference on Interactive Computer Aided Learning (ICL 2005). Citeseer (2005)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Kim, B., et al.: Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning (2017). https://api.semanticscholar.org/CorpusID:51737170
14. Linson, A., Xu, Y., English, A.R., Fisher, R.B.: Identifying student struggle by analyzing facial movement during asynchronous video lecture viewing: towards an automated tool to support instructors. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education, pp. 53–65. Springer, Cham (2022)

15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 4768–4777. Curran Associates Inc., Red Hook, NY, USA (2017)
16. Graziani, M., Andrearczyk, V., Marchand-Maillet, S., Müller, H.: Concept attribution: explaining CNN decisions to physicians. Comput. Biol. Med. **123**, 103865 (2020). https://doi.org/10.1016/j.compbiomed.2020.103865, https://www.sciencedirect.com/science/article/pii/S0010482520302225
17. Ma, Q., Shen, H., Koedinger, K., Wu, S.T.: How to teach programming in the AI era? USING LLMs as a teachable agent for debugging. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education, pp. 265–279. Springer, Cham (2024)
18. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 1135–1144. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778
19. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: why did you say that? Visual explanations from deep networks via gradient-based localization (2016). https://api.semanticscholar.org/CorpusID:217472699
20. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: Workshop at International Conference on Learning Representations (2014)
21. Sonkar, S., Liu, N., Mallick, D.B., Baraniuk, R.G.: Marking: visual grading with highlighting errors and annotating missing bits. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education, pp. 309–323. Springer, Cham (2024)
22. Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A.: Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med. Image Anal. **79**, 102470 (2022)
23. Wang, Y., Zhang, T., Guo, X., Shen, Z.: Gradient based feature attribution in explainable AI: a technical review (2024). https://arxiv.org/abs/2403.10415
24. Yao, Y.: Concept formation and learning: a cognitive informatics perspective. In: Proceedings of the Third IEEE International Conference on Cognitive Informatics, pp. 42–51 (2004). https://doi.org/10.1109/COGINF.2004.1327458
25. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480 (2022)
26. Zhou, Y., Booth, S., Ribeiro, M.T., Shah, J.: Do feature attribution methods correctly attribute features? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 9, pp. 9623–9633 (2022). https://doi.org/10.1609/aaai.v36i9.21196, https://ojs.aaai.org/index.php/AAAI/article/view/21196

Understanding Public School Enrollment Trends in the Northeast of Brazil: Unraveling Features Through AI Related to the Learning Process for Low-Income Students

Artur Cavalcante de Jesus[1], Leonardo Marques[1], Diego Carvalho do Nascimento[2], and Bruno Almeida Pimentel[1(✉)]

[1] Universidade Federal de Alagoas, Maceió, Brazil
{acj,brunopimentel}@ic.ufal.br, leonardo.marques@cedu.ufal.br
[2] NEOMA Business School, 76825 Mont-Saint-Aignan, France
diego.nascimento@neoma-bs.fr

Abstract. Accurate enrollment forecasts are vital for evidence-based educational planning. We train Long Short-Term Memory (LSTM) networks on ten years of Brazilian School Census data to predict public-school enrolment in Maceió, a low-income urban centre in Brazil's Northeast. To support equitable resource allocation by the National Textbook and Educational Material Programme (PNLD), we couple the model with Explainable-AI techniques—SHAP values and Granger causality—to uncover the drivers of enrolment dynamics at preschool, elementary, and high-school levels. Results show that infrastructure quality (adapted restrooms, multimedia equipment) is the main lever for preschool access, teacher availability and libraries sustain elementary retention, while science labs, sports facilities, and ICT resources shape high-school demand. Intervariable analysis reveals that targeted budget cuts can trigger indirect declines in enrolment, underscoring the need for holistic investment.

Keywords: School enrollment forecasting · Long Short-Term Memory · Explainable AI · Educational equity · Low-income students

1 Introduction

Efficient educational management depends on identifying the drivers of enrollment change [1,3,23]. Brazil's annual School Census—the nation's most detailed basic-education database—records enrollments, infrastructure, teacher supply, and student socioeconomic profiles, providing the evidence base for strategic planning. As Lima &Sousa [14] observe, the Census data "support an educational diagnosis for Brazil, guiding strategies for student access, retention, and success" [14]. Hence, longitudinal analysis of these data not only tracks enrollment trends but also exposes regional inequalities and informs corrective policies [20].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 454–461, 2025.
https://doi.org/10.1007/978-3-031-98462-4_57

Maceió illustrates these challenges vividly. INEP's Educational Indicators Panel ranked the city among the nation's worst for dropout, with 2.1% of municipal elementary pupils leaving school in 2019 [21], and UNICEF notes that, despite recent declines, pronounced regional disparities persist [12]. Age-grade distortion is likewise severe: 40.9% of elementary students—about 47,000 children—were at least two years overage in 2016, dropping only to 31.7% by 2019 [14]. These figures highlight the need for detailed analysis of the factors driving enrollment instability [12].

This article models School Census time-series data (2013–2022, Maceió) with Long Short-Term Memory (LSTM) [9,10] neural networks to forecast enrollments in preschool, elementary, and high-school levels. LSTM captures historical patterns and temporal dependencies [2], delivering accurate projections. Granger causality [19] and SHAP-based eXplainable AI [18] then expose the variables exerting the strongest causal influence on enrollment dynamics. The resulting insights inform data-driven policy design, more efficient resource allocation, and provide a methodological template for other municipalities facing similar challenges.

This paper is organized as follows. Section 2 reviews the literature on machine learning, model interpretability, and relevant studies in educational data analysis. Section 3 outlines the methodology and data processing steps employed in this study. Section 4 presents the results. Finally, Sect. 5 concludes the study and proposes directions for future research.

2 Methods

Within the time series (TS) modeling family, Long Short-Term Memory (LSTM) networks extend recurrent neural networks by using gated units that mitigate vanishing or exploding gradients and thus capture long-range dependencies [6,11]. Although effective, their high-dimensional parameter space hampers interpretability, motivating the use of explainability techniques.

Explainable AI (XAI) makes machine-learning decisions transparent—an essential requirement as models increasingly shape high-stakes outcomes [16]. Deep networks, while accurate, act as "black boxes," prompting concerns over bias and accountability [22]. Typical XAI outputs include [13] (i) visual cues such as heatmaps, (ii) concise textual rationales, and (iii) example-based comparisons. Yet a tension remains: simple, inherently interpretable models may overlook complex relationships, whereas highly accurate deep models are much harder to explain [22].

Recent approaches suggest explainable-by-design methods and causal AI techniques to uncover causal relationships in machine learning models [22]. This work will focus on SHAP (*SHapley Additive Explanations*) [15] and Granger Causality for analyzing variable impacts and causal relationships in temporal contexts [5,7,8,17].

3 Experiments

The data used in this study comes from the School Census, conducted annually by the National Institute for Educational Studies and Research Anísio Teixeira (INEP), and can be found on the INEP website. The dataset includes detailed information on schools, classes, teachers, and enrollments from municipalities across Brazil.

Preprocessing used encode categorical attributes numerically, and scale every predictor with min-max normalization to [0,1]—a step that mitigates the scale sensitivity of LSTM networks. The normalized records were then reshaped via sliding-window segmentation into input–output sequences, preserving temporal order and long-range dependencies. The resulting dataset is consistent and ready for LSTM training.

A 3-year sliding window converts the School Census time series into sequences that predict the following year, reflecting the assumed persistence of factors driving enrollment and capturing seasonality.

The network, implemented in Keras, stacks two LSTM layers (300 units with return_sequences, dropout; then 200 units), followed by a single-unit dense output. Trained with Adam and MSE on an 80/20 split, 100 epochs, and batch size 256, this configuration balances efficiency and accuracy while modeling long-range dependencies.

The results were assessed using standard regression metrics commonly adopted in the machine learning field, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2) [4].

Additionally, the work also focused on correlation and causality analysis. The Granger causality method was applied to verify dependency relationships between the considered variables, while SHAP provided insights into the contribution of each variable to the model's results.

4 Results

Top-chart Fig. 1 illustrates the model's performance in predicting enrollments for Early Childhood Education. There is a reasonable alignment between the predicted and actual values, indicating that the model can capture part of the trend in enrollments for this segment.

The LSTM models showed varying performance across educational stages. For Early Childhood Education (E.C.E), the results were: MAE of 137.41, RMSE of 166.32, MSE of 27,663.03, and R^2 of 0.51. The Elementary School (E.S.) model achieved results with a MAE of 6,534.19, RMSE of 7,708.74, MSE of 59,424,611.42, and R^2 of 0.72. In High School (H.S.), the model reported a MAE of 306.01, RMSE of 454.06, MSE of 206,174.0, and R of 0.60.

Finally, the following results will be dedicated to explaining each model (and features causality) to each educational level detailing the main findings.

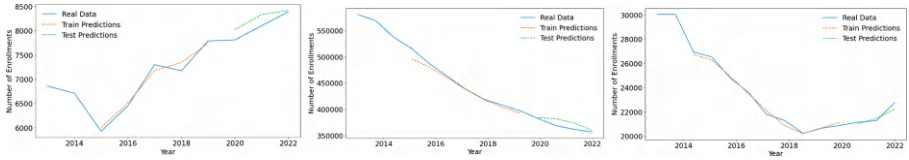


Fig. 1. Comparison between actual and predicted enrollments. Top-Chart describes the dynamic of Early Childhood Education records. Bottom-Charts Elementary School (left-hand) and High School (right-hand).

4.1 Early Childhood Education

Figures 2 present variables' importance in the LSTM adjusted model and interpretability analyses conducted using SHAP values to identify the variables that most influence the number of enrollments in Early Childhood Education.

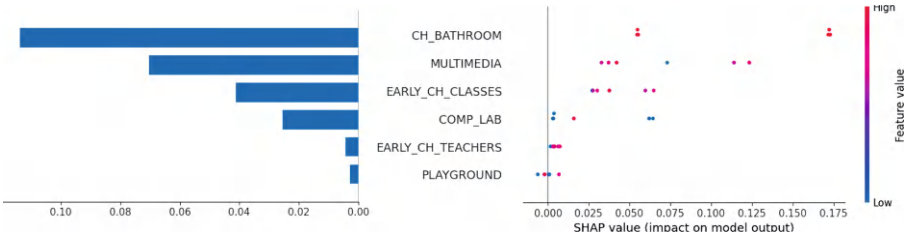


Fig. 2. The XAI Feature Importance (left-hand) and Beeswarm Plot (right-hand) using the LSTM model related to Early Childhood Education (Author, 2024). (Color figure online)

The analysis reveals that presence of adequate bathroom facilities for Early Childhood Education (`CH_BATHROOM`) is the most influential variable. Schools with this infrastructure (red points) tend to have positive SHAP values (to the right), indicating a positive impact on enrollments, while schools without this infrastructure (blue points) have negative or near-zero SHAP values, suggesting a negative or negligible impact. The variable multimedia equipment (`MULTIMEDIA`) emerges as the second most important variable, with a similar pattern: schools with more multimedia equipment (red points) correlate with positive SHAP values, suggesting a positive impact on the attractiveness of the school. The absence of such equipment (blue points) is associated with lower SHAP values.

4.2 Elementary School

According to the XAI's results (Fig. 3), the number of elementary school teachers (`ELEM_SCHOOL_TEACHERS`) is revealed to be the most influential variable. Schools with more teachers (red points) tend to have positive SHAP values (to the right),

indicating an increase in enrollments, while schools with fewer teachers (blue points) show negative or close-to-zero SHAP values, suggesting a negative or null impact. The number of elementary school classes (`ELEM_SCHOOL_CLASSES`) appears as the second most important variable, with a similar pattern: more classes (red points) correlate with positive SHAP values, impacting enrollments positively.

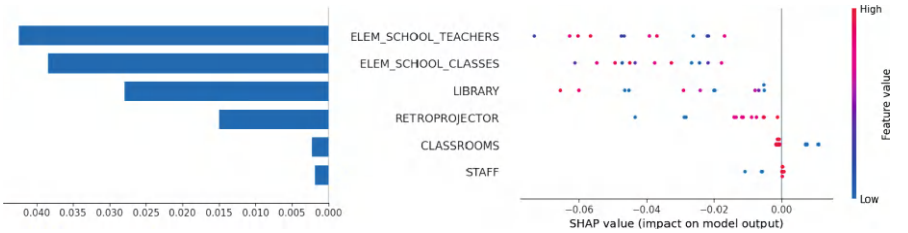


Fig. 3. The XAI Feature Importance (left-hand) and Beeswarm Plot (right-hand) using the LSTM model related to Elementary School (Author, 2024). (Color figure online)

4.3 High School

The XAI related to the LSTM model from the High School analysis (Fig. 4) reveals that the number of high school classes (`HS_CLASSES`) is the most influential variable. Schools with more classes (red points) tend to have positive SHAP values (to the right), indicating an increase in enrollments, while schools with fewer classes (blue points) have negative SHAP values (to the left) or values close to zero, suggesting a negative or null impact. The presence of a sports court (`SPORTS_COURT`) emerges as the second most important factor. Schools that offer a sports court (red points) tend to display positive SHAP values, which points to a favorable effect on enrollments, although some dispersion does occur.

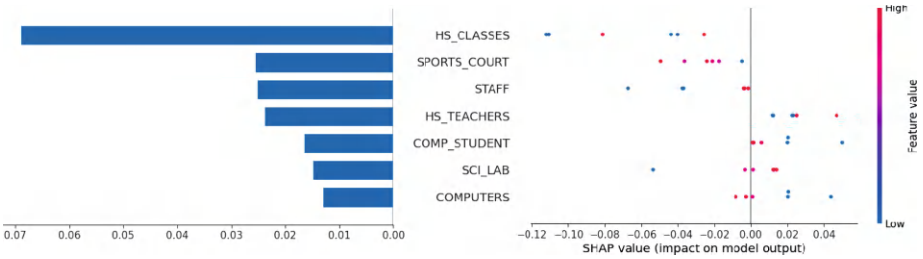


Fig. 4. The XAI Feature Importance (left-hand) and Beeswarm Plot (right-hand) using the LSTM model related to High School (Author, 2024). (Color figure online)

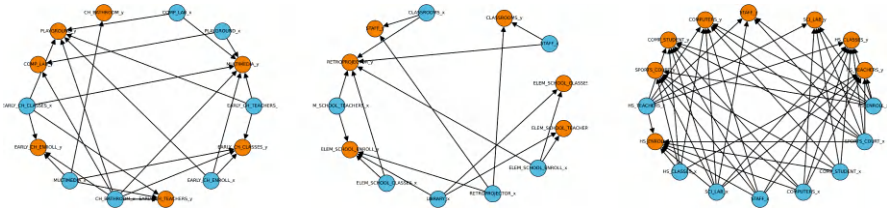


Fig. 5. Granger Causality through a probabilistic graphical representation (orange color response variables and light-blue explanatory variables): The left-hand is based on Early Childhood Education, the center is the Elementary School, and the right-hand is the High School features. (Color figure online)

Figure 5 shows the results of the Granger causality test applied to the adopted variable to bring an explanation into the LSTM results. The most important feature from the LSTM model for the E.C.E. was also the most connected variable from the graph (`CH_BATHROOM`), the highest degree in the network, and the feature (`MULTIMEDIA`) is the second most relevant feature with the highest closeness. Based on the E.S. data, the variable within (receiving) most connections (`RETROPROJECTOR`), the highest degree and closeness in the network, from the Granger test was also the less important from the LSTM model (moreover, its information is conditional on other variables). Finally, the H.S. Bipartite graph (explainable vs response features) results from an almost complete graph, describing the difficulty of revealing the hidden patterns, and suggesting hidden counterfactual features.

5 Conclusion

This study aimed to analyze the factors influencing enrollment numbers in early childhood, elementary, and high school education in the public network of Maceió, using LSTM neural networks for time series modeling and Explainable AI techniques such as SHAP and Granger Causality to interpret the most relevant variables. Using data from the INEP School Census covering a ten-year period, patterns were identified that help understand the evolution of enrollments and provide valuable insights for more efficient educational management.

The results suggest that factors like school infrastructure, teacher availability, technology access, and extracurricular activities significantly influence student retention. A link between structural and pedagogical aspects highlights the need for integrated planning that avoids harmful budget cuts. Limitations include the focus on Maceió, limiting generalization, and the exclusive use of the LSTM model, which could be compared with others. Future studies should explore broader regions, include socioeconomic variables, and develop predictive tools like educational dashboards to support real-time, data-driven decision-making.

References

1. Alam, A., Bowden, A.B.: Advancing high school dropout predictions using machine learning. In: International Conference on Artificial Intelligence in Education, pp. 334–341. Springer, Cham (2024)
2. Alizadegan, H., Rashidi Malki, B., Radmehr, A., Karimi, H., Ilani, M.A.: Comparative study of long short-term memory (LSTM), bidirectional LSTM, and traditional machine learning approaches for energy consumption prediction. Energy Explor. Exploit. **43**(1), 281–301 (2025)
3. Belfield, C.R., Levin, H.M.: The Price We Pay: Economic and Social Consequences of Inadequate Education. Brookings Institution Press (2007)
4. Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination r-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. **7**, e623 (2021)
5. Dantas, F., Weydmann, C.L.: Carne de frango: uma análise da relação entre os preços dos produtores e de exportação. Revista de Economia e Agronegócio **7**(1), 31–53 (2009)
6. Elshewey, A.M., Abed, A.H., Khafaga, D.S., Alhussan, A.A., Eid, M.M., El-Kenawy, E.: Enhancing heart disease classification based on greylag goose optimization algorithm and long short-term memory. Sci. Rep. **15**(1), 1277 (2025)
7. Farias, H.P., Sáfadi, T.: Causalidade entre as principais bolsas de valores do mundo. RAM. Revista de Administração Mackenzie **11**(2), 96–122 (2010)
8. Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: J. Econometric Soc. 424–438 (1969)
9. Graves, A., Graves, A.: Long short-term memory. In: Supervised Sequence Labelling with Recurrent Neural Networks, pp. 37–45 (2012)
10. Guo, Q., He, Z., Wang, Z.: Assessing the effectiveness of long short-term memory and artificial neural network in predicting daily ozone concentrations in Liaocheng city. Sci. Rep. **15**(1), 6798 (2025)
11. Hochreiter, S., Schmidhuber, J.: LSTM can solve hard long time lag problems. In: Advances in Neural Information Processing Systems, vol. 9 (1996)
12. das Nações Unidas para a Infância, U.F.: O mapa da exclusão escolar em Maceió. UNICEF, Brasília (2020), disponível em: https://www.unicef.org/brazil/relatorios/plataforma-dos-centros-urbanos-2017-2020/mapa-exclusao-escolar-maceio
13. Kalasampath, K., Spoorthi, K., Sajeev, S., Kuppa, S.S., Ajay, K., Angulakshmi, M.: A literature review on applications of explainable artificial intelligence (XAI). IEEE Access (2025)
14. Lima, A.A.A.A.d., Sousa, F.P.d.: Censo escolar da educação básica: Uma referência para elaboração de políticas públicas e transferência de recursos para educação pública. Revista Com Censo, Vol. 1, n.1, 1ª ed., Brasília-DF, pp. 94–102, dezembro de 2014 (2014). Disponível em: http://www.cre.se.df.gov.br/ascom/documentos/suplav/revista_comcenso/artigo_censo_escolar_da_ed_basica.pdf
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), pp. 4765–4774. Curran Associates Inc. (2017)
16. Mathew, D.E., Ebem, D.U., Ikegwu, A.C., Ukeoma, P.E., Dibiaezue, N.F.: Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human. Neural Process. Lett. **57**(1), 16 (2025)

17. Phillips, P., Perron, P.: Testing for a unit root in time series regression. Biometrika **75**(2), 335–346 (1988)
18. Salih, A.M., et al.: A perspective on explainable artificial intelligence methods: shap and lime. Adv. Intell. Syst. **7**(1), 2400304 (2025)
19. Shojaie, A., Fox, E.B.: Granger causality: a review and recent advances. Annu. Rev. Stat. Appl. **9**(1), 289–319 (2022)
20. Souza, R.M.d.O., Oliveira, E.A.M.: O censo escolar no contexto da democratização da educação básica e do pacto federativo brasileiro. In: XVI ENDIPE - Encontro Nacional de Didática e Práticas de Ensino. vol. Livro 3. Junqueira&Marin Editores, Campinas (2012)
21. TATU, A.: Maceió é a capital com maior abandono escolar do país, aponta painel (2021). disponível em: https://www.agenciatatu.com.br/noticia/maceio-e-a-capital-com-maior-abandono-escolar-do-pais-aponta-painel/
22. Van der Velden, B.: Explainable AI: current status and future potential. Eur. Radiol. **34**, 1187–1189 (2024). https://doi.org/10.1007/s00330-023-10121-4
23. Wang, Q., Hsiao, Y.Y., Hushman, C., Armstrong, J.: The effectiveness of dropout intervention programs among k-12 students: a meta-analysis. J. Educ. Students Placed at Risk (JESPAR), pp. 1–31 (2024)



A Study of the Influence of the Attention Mechanism on Knowledge Tracing

Yuyan Wu[1(✉)], Miguel Arevalillo-Herráez[1], and David Arnau[2]

[1] Departament d'Informàtica, Universitat de València, Valencia, Spain
{yuyan.wu,miguel.arevalillo}@uv.es
[2] Departament de Didàctica de la Matemàtica, Universitat de València, Valencia, Spain
david.arnau@uv.es

Abstract. Knowledge Tracing (KT) enables adaptive learning systems to offer personalized learning experiences by modeling student learning behaviors. While recent studies in KT have shown promising results with attention-based models, the performance impact of the attention mechanism has not been sufficiently explored. To fill this gap, we aim to investigate the effects of different attention mechanisms in KT models. Through extensive experiments, we assess the strengths and limitations of five attention mechanisms across three benchmark datasets. Our results indicate that Talking-Heads Multi-Head attention outperforms the other four attention mechanisms, with the most significant improvements observed in terms of AUC, with an average performance increase exceeding 3% over the next best-performing method.

Keywords: Knowledge Tracing · Attention Mechanism · Adaptive Learning

1 Introduction

Adaptive learning is fundamental to intelligent education as it adapts to individual student differences and learning preferences, enabling the development of personalized tutoring strategies that optimize learning experience and outcomes. Knowledge Tracing (KT) is a crucial component of adaptive learning that assesses student abilities and predicts future performance based on their learning history. For example, knowledge tracing models predict students' correctness on a given exercise, allowing the system to personalize the learning path by providing targeted feedback or suggesting relevant resources.

The evolution of KT methods is closely linked to the changing demands for educational data. Early KT relied on probabilistic Bayesian models [5,11] and logistic approaches [1,15] due to limited educational data and computational resources. However, the widespread use of large-scale educational platforms and massive behavioral data has exposed the limitations of traditional methods: their shallow modeling cannot analyze complex learning patterns, hindering prediction

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 462–469, 2025.
https://doi.org/10.1007/978-3-031-98462-4_58

accuracy. Deep learning addresses this issue via nonlinear modeling and large parameterization, enabling the capture of complex behaviors and more accurate knowledge state prediction. Deep learning-based methods can be classified into three categories based on their modeling paradigms: sequential neural network knowledge tracing captures local temporal dependencies with hidden state propagation [3,16]; attention-based knowledge tracing uses self-attention for global weighting of historical interactions and long-term dependency modeling [7,8,12]; and graph-based knowledge tracing models the structured evolution of knowledge states using concept association graphs [6,18]. Among these, the attention-based knowledge tracing model stands out as a powerful baseline, excelling on mainstream knowledge tracing datasets due to its ability to parallelize sequence processing and flexibly weigh feature interactions.

Attention mechanisms are core to attention-based KT, but previous studies often change both input representation and attention mechanisms, making it hard to isolate their individual effects. To fill this gap, we systematically compare the performance of five attention mechanisms under controlled conditions, including both explored and unexplored approaches, to provide insights for optimizing attention-based KT methods.

2 Related Work

Attention-based knowledge tracing models use transformer modules to model students' learning sequences, evolving to incorporate more complex learning behaviors. Early models, such as the Self-Attentive Knowledge Tracing (SAKT) [13] and SimpleKT [12], employed multi-head self-attention to capture long-range dependencies, but did not explicitly model forgetting. To address this, context-aware Attentive Knowledge Tracing (AKT) [7], SparseKT [8], and Forgetting-aware Linear Bias (FoLiBi) [9] introduced various sparse attention mechanisms, recognizing that longer learning intervals increase the likelihood of forgetting. Relation-aware self-attention for Knowledge Tracing (RKT) [14] further enhanced KT by incorporating exercise text information into a relation-aware attention mechanism. The Hierarchical Transformer model for Session-aware Knowledge Tracing (HiTSKT) [10] incorporated session-level information for adaptability to rapid changes and proposed a power-law-decay attention mechanism. More recently, Zhan et al. [20] introduced Large Language Model for Knowledge Tracing (LLMKT), applied autoregressive LLMs to KT, and analyzed the effect of model size.

3 Methodology

Problem Definition. Let us assume a sequence of student learning interactions $S = \{(e_1, r_1), (e_2, r_2), \ldots, (e_{T-1}, r_{T-1})\}$, where e_t represents the exercise presented at time t, and r_t denotes the corresponding student response ($r_t = 1$ for correct and $r_t = 0$ for incorrect). The goal of KT is to predict the probability that the student will answer the exercise e_T correctly, i.e., $P(r_T = 1|S)$.

Input Representation. Due to the sparsity of exercises, we use concepts rather than exercises to represent student interactions. Specifically, let z_{c_k} be the latent representation of the concept c_k, r_{e_j} be the latent vector representing the students' response on exercise e_j, and y_t be the latent representation of the interaction at time t. We define:

$$z_{c_k} = W_c \cdot e_{c_k}, \quad r_{e_j} = W_e \cdot r_j, \quad y_t = z_{c_k} \oplus r_{e_j}, \tag{1}$$

where $W_c \in \mathbb{R}^{d \times ||C||}$ and $W_e \in \mathbb{R}^{d \times 2}$ are learnable linear transformation operations, and $\oplus$ denotes element-wise addition. Here, $||C||$ denotes the total number of concepts and d represents the hidden dimension.

Model Architecture. We use the standard architecture presented in [12], which consists of a two-layer transformer encoder and a prediction network. To study the effect of attention mechanisms, we replace the original attention component with different variants in our experiments.

Attention Mechanisms. We evaluate the effectiveness of standard multi-head attention [19], attention with linear bias multi-head attention [9], monotonic multi-head attention [7], sparse multi-head attention [4,8], and talking-heads multi-head attention [17].

Standard Multi-Head Attention (MHA) was introduced by Vaswani et al. [19], computing multiples scaled dot-product attention heads in parallel. In this method, the scaled dot-product attention is computed as follows:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, V represent the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors.

Monotonic Multi-Head Attention (MMHA), proposed for KT by Ghosh et al. [7], introduced temporal monotonicity constraints by applying an exponential decay to the attention weights, which reduced the influence of distant exercises, as shown in the following equation:

$$\text{softmax}\left(\frac{\exp\left(-\theta \cdot d(t,\tau)\right) q_t k_\tau^T}{\sqrt{d_k}}\right) \tag{3}$$

where $\theta \in \mathbb{R}^+$ is a trainable decay coefficient, $d(t, \tau)$ denotes the temporal distance between the current time steps t and τ.

Attention with Linear Bias Multi-Head Attention (AMHA) enhances local attention by adding position-specific bias terms based on relative temporal distances, mimicking cognitive forgetting patterns, making it ideal for KT tasks. In the KT-specific implementation FoLiBi [9], the attention weights are calculated as:

$$\text{softmax}\left(\frac{q_t k_\tau^T + \lambda_h \cdot (t - \tau)}{\sqrt{d_k}}\right), \tag{4}$$

where λ_h controls the impact of forgetting behavior on the h-th attention head out of H heads, t is the current time step, and $\tau \in \{1, \ldots, t-1\}$ represents a past time step.

Sparse Multi-Head Attention (SMHA) improves computational efficiency by selecting only the top-k most relevant attention weights, effectively reducing noise from redundant connections in MHA [4]. Huang et al. [8] proposed a KT-optimized variant called top-k sparse attention, which operates as follows:

$$softmax\left(\frac{QK^T \odot M}{\sqrt{d_k}}\right), \tag{5}$$

where $\odot$ represents the element-wise product operator and M is a binary mask defined as:

$$M_{i,j} = \begin{cases} 1 & \text{if } (QK^\top/\sqrt{d_k})_{i,j} \text{ ranks in top-}k \\ -\infty & \text{otherwise} \end{cases} \tag{6}$$

Talking-heads Multi-Head Attention (TMHA) enhances MHA performance by leveraging cross-head information [17]. This approach achieves this by introducing two learnable projection matrices before and after the softmax operation:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} W_{pre}\right) W_{post} V \tag{7}$$

where W_{pre} and W_{post} transform attention scores across heads before and after softmax.

4 Experimental Setting

Datasets. To evaluate the effectiveness of different attention mechanisms, we conduct experiments on three real-world educational datasets: ASSISTments2012[1], ASSISTments2017[2] and Junyi. The first two datasets originate from the ASSISTments online math tutoring system, while the Junyi dataset [2] was collected from Junyi Academy, a Chinese e-learning platform founded in 2012 using open-source code from Khan Academy. Following prior works [7,8,12], we fixed the student response sequence length to 200, discarding sequences with fewer than five elements. Longer sequences were split into non-overlapping chunks of up to 200, while shorter ones were right-padded with zeros. The number of sequences after preprocessing is presented in the "# of Sequences" row of Table 1. For computational efficiency, we randomly sampled 5 000 out of the 247 606 students from the Junyi dataset. Dataset statistics before and after preprocessing are summarized in Table 1.

[1] https://sites.google.com/site/assistmentsdata/2012-13-school-data-with-affect.
[2] https://sites.google.com/view/assistmentsdatamining/.

Table 1. Data statistics. "raw" and "preprocessed" refer to the initial and preprocessed data statistics

Datasets	ASSISTments2012		ASSISTments2017		Junyi	
	raw	preprocessed	raw	preprocessed	raw	preprocessed
# of Students	46 674	25 266	1 709	1 708	247 606	5 000
# of Quest.	179 999	50 918	3 162	3 162	721	700
# of KCs	256	245	102	102	39	39
# of Interactions	6 123 270	2 621 015	942 816	942 711	25 925 984	1 515 140
# of Sequences	–	30 991	-	5 516	–	11 038
Correct Label rate	67.68 %	69.65 %	37.27 %	37.27 %	82.79 %	82.86 %

Evaluation Metrics and Implementation Details. We evaluate model performance using four standard metrics commonly adopted in KT research: Area Under the ROC Curve (AUC) for probabilistic ranking accuracy, Accuracy (ACC) for binary classification correctness, the F1-score for balancing precision and recall under class imbalance, and Root Mean Squared Error (RMSE) for measuring continuous prediction errors. We apply 5-fold cross-validation, using 20% of the data for testing, and the remaining 80% split 9:1 into training and validation sets. The model architecture employs a hidden state dimension of 256 and 8 attention heads. Key hyperparameters include a batch size of 128, learning rates of 10^{-3} and 10^{-4} selected via grid search, a dropout rate of 0.2 on hidden layers, and a maximum of 200 training epochs with early stopping after 10 epochs of no improvement.

5 Results

Table 2 presents the 5-fold average of different attention mechanisms across three datasets, with the best results highlighted in bold. Our results demonstrate that TMHA achieves superior performance across all datasets. Taking MHA as the baseline model, TMHA improves AUC by 4.1%, 3.3%, and 1.8% on ASSISTments2012, ASSISTments2017, and Junyi, respectively.

Additionally, the performance of sparse attention is lower than that of MHA, possibly due to the sparsity of the top-k selection strategy, which may be inadequate in this case (we set k=5). Regarding the results of AMHA and MMHA, both methods outperform MHA on the ASSISTments datasets but perform worse on the Junyi dataset. Overall, modifying the attention weight mechanism alone does not lead to a significant improvement in performance, as the results across different methods remain relatively similar.

From the perspective of dataset fitting, attention-based models exhibit good performance on the ASSISTments2012 and Junyi datasets, with F1-scores remaining around 0.8 and 0.9, even under imbalanced classification conditions. However, there is still room for improvement in ASSISTments2017, as the F1-score is around 0.5 for all tested methods.

Table 2. The overall performance comparison on the three real-world datasets.

Datasets	Metrics	MHA	AMHA	MMHA	SMHA	TMHA
ASSISTments2012	AUC	0.7338	0.7390	0.7362	0.7289	**0.7749**
	ACC	0.7370	0.7395	0.7385	0.7357	**0.7512**
	F1-score	0.8297	0.8307	0.8295	0.8315	**0.8351**
	RMSE	0.4230	0.4215	0.4222	0.4246	**0.4096**
ASSISTments2017	AUC	0.6977	0.7071	0.7032	0.6849	**0.7309**
	ACC	0.6790	0.6832	0.6827	0.6755	**0.6965**
	F1-score	0.4684	0.4760	0.4596	0.4403	**0.5258**
	RMSE	0.4547	0.4521	0.4527	0.4579	**0.4447**
Junyi	AUC	0.7513	0.7467	0.7493	0.7459	**0.7690**
	ACC	0.8452	0.8432	0.8443	0.8448	**0.8474**
	F1-score	0.9128	0.9114	0.9123	0.9129	**0.9141**
	RMSE	0.3455	0.3476	0.3460	0.3463	**0.3416**

In addition, TMHA introduces complexity in interpretability by allowing interactions between heads, making it harder to assess each head's contribution through attention weights alone. This reduces its interpretability compared to other approaches. With regard to model complexity, TMHA has a slight increase in learnable parameters, adding only 288 more than the MHA model, which has a negligible impact on computational cost. For example, the talking-heads attention model used in our experiments had over 800 000 parameters.

6 Conclusion

In this work, we evaluated five attention mechanisms under the same conditions and a shared base architecture for knowledge tracing. Experiments on three large-scale educational datasets (ASSISTments2012, ASSISTments2017, and Junyi) show that Talking-Heads Multi-Head attention consistently outperforms the others. In contrast, methods with forgetting mechanisms based on interaction distances offer no clear gains over the baseline, likely because input data limited to concept IDs and binary responses fails to capture forgetting effectively. Time-related features such as lag time and interaction duration may provide better representations of these temporal dynamics. While large datasets can support complex models, many real-world applications face data limitations, particularly in cold-start scenarios where new users or items lack sufficient historical data, which hinders scalability. Transparency in model decision-making also remains a key challenge for deep learning approaches. These insights motivate our future work on Talking-Heads Multi-Head attention models that aim to balance accuracy with interpretability.

Acknowledgments. This research has been supported by project TED2021-129485B-C42, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR; project PID2023-150960NB-I00, funded by the Spanish Ministry of Science, Innovation and Universities and the European Union; project AICO/2023/065, funded by Valencian Regional Government (Spain); and grant ACIF/2021/439, funded by the Conselleria de Educación, Universidades y Empleo de la Generalitat Valenciana.

References

1. Cen, H., Koedinger, K., Junker, B.: Learning factors analysis – a general method for cognitive model evaluation and improvement. In: Ikeda, M., Ashley, K.D., Chan, T.W. (eds.) Intelligent Tutoring Systems. pp. 164–175. Springer Berlin Heidelberg, Berlin (2006). https://doi.org/10.1007/11774303_17
2. Chang, H.S., Hsu, H.J., Chen, K.T., et al.: Modeling exercise relationships in e-learning: a unified approach. In: EDM, pp. 532–535 (2015)
3. Chen, J., Liu, Z., Huang, S., Liu, Q., Luo, W.: Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations. In: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2023/IAAI 2023/EAAI 2023. AAAI Press (2023). https://doi.org/10.1609/aaai.v37i12.26661
4. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509 (2019)
5. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User Model. User-Adap. Inter. **4**, 253–278 (1994)
6. Cui, C., et al.: Dgekt: a dual graph ensemble learning method for knowledge tracing. ACM Trans. Inform. Syst. **42**(3) (Jan 2024). https://doi.org/10.1145/3638350
7. Ghosh, A., Heffernan, N., Lan, A.S.: Context-aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2020, pp. 2330–2339. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3394486.3403282
8. Huang, S., Liu, Z., Zhao, X., Luo, W., Weng, J.: Towards robust knowledge tracing models via k-sparse attention. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, pp. 2441–2445. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3539618.3592073
9. Im, Y., Choi, E., Kook, H., Lee, J.: Forgetting-aware linear bias for attentive knowledge tracing. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, pp. 3958–3962. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3583780.3615191
10. Ke, F., et al.: Hitskt: a hierarchical transformer model for session-aware knowledge tracing. Knowl.-Based Syst. **284**, 111300 (2024). https://doi.org/10.1016/j.knosys.2023.111300
11. Käser, T., Klingler, S., Schwing, A.G., Gross, M.: Dynamic bayesian networks for student modeling. IEEE Trans. Learn. Technol. **10**(4), 450–462 (2017). https://doi.org/10.1109/TLT.2017.2689017

12. Liu, Z., Liu, Q., Chen, J., Huang, S., Luo, W.: simplekt: a simple but tough-to-beat baseline for knowledge tracing. arXiv preprint arXiv:2302.06881 (2023)
13. Pandey, S., Karypis, G.: A self-attentive model for knowledge tracing. In: 12th International Conference on Educational Data Mining, EDM 2019, pp. 384–389. International Educational Data Mining Society (2019)
14. Pandey, S., Srivastava, J.: Rkt: Relation-aware self-attention for knowledge tracing. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM 2020, pp. 1205–1214. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3340531.3411994
15. Pavlik, P.I., Cen, H., Koedinger, K.R.: Performance factors analysis –a new alternative to knowledge tracing. In: Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling, pp. 531–538. IOS Press, NLD (2009)
16. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., Sohl-Dickstein, J.: Deep knowledge tracing. In: Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1, NIPS 2015, pp. 505–513. MIT Press, Cambridge, MA, USA (2015)
17. Shazeer, N., Lan, Z., Cheng, Y., Ding, N., Hou, L.: Talking-heads attention. arXiv preprint arXiv:2003.02436 (2020)
18. Tong, S., Liu, Q., Huang, W., Hunag, Z., Chen, E., Liu, C., Ma, H., Wang, S.: Structure-based knowledge tracing: an influence propagation view. In: 2020 IEEE International Conference on Data Mining (ICDM), pp. 541–550 (2020). https://doi.org/10.1109/ICDM50108.2020.00063
19. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)
20. Zhan, B., et al.: Knowledge tracing as language processing: a large-scale autoregressive paradigm. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education, pp. 177–191. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-64302-6_13



Similarity-Based Scoring Model for Handwritten Answers in Japanese Workbooks

Takahiro Saito[1], Hung Tuan Nguyen[1], Kanako Komiya[1(✉)], Tsunenori Ishioka[2], and Masaki Nakagawa[1]

[1] Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan
s245145v@st.go.tuat.ac.jp, {kkomiya,fx7297}@go.tuat.ac.jp, nakagawa@cc.tuat.ac.jp

[2] National Center for University Entrance Examinations, 2-19-23 Komaba, Meguro-ku, Tokyo 153-8501, Japan
tunenori@rd.dnc.ac.jp

Abstract. This paper proposes a similarity-based grading system for handwritten answers to short-answer questions in Japanese workbooks. While previous studies have predominantly focused on models trained specifically for individual questions, this study utilizes a similarity-based approach, in which both the student's answer and the model answer are input into the model, allowing it to learn the semantic differences between them. This approach enables the model to handle multiple questions within a single training session, thereby improving grading accuracy even with a limited dataset and allowing the system to effectively score answers to unseen questions. We utilized three types of foundational models: a Japanese pre-trained BERT model, a Japanese pre-trained Sentence-BERT model, and ChatGPT. For training and evaluation, answer data from Japanese workbooks used by junior high school students was employed. The experimental results demonstrated that similarity-based scoring using a model answer is effective for grading short-answer questions in Japanese workbooks. In addition, the model achieved satisfactory results even when applied to text automatically recognized from handwritten answers to unseen questions, demonstrating its practical applicability in real-world educational settings.

Keywords: Scoring · Japanese · Similarity · Handwritten · ChatGPT

1 Introduction

Recently, excessive educator workload has emerged as a societal issue in Japan, leading to initiatives to automate grading using NLP technologies. Traditional automatic grading methods use separate models for each question, requiring substantial training data per question and limiting practical application in schools.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 470–477, 2025.
https://doi.org/10.1007/978-3-031-98462-4_59

This study proposes a similarity-based grading model that evaluates differences between student and model answers, leveraging BERT [8], Sentence-BERT [16], and ChatGPT as foundational models. The method targets short-answer questions with handwritten answers, using a model answer as reference. This approach supports simultaneous training on multiple questions, enhances grading accuracy with limited data, and enables scoring of unseen questions.

We used answer data collected from Japanese-language workbooks for junior high school students in Japan for training and evaluation. The dataset contains both manually transcribed and automatically recognized handwritten answers, supporting robust system performance in practical scenarios.

We evaluated the system's scoring of short-answer questions. Results show that similarity-based scoring with a model answer offers reliable grading, even for automatically transcribed handwritten answers. These findings demonstrate the model's potential for practical use in educational contexts.

The contributions of this paper are: (1) This study introduces the first similarity-based scoring method for grading diverse short-answer questions in Japanese workbooks. (2) We hypothesized and empirically confirmed that this system effectively grades answers to unseen questions. (3) Experimental results verify the effectiveness of the similarity-based method for grading short-answer questions, including those from automatically recognized handwritten text.

2 Related Work

Oka et al. [15] addressed automatic grading of handwritten Japanese short-answer questions by integrating $\text{BERT}_{\text{base}}$ with a handwritten character recognition system developed by Hung et al. [13], constructing question-specific models using approximately 60,000 answers per item.

Automatic grading has also been explored not only in short-answer questions [1,3,17], but also in essay scoring [7,9–12,18], and mathematics word problems [6], where most approaches build separate models for each question. Such approaches require substantial training data per item, limiting practical adoption in schools outside large-scale assessments.

Another line of research involves similarity-based scoring, which has gained attention as an alternative to conventional instance-based methods. Bexte et al. [2,4,5] introduced an efficient similarity-based scoring architecture and demonstrated, among other things, that it achieves performance comparable to BERT classification models on the Automated Student Assessment Prize (ASAP) dataset.

However, these studies primarily targeted known questions. This study demonstrates that similarity-based scoring enables the grading of unseen questions with a single model trained on multiple tasks.

3 Data

We extracted 25 short-answer questions from three Japanese junior high school workbooks, creating a dataset of 1,547 student answers (55–66 per question)[1]. After removing invalid answers, 1,372 answers remained for analysis. Each answer is represented as a tuple: (model answer, student answer, correctness label). The dataset includes both manually and automatically transcribed answers, with the latter generated using iLabo's handwriting recognition system[2], fine-tuned on private handwriting data for unstable handwriting [14].

The character recognition rate (CRR) and word recognition rate (WRR) for automatically transcribed answers are 74.68% and 51.18%, respectively, which are defined as one minus the character error rate (CER) and word error rate (WER). CER and WER are calculated using the Levenshtein distance[3] divided by the total number of characters or words in the reference answer. Word units for WER are obtained via MeCab[4], a Japanese morphological analyzer.

Correctness labels (1 or 0) were labeled by teachers, with 64.5% of dataset answers deemed correct. The dataset was partitioned into five sets of five questions each and split 3:1:1 into training, validation, and evaluation sets, enabling grading accuracy assessment on previously unseen questions.

4 Methods

In the BERT-based method, each answer pair was concatenated and tokenized to produce a token ID sequence, by inserting a [SEP] token between the model and student answers. This sequence was fed into BERT, yielding a 768-dimensional embedding. Following Oka et al. [15], [CLS] token embeddings from layers 9–12 were extracted and concatenated, resulting in a 3,072-dimensional representation. This vector was fed into a linear classifier (3,072 to 2 dimensions), followed by a Softmax function. If the output exceeded the predefined threshold (0.5 in this study), the answer was classified as correct; otherwise as incorrect.

In the Sentence-BERT method, each answer was tokenized separately to generate individual token ID sequences and then fed into BERT, resulting in two 768-dimensional embeddings. Sentence representations were computed by averaging the token embeddings across all dimensions. The cosine similarity between the sentence vectors of the model and student answers was calculated; answers with similarity above the predefined threshold (0.5 in this study) were classified as correct; otherwise as incorrect.

In the ChatGPT-based approach, we employed GPT-4o and GPT-4o-mini through OpenAI's official Python library[5]. Grading utilized the following prompt.

[1] The answers were collected with the approval of the University's Ethics Review Committee for Research Involving Human Subjects (No.220707-04111).
[2] https://ilabo.biz/wp/.
[3] The minimum number of operations needed to transform one string into another.
[4] https://github.com/SamuraiT/mecab-python3.
[5] https://pypi.org/project/openai/, version 1.59.6.

```
You are a grader. Read the provided model answer and the student's
answer. If you consider the student's answer correct, output ‘‘1";
if you consider it incorrect, output ‘‘0". The output must be a
single character, either ‘‘1" or ‘‘0". Even if the judgment is
difficult, choose and output the option that seems more likely.
```

5 Experiments

5.1 Training

During training, each foundational model was fine-tuned to develop a scoring system. For BERT[6] and Sentence-BERT[7], AdamW was used as the optimizer and BCELoss as the loss function. The learning rate was determined via grid search[8] and scheduled with a cosine annealing strategy, gradually decaying the learning rate to zero. The batch size was 16, and the number of epochs was 20. All other parameters were set to default values. For both GPT-4o and GPT-4o-mini[9], fine-tuning used only default settings.

Model performance was evaluated using accuracy (ACC), the proportion of instances where the model correctly predicted the answer's label.

5.2 Options

We conducted two experiments: one using manual and another using automatic transcriptions of student answers. The aim was to assess the grading performance of four models—BERT, Sentence-BERT, GPT-4o, and GPT-4o-mini—across both transcription types. Training data included both manual and automatic transcriptions. For Sentence-BERT and ChatGPT, zero-shot performance was also evaluated. Overall, we compared 11 methods, as detailed in Table 1, which summarizes the experimental setups and results.

6 Results

Table 1 presents the macro-averaged accuracies over 5-fold trials for manual and automatic transcriptions. The best results are underlined, and each foundational model's highest accuracy is in bold. BERT and Sentence-BERT experiments were repeated five times under identical conditions, and averages are reported. ChatGPT models were evaluated only once.

In the grading of the manually transcribed answers, the best-performing method was **GPT_4o-auto**, followed by **GPT_4o-manual**. In the grading of

[6] https://huggingface.co/cl-tohoku/bert-base-japanese-v3.
[7] https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2.
[8] 5e-6 for **BERT-manual**, 2e-6 for **BERT-auto**, 1e-6 for two SBERT-based methods.
[9] `gpt-4o-2024-08-06` and `gpt-4o-mini-2024-07-18`.

Table 1. Accuracies of Scoring Systems

Methods	Foundational models	Training data	ACC for Manual Transcription	ACC for Auto Transcription
BERT-manual	BERT	Manual	**75.8%**	68.1%
BERT-auto		Automatic	75.1%	**71.3%**
SBERT-zero	Sentence-BERT	N/A	75.0%	**70.6%**
SBERT-manual		Manual	74.6%	67.9%
SBERT-auto		Automatic	**75.3%**	69.5%
GPT_4o-zero	GPT_4o	N/A	76.2%	50.8%
GPT_4o-manual		Manual	81.9%	75.7%
GPT_4o-auto		Automatic	**82.6%**	**81.1%**
GPT_4o-mini-zero	GPT_4o-mini	N/A	72.9%	52.4%
GPT_4o-mini-manual		Manual	79.9%	75.6%
GPT_4o-mini-auto		Automatic	**81.1%**	**78.2%**

the automatically transcribed answers, **GPT_4o-auto** also achieved the best performance, followed by **GPT_4o-mini-auto**.

When using **GPT_4o-auto**, the grading accuracy reached 82.6% for manually transcribed answers and 81.1% for automatically transcribed answers. Given that the model had never encountered answers to the same question during the training phase, these results are remarkable.

As a reference, the accuracy of the grading done by parents compared to the gold labels assigned by teachers was 87.3%. The gap between our automated grading system and humans is narrowing considerably.

7 Discussion

First, we examine the grading of manually transcribed answers in the left column of Table 1. For all four foundational models, there was no significant difference[10] between models trained on manually versus automatically transcribed data. This finding is consistent with Oka et al. [15] and likely indicates that fine-tuning enables models to accommodate handwriting recognition errors, minimizing their impact on scoring performance.

Next, we assessed the impact of fine-tuning by comparing zero-shot and fine-tuned models. Fine-tuning yielded significant accuracy improvements for both ChatGPT models, but not for Sentence-BERT, likely because Sentence-BERT is already trained to assess sentence similarity.

We then compared the performance of different foundational models. Regardless of training data, fine-tuned ChatGPT models significantly outperformed both BERT-based models. In the zero-shot setting, no significant difference was found between Sentence-BERT and ChatGPT models. Similarly, BERT

[10] All statistical significance was assessed via the chi-square test at the 0.05 level.

and Sentence-BERT showed no significant difference, indicating limited distinction between the two BERT variants. Comparing GPT_4o and GPT_4o-mini, fine-tuning removed any significant difference; however, in the zero-shot setting, GPT_4o significantly outperformed GPT_4o-mini. This outcome aligns with the expected ranking of foundational models in terms of their capability.

Next, we examine the evaluation results for automatically transcribed answers in the right column of Table 1. With the exception of GPT_4o-mini, fine-tuning on automatically transcribed data significantly outperformed fine-tuning on manually transcribed data. This suggests that fine-tuning solely with clean text is insufficient to address recognition errors in transcribed answers.

Furthermore, we also analyzed the effect of fine-tuning compared to zero-shot performance. In both ChatGPT models, significant performance improvements were obtained through fine-tuning regardless of the type of training data. However, as with the manual transcription data scoring, Sentence-BERT did not show improved scoring performance.

We also investigated foundational model differences. ChatGPT generally achieved higher accuracy, but in the zero-shot setting, **SBERT-zero** outperformed the ChatGPT models. When comparing BERT and Sentence-BERT, BERT was superior only after fine-tuning on automatic transcription data. For GPT_4o versus GPT_4o-mini, unlike in manual transcription scoring, no significant difference emerged regardless of fine-tuning. This suggests that, since foundational models are pre-trained on clean text, GPT_4o's enhanced pretraining does not confer an advantage for noisy data.

Finally, we compare the scoring performance for manually and automatically transcribed answers. In both experiments, **GPT_4o-auto** outperformed all other methods. Despite differences in the test data, there was no significant difference in performance between two experiments. However, for **GPT_4o-manual**, accuracy was significantly higher when evaluating manually transcribed answers than automatically transcribed ones. This indicates that fine-tuning on noisy data does not degrade performance on clean data, but scoring noisy data effectively requires training on similar noise levels whenever possible.

8 Conclusion

We proposed a similarity-based scoring system for evaluating short-answer questions in Japanese workbooks. By comparing student and model answers, the system can grade previously unseen questions. We conducted experiments with two data types: manually and automatically transcribed handwritten answers. Four foundational models were assessed in both zero-shot and fine-tuned settings. The results showed: (1) Similarity-based methods were effective in evaluating answers to unknown problems, achieving a maximum accuracy rate of 82.6%. (2) Zero-shot accuracy was low for automatically transcribed data, but fine-tuning markedly improved performance. (3) Fine-tuning on noisy data does not diminish performance for clean data, but scoring noisy data requires training for noisy data. These findings underscore the system's applicability in real-world educational contexts.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Numbers 23H03511 and 24H00738. We would like to thank Toshihiko Horie from Wacom Co., Ltd., who gave us the text data set of questions.

References

1. Asazuma, Y., Funayama, H., Matsubayasi, Y., Mizumoto, T., Inui, K.: Verification of consistency with scoring criteria for the descriptive answer grading model. In: 29th The Association for Natural Language Processing, pp. 1868–1873 (2023)
2. Bexte, M., Horbach, A., Schützler, L., Christ, O., Zesch, T.: Scoring with confidence? – exploring high-confidence scoring for saving manual grading effort. In: Kochmar, E., Bexte, M., et al. (eds.) Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), pp. 119–124. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), https://aclanthology.org/2024.bea-1.11/
3. Bexte, M., Horbach, A., Zesch, T.: Implicit phenomena in short-answer scoring data. In: Roth, M., Tsarfaty, R., Goldberg, Y. (eds.) Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language, pp. 11–19. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.unimplicit-1.2, https://aclanthology.org/2021.unimplicit-1.2
4. Bexte, M., Horbach, A., Zesch, T.: Similarity-based content scoring - how to make S-BERT keep up with BERT. In: Kochmar, E., Burstein, J., Horbach, A., Laarmann-Quante, R., Madnani, N., Tack, A., Yaneva, V., Yuan, Z., Zesch, T. (eds.) Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pp. 118–123. Association for Computational Linguistics, Seattle, Washington (Jul 2022). https://doi.org/10.18653/v1/2022.bea-1.16, https://aclanthology.org/2022.bea-1.16/
5. Bexte, M., Horbach, A., Zesch, T.: Similarity-based content scoring - a more classroom-suitable alternative to instance-based scoring? In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, pp. 1892–1903. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.119, https://aclanthology.org/2023.findings-acl.119/
6. Cahill, A., Fife, J.H., Riordan, B., Vajpayee, A., Galochkin, D.: Context-based automated scoring of complex mathematical responses. In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 186–192 (2020)
7. Carvalho, R., Lins, L.F., Rodrigues, L., Miranda, P., Oliveira, H., Cordeiro, T., Bittencourt, I.I., Isotani, S., Mello, R.F.: Exploring nlp and embedding for automatic essay scoring in the portuguese. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, pp. 228–233. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-64315-6_18
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT, pp. 4171–4186 (2019)

9. Feng, H., et al.: Leveraging large language models for automated chinese essay scoring. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education, pp. 454–467. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-64302-6_32
10. Galhardi, L., et al.: Contextual features for automatic essay scoring in portuguese. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. pp. 270–282. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-64315-6_23
11. Latif, E., Fang, L., Ma, P., Zhai, X.: Knowledge distillation of llms for automatic scoring of science assessments. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, pp. 166–174. Springer Nature Switzerland, Cham (2024).https://doi.org/10.1007/978-3-031-64312-5_20
12. Li, W., Liu, H.: Applying large language models for automated essay scoring for non-native Japanese. Palgrave Commun. **11**(1), 1–15 (December 2024) https://doi.org/10.1057/s41599-024-03209-, https://ideas.repec.org/a/pal/palcom/v11y2024i1d10.1057_s41599-024-03209-9.html
13. Nguyen, H.T., Nguyen, C.T., Nakagawa, M.: Icfhr 2018 – competition on vietnamese online handwritten text recognition using hands-vnondb (vohtr2018). In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 494–499 (2018). https://doi.org/10.1109/ICFHR-2018.2018.00092
14. Nguyen, H.T., Truong, T.N., Ly, N.T., Nakagawa, M., Horie, T.: Automatic scoring for handwritten answers from elementary to junior high school grades in japan. In: IEICE Tech. Rep. IEICE, Japan (2025)
15. Oka, H., Nguyen, H.T., Nguyen, C.T., Nakagawa, M., Ishioka, T.: Fully automated short answer scoring of the trial tests for common entrance examinations for japanese university. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education, pp. 180–192. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-11644-5_15
16. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1410, https://aclanthology.org/D19-1410
17. Sato, T., Funayama, H., kazuaki, H., Asazuma, Y., Inui, K.: Explanation of the automatic grading results based on the reference sections. In: 28th The Association for Natural Language Processing, pp. 459–464 (2022)
18. Yoshida, L.: The impact of example selection in few-shot prompting on automated essay scoring using gpt models. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. pp. 61–73. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-64315-6_5



Exploring Generative AI in Learning: Neurodivergent and Disabled Students' Voices in Higher Education

Yun He(✉), Elaine Chen, and Richard Fletcher

Nottingham Trent University, Nottingham, UK
{yun.he,elaine.chen,richard.fletcher}@ntu.ac.uk

Abstract. In recent years, the number of neurodivergent and disabled students enrolled in universities in the UK has increased. These students often face challenges related to cognitive processing, self-regulation, and social acceptance. Generative AI (GenAI) chatbots have gained significant attention and show potential to offer personalised support to these students. Despite this growing focus on GenAI, the perspectives of disabled and neurodivergent students remain largely overlooked. This paper amplifies the voice of these students through a survey in one of the largest UK universities. It explores the way they perceive and use GenAI to support their studies, and how GenAI complements the institutional support that they can receive. We develop a framework based on information processing theory for analysing open-ended questions of the survey. The analysis confirmed the biggest challenges faced by these students and their primary concerns about GenAI.

Keywords: Generative AI · neurodiversity · specific learning difficulties

1 Introduction

In 2021/22, 16% of students enrolled in UK higher education institutions (HEIs) reported having a known disability. Among these, 32% of the students had a specific learning difficulty (SpLD) such as dyslexia, dyspraxia or ADHD [13]. Current interventions for supporting neurodivergent and disabled students in higher education are mainly: assistive technologies to compensate students' difficulties with reading and writing, direct one-to-one instructional assistance for students to discuss their difficulties with a personal peer tutor or professional, strategic instructions to improve their cognitive process and self-efficacy, and comprehensive support programs integrating various interventions to promote their overall academic skills [25].

While these interventions are helpful, some of them follow a "one-size-fit-all" strategy which does not fit individual student needs [5,16]. AI-driven tools can help address individual learning challenges and promote academic engagement without students explicitly disclosing their learning difficulties [2]. They

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 478–485, 2025.
https://doi.org/10.1007/978-3-031-98462-4_60

can offer capabilities such as adaptive learning platforms that analyse students' strengths and areas for improvement, to deliver personalised content that enhances their learning experience [8]. The rise of generative artificial intelligence (GenAI) has further revolutionised this support by introducing features such as real-time text-to-speech tools, content summarisation, grammar assistance, and note-taking aids [3,7]. Using these innovations, learning can be made more accessible and engaging for neurodivergent and disabled students.

In this study, we apply information processing theory, a cognitive theory that emphasises how people receive, process and retrieve information [21], to analyse how current GenAI tools can affect the ability of neurodivergent and disabled students to encode, store and process knowledge. This framework has been widely applied in research to understand learners' individual cognitive process and learning preference [17,18,22–24]. We extend the theory with an additional layer of interaction between learners' internal state with the external environment and explore how GenAI affects users' cognitive load, working memory, and ability to engage with information effectively. Applying this perspective, we examine the benefits and limitations of current GenAI tools to support the diverse learning needs of these students.

This paper explores (1) how students with neurodiversity and disabilities in higher education use GenAI in their learning and (2) how GenAI tools address the learners' challenges. It starts with a brief review of the literature on AI for supporting neurodivergent and disabled students in higher education, and the theoretical underpinning in Sect. 2. Then, Sect. 3 presents the methodological development followed by the main results in Sect. 4. It finishes with a discussion and conclusion in Sect. 5.

2 Related Works

Various technologies have been developed to support students with learning difficulties and disabilities, under the names of AI, assistive technologies, and smart/intelligent tutoring systems. They can support students' intellectual and emotional development, facilitate the collection, evaluation, and communication of data from different stakeholders, enhance the enjoyment of teaching and learning process, or improve diagnostic processes [20]. Recently, the rise of GenAI is transforming how these students engage with learning by offering personalised content creation and adaptive communication. GenAI tools can generate human-like texts and images in response to users' prompts, which offers new landscape for teaching and learning. Various research revealed that students in higher education generally have a positive attitude towards GenAI in teaching and learning [4,12]. They see GenAI tools as a study companion, using them for grammar checks, clarifying complex concepts, and receiving immediate feedback on their work. Although there has been extensive research on the benefits and challenges of integrating GenAI in higher education, the perspectives of learners with disabilities remain largely underexplored.

Not all GenAI tools were designed to serve disabled users effectively [1]. Neurodivergent students who used GenAI tools felt that content accessibility could be improved [15]. Further accessibility tests indicated that GenAI tools were not trained with accessibility in mind [10]. GenAI tools that fail to accommodate varied cognitive styles present challenges to neurodivergent and disabled students. Complex interfaces and unclear feedback can increase burden on working memory and attention, decreasing comprehension and engagement.

3 Methods

We designed a survey to collect primary data from one of the largest universities in the UK, with about 9500 students with neurodiversity, disability or having other long term health conditions. The questionnaire was co-designed with a dyslexic student and an accessibility specialist, after institutional ethics approval. The 53-question survey covers demographics, GenAI awareness, use for learning support, non-academic advice, teaching, assessment, AI detection, plagiarism, and employment. Data were collected using Microsoft Forms between March and May 2024. A pilot with six additional questions tested the survey before the April–May 2024 rollout.

54 valid responses were collected from a range of dyslexic, neurodivergent and disabled students, statistically representative of the targeted student population in the university under study. Of the respondents, 24 reported dyslexia, 20 autism, 17 ADHD, 17 mental health conditions, 8 other long-term health issues, and 3 other SpLDs. Non-parametric statistical analysis (Kendall τ, Fisher's exact test and Mann-Whitney U test) was used to test for independence and differences between groups of variables [19]. The quantitative analysis was performed using Python and MS Excel. A thematic coding of open-ended questions was carried out and cross-validated by separate team members.

A preliminary framework has been developed to further analyse the coded answers based on information processing theory and basic concepts in memory and learning [6,21], with the aim of analysing student' usage and experience of GenAI in relation to their specific learning difficulties [11] and consideration of the Wechsler Adult Intelligence Scale IV. This framework models each student as an information processing agent exchanging information with their external learning environment. Internal to the learner are their learning difficulties such as limited working memory and processing speed. External to the learner is their environment, which can be either physical environment such as building, classroom, devices etc., or people surrounding them, such as their peers, teachers or other social aspects. Their daily learning tasks are categorised into inputs and outputs. Inputs include searching for information and receiving visual and auditory information, representing the information flow from the external environment to the learner. Outputs include writing and presenting, representing the information flow from the learner to the external environment. To manage their learning task despite their learning difficulties, coping strategies and self-organisation methods are developed as an interface between the learner and their

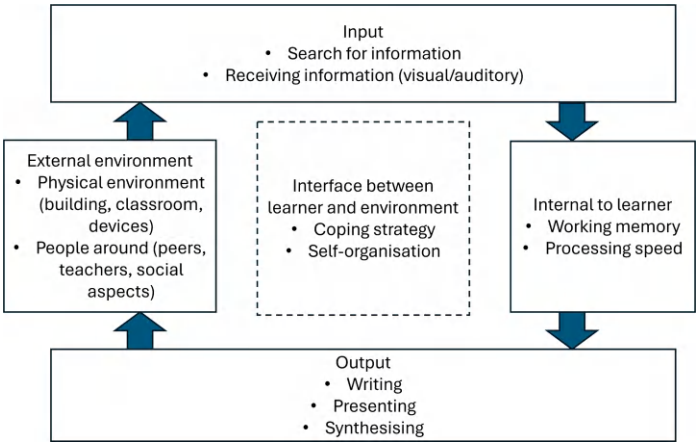


Fig. 1. Analytical framework based on information processing theory

environment. This framework is depicted in Fig. 1, where the arrows represent the information flow.

4 Main Results

We found that while 54% of respondents considered themselves to be 'confident' computer users, only 26% considered themselves confident in GenAI. Most respondents (65%) trust the outputs of GenAI 'some of the time' with the remainder being split between trusting most of the time (20%) and hardly ever or never trusting (15%). There is no significant evidence showing that more frequent users of GenAI feel more confident in using it ($\tau = 0.0792, p = 0.482$), but there is a positive correlation between frequency of use and user's trust in the outputs of GenAI ($\tau = 0.334, p < 0.005$). The most acceptable uses of GenAI are explaining concepts, suggesting research ideas, and summarising articles, with acceptance rates outpacing actual usage rates, consistent with previous findings with neurotypical students [9].

Table 1 outlines themes from students' perceived challenges, benefits, risks, and GenAI use cases based on the framework in Fig. 1. By aggregating the challenges under the analytical framework, we can see that students generally face challenges in input and output, while students with ADHD, mental health conditions and dyslexia also have more difficulties interacting with their environment. Dyslexic students' internal information processing issues are also highlighted.

GenAI was perceived beneficial to tackle different aspects of learning-related tasks, especially explaining concepts from a neurodivergent sense (33%), improving their information processing speed (26%) and summarising information (20%). However, GenAI would not help to improve their memory nor their presentation skills. Students' perceived benefits are also statistically related to their perceived learning challenges ($p < 0.05$). Students also raised concerns such as

Table 1. Summary of main themes extracted from open-text questions

	Category	Challenges	Perceived benefits	Perceived risks	GenAI Tools	
					Helpful	Not helpful
Input	Search for information	Search for information; Summarising and organising ideas	Summarising	Misinformation, accuracy or bias; Need of own skills to effectively search	Summarising text; Explaining things	Verbose, vague, repetitive; Lack of source; Reliability not ensured
	Receiving information: Auditory or visual	Volume of reading; Keeping up in lectures; Clarity, simplicity or comprehension	Simplifying / clarifying from a neurodivergent perspective; Explaining through spoken language	Over explained or too vague	Recording and transcription; Breaking down text; Understanding communications	Accuracy of recording and transcription; Fact checking;
Internal	Working memory	Memory			Ineffectiveness	
	Processing information	Information processing speed	Improving information processing speed		Time saving & improving information processing speed	
Interface	Coping strategy	Mental health, burnout	Avoid asking for others' help			Prior knowledge to use; distraction
	Self-organisation	Organising time and time management			Providing starting point on writing, brainstorming	Time management, prior time investment to use
External	Learning environment	Physical surroundings (rooms, devices etc)				
	Other people in the environment	Group work, social aspects	Personalised help		Reduce pressure of keep asking questions	Replacing social contact and personal input
Output	Writing	Writing	Structuring; Spelling, punctuation and grammar	Overuse and over reliance; Repetitive buzz words	Plan and structure; Rewriting and proofreading	Problem shifts to determining or filtering what is important and correct to include
	Presentation	Presenting publicly				Lacks support for presentation
	Knowledge synthesis and understanding			Inauthentic learning; ownership of inputs; plagiarism		

prior investment to effective use (28%), risk of plagiarism (35%), misinformation (41%), and inauthentic learning (43%). One student with autism said that GenAI does not help with their "extremely methodical ways of doing things" unless they do deep prompt engineering. In addition, 46% of respondents became more concerned about plagiarism due to GenAI, and the students became more negative about GenAI's impact on their career aspirations in longer term. Students' views are divided on whether GenAI support creativity, with some (14.8%) found it beneficial while others thought "we risk cheapening the time and effort of other students and hampering their potential to succeed in creatively driven and opinion-based degrees". Despite concerns about GenAI, 46% students agreed that GenAI tools are well-designed to accommodate their special needs, with 43% neutral and 11% disagreed.

5 Discussions and Conclusion

Our study provides a comprehensive understanding of the views of neurodivergent and disabled students about the use of GenAI in higher education. Students found GenAI useful for study, but many lacked confidence and remained sceptical of its information. They were concerned about the risks of inauthentic learning and plagiarism, as well as the potential negative impact on their future careers. Students' perceptions of GenAI were closely linked to their learning difficulties and the challenges they have met during their studies. Our study confirmed the need to design teaching programs which are equitable and inclusive, considering students' various learning abilities and digital literacy. The analytical framework gives us a unified view over students' learning challenges regardless of their neurodivergent identities and these challenges can be similar to those faced by neurotypical students. More analytical works are needed for university teaching and support staff to build targeted tools and support programmes, addressing students' individual learning challenges.

Our analysis shows that GenAI supports neurodivergent and disabled students with their input and output tasks, such as summarising information, brainstorming, providing a starting point in writing, structuring, and proofreading, which improves their information processing speed and indirectly enhances self-organisation. However, GenAI has little impact on their physical environment, their internal self (their memory), nor the interaction between these two. It also does not help with presentation or communication skills. Since these GenAI tools generally do not improve how the physical environment functions for neurodivergent students, their support is limited to internal cognitive and organisational aspects rather than external environmental conditions. The use of GenAI can keep students isolated in their own world rather than asking others for help, in line with [14]. Alternatively, it may encourage students to rely on Gen AI even more if they find its answers more effective. As AI becomes more advanced, students may need to re-adapt their personal coping strategies to better live in an AI-driven world, potentially requiring support in this adaptation process. In this context, we may infer that a 'human touch' across various aspects of the university experience holds significant value.

Although insightful, this study has limitations. Its sample is limited to one university in the UK. Future plans include expanding the survey and further developing the analytical models to incorporate different types of information that consider various functions and sensory modalities. Another limitation is in the timeliness of the study, since the acceptance and awareness of GenAI tools may have changed since the data collection in first half of 2024. Thus, a longitudinal study examining the evolution of student use and experience with GenAI would be a valuable direction for future research.

References

1. Andrew Jenks, L.O., Guberman, D.: Emergent questions of access: disability and the integration of generative ai in teaching and learning. Higher Educ. Res. Develop., 1–14 (2024). https://doi.org/10.1080/07294360.2024.2442611
2. Barua, P.D., et al.: Artificial intelligence enabled personalised assistive tools to enhance education of children with neurodevelopmental disorders—a review. Inter. J. Environ. Res. Public Health **19**(3) (2022). https://doi.org/10.3390/ijerph19031192
3. Bhuvan Botchu, K.P.I., Botchu, R.: Can chatgpt empower people with dyslexia? disability and Rehabilitation: Assistive Technol. **19**(5), 2131–2132 (2024). https://doi.org/10.1080/17483107.2023.2256805
4. Chan, C.K.Y., Hu, W.: Students' voices on generative ai: perceptions, benefits, and challenges in higher education. Intern. J. Educ. Technol. Higher Educ. **20**(43) (2023)
5. Clouder, L., Karakus, M., Cinotti, A., Ferreyra, M.V., Fierros, G.A., Rojo, P.: Neurodiversity in higher education: a narrative synthesis. High. Educ. **80**(4), 757–778 (2020). https://doi.org/10.1007/s10734-020-00513-6
6. D'Souza, D., Avati, A.: Memory and Learning: Basic Concepts, pp. 227–240. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-72400-9_11
7. Fletcher, R., Bilson, U.: Disabled student's allowances: what isawarded and what is the impact? J. Inclusive Pract. in Further Higher Educ. (2019)
8. Folgieri, R., De Vecchi Galbiati, P., Dei Cas, L., Lucchiari, C.: A cognitive-driven BCI-based e-learning platform for learning disorders: A preliminary study. In: Project and Design Literacy as Cornerstones of Smart Education: Proceedings of the 4th International Conference on Smart Learning Ecosystems and Regional Development, pp. 235–246. Springer (2020). https://doi.org/10.1007/978-981-13-9652-6_21
9. Freeman, J.: Provide or punish? students' views on generative ai in higher education, https://www.hepi.ac.uk/2024/02/01/provide-or-punish-students-views-on-generative-ai-in-higher-education/, hEPI Policy Note 51
10. Glazko, K.S., Yamagami, M., Desai, A., Mack, K.A., Potluri, V., Xu, X., Mankoff, J.: An autoethnographic case study of generative artificial intelligence's utility for accessibility. In: Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2023. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3597638.3614548
11. Grant, D.: That's the Way I Think: Dyslexia, Dyspraxia and ADHD Explained. Routledge, 3 edn. https://doi.org/10.4324/9781315647005
12. Heather Johnston, Rebecca F. Wells, E.M.S.T.B., Parsons, B.N.: Student perspectives on the use of generative artificial intelligence technologies in higher education **20**(2)
13. HESA: Higher education student statistics: Uk, 2021/22 - student numbers. https://www.hesa.ac.uk/news/19-01-2023/sb265-higher-education-student-statistics/numbers
14. Joseph Crawford, Kelly-Ann Allen, B.P., Cowling, M.: When artificial intelligence substitutes humans in higher education: the cost of loneliness, student success, and retention. Stud. Higher Educ. **49**(5), 883–897 (2024). https://doi.org/10.1080/03075079.2024.2326956

15. K. Lea, R.B., Howe, R.: Report: exploring student perspectives on generative artificial intelligence tools at the university of northampton: a survey-based study
16. Lisa Jacobs, Adrian Parke, F.Z.C.H., Angeli, A.D.: Learning at school through to university: the educational experiences of students with dyslexia at one uk higher education institution. Disability Soc. **37**(4), 662–683 (2022). https://doi.org/10.1080/09687599.2020.1829553
17. McLeod, D.B.: Information-processing theories and mathematics learning: the role of affect. Int. J. Educ. Res. **14**(1), 13–29 (1990). https://doi.org/10.1016/0883-0355(90)90014-Y
18. Mwambe, O.O.: Deployment of information processing theory to support adaptive e-learning systems: feasibility study. Comput. Human Behav. Rep. **14**, 100420 (2024). https://doi.org/10.1016/j.chbr.2024.100420
19. Myles Hollander, D.A.W., Chicken, E.: Nonparametric Statistical Methods. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 2 edn. https://doi.org/10.1002/9781119196037
20. Rice, M.F., Dunn, S.: The use of artificial intelligence with students with identified disabilities: a systematic review with critique. Comput. Sch. **40**(4), 370–390 (2023). https://doi.org/10.1080/07380569.2023.2244935
21. Swanson, H.L.: Information processing theory and learning disabilities: an overview. J. Learn. Disabil. **20**(1), 3–7 (1987). https://doi.org/10.1177/002221948702000102
22. Wang, L.C., Li, X., Chung, K.K.H.: Relationships between test anxiety and metacognition in chinese young adults with and without specific learning disabilities vol. 71, pp. 103–126
23. Wang, S.H.: The application of information processing theory in higher education . In: 2021 2nd International Conference on Education, Knowledge and Information Management (ICEKIM).,pp. 930–933. IEEE Computer Society, Los Alamitos (Jan 2021). https://doi.org/10.1109/ICEKIM52309.2021.00209
24. Yassin, A., Mabanja, A.: Integrating information processing theory with artificial intelligence for enhanced learning outcomes. In: International Conference on Applied Social Sciences in Education, vol. 1(1), pp. 1–13 (Aug 2024). https://doi.org/10.31316/icasse.v1i1.6902
25. Zeng, W., Ju, S., Hord, C.: A literature review of academic interventions for college students with learning disabilities. Learn. Disabil. Q. **41**(3), 159–169 (2018). https://doi.org/10.1177/0731948718760999

Author Index

A
Abboud, Ralph 220
Aboukacem, Abdelghafour 364
Acunin, Adam 171
Adeniran, Adetunji 53
Agarwal, Snigdha 179
Akhuseyinoglu, Kamil 276
Aleven, Vincent 244
Alhathli, Manal 53
Almoubayyed, Husni 309, 356
Ambrose, Alex 171
Arevalillo-Herráez, Miguel 195, 462
Arnau, David 195, 462
Arnau-González, Pablo 195
Ayik, Bilgehan 268
Azevedo, Roger 325

B
Baeza, Cynthia 85
Baker, Ryan S. 356, 398
Baraniuk, Richard 203
Barany, Amanda 398
Barria-Pineda, Jordan 276
Belitz, Clara 356
Bergou, El Houcine 364
Berrada, Ismail 364
Bhushan, Shambhavi 220
Bian, Cunling 140
Bisconti, Piercosma 77
Blanchard, Nathaniel 446
Borchers, Conrad 220, 244
Bosch, Nigel 356
Botelho, Anthony F. 431
Boubaker, Anis 380
Brosnihan, Annamarie 325
Brusilovsky, Peter 276
Buddarapu, Venkata Nagaraju 372
Budzynska, Katarzyna 332

C
Cade, Whitney 301
Cai, Chang 27, 116
Cai, Minghao 260
Carvalho, Paulo F. 244
Castro, Mariana 85
Cesaroni, Valeria 77
Chartofylaka, Lamprini 380
Chawla, Nitesh 171
Chen, Elaine 478
Chen, Kaiquan 140
Chen, Si 171
Christie, S. Thomas 309
Chudziak, Jaroslaw A. 332
Cock, Jade Mai 60
Cohausz, Lea 108
Coker, Berk 228
Conati, Cristina 133
Costa, Newarney 406
Cutler, Ella 309

D
D'Mello, Sidney K. 301
Davidesco, Ido 19
Davis, Richard 60
de Jesus, Artur Cavalcante 454
Decker, Stefan 36
Delevaux, Valentine 60
Delgado, Tara 325
Demmans Epp, Carrie 260
Dempsey, Claire W. 301
Dieing, Thilo I. 108
Dijkstra, Rick 317
Ding, Jihong 236
do Nascimento, Diego Carvalho 454
Du, Sixuan 27
Duell, Jamie 116

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15881, pp. 487–490, 2025.
https://doi.org/10.1007/978-3-031-98462-4

E
El Mawas, Nour 380
Elhayany, Mohamed 3
Elouazizi, Noureddine 133
Endedijk, Maaike D. 156

F
Fan, Xiuyi 27, 116
Fan, Yizhou 293
Fancsali, Stephen E. 356
Faraji, Abdolali 187
Fathi, Fatemeh 36
Feng, Haiyue 27
Fletcher, Richard 478

G
Gagnon, David J. 398
Gajewska, Ewelina 332
Galletti, Martina 77
Ganaiem, Eman 439
Gašević, Dragan 293, 406
Gatz, Erin 220
Goel, Ashok K. 92
Gorgun, Guher 260
Graesser, Art 301
Guha, Alina 85
Guo, Liyuan 236
Guo, Shuchen 268
Gupta, Shivang 220

H
Haber, Nick 11
Haeri, Maryam A. 156
Halim, Noha Lea 100
Hays, Patrick 301
He, Yun 478
Hidalgo, Camila 100
Hinne, Max 317
Hong, Mengze 148
Hong, Shengxin 27
Houk, Alex 220
Huang, Xiaoshan 44
Huwer, Johannes 124

I
Inclezan, Daniela 390
Iraqi, Youssef 364
Ishioka, Tsunenori 470

J
Jawahar, C. V. 179
Jin, Lingxi 148

K
Kakarla, Sanjit 220
Karimi, Maral 325
Karumbaiah, Shamya 85
Käser, Tanja 60, 439
Kin Ho, Weng 116
Kismihók, Gábor 187
Kocabagli, Kemal Berk 228
Koedinger, Kenneth R. 220, 244
Komiya, Kanako 470
Krantz, Alexandre 414
Krishnaswamy, Nikhil 446
Kwong Lee, Bernett Teck 116
Kyriakoglou, Revekka 380

L
Lajoie, Susanne P. 44
Lallé, Sébastien 380
Lang, James 171
Latif, Ehsan 268
Le, Khiem 171
Lee, HaeJin 356
Lee, Hyosun 19
Lee, Ji-Eun 431
Lee, Seiyon M. 431
Lee, Zack 309
Lehman, Blair 301
Li, Fang 116
Li, Hongming 431
Li, Tengju 140
Li, Wengxi 11
Li, Zhaohui 285
Lin, Baicheng 148
Lin, Jiaju 285
Lins, Rafael 406
Liu, Lei 268
Liu, Naiming 203
Liu, Ninghao 268
Liu, Siyuan 27, 116
Liu, Suyi 19
Liu, Xiner 398
Liu, Xue 44
Liu, Yushuang 19
Lobczowski, Nikki G. 414

Loch, Frieder 124
Lu, Weigang 140
Lynch, Collin 372
Lyu, Qianru 244

M
Ma, Boxuan 236
Mahinpei, Romina 133
Marano, Cameron 325
Marques, Leonardo 454
Mason, Lisa 446
McLaren, Bruce 19
Medeiros Machado, Guilherme 380
Meinel, Christoph 3
Mekouar, Loubna 364
Mello, Rafael 406
Metcalf, Shari J. 398
Metoyer, Reid 171
Metoyer, Ronald 171
Minematsu, Tsubasa 348
Minghui Chen, Daisy 116
Minn, Sein 423
Mirhaji, Erfan 446
Moein, Mohammad 187
Molavi, Mohammadreza 187
Molenaar, Inge 317
Molnar, Izzy 171
Moreno, Matthew 325
Mou, Yongli 36

N
Nakagawa, Masaki 470
Nam, Sungjin 252
Namboodiri, Vinay 179
Nasiar, Nidhi 356
Nath, Debarshi 293
Ng, Olivia 116
Nguyen, Hung Tuan 470
Nkambou, Roger 423

O
Ocumpaugh, Jaclyn 356
Olney, Andrew M. 301
Opara, Chidimma 212
Orozco Vasquez, Ivan 133
Otten, David M. 156

P
Papadopoulos, Pantelis M. 156
Pasqua, Eleonora 77
Pea, Roy 11
Peng, Cindy 244
Perez, Matthew 228
Person, Natalie 301
Pimentel, Bruno Almeida 454
Prádanos, Luis I. 390
Proust-Androwkha, Sonia 380
Psyché, Valéry 380
Purandare, Mitra 124

R
Rafferty, Anna N. 309
Rajendran, Ramkumar 293
Richey, J. Elizabeth 19
Ring, Josh 317
Ritter, Steve 356
Roch, Neele 68
Rodrigues, Luiz 406
Rolim, Vitor 406
Roll, Ido 60, 439
Roman, Diego 85
Rothlin, Tobias 124

S
Saito, Takahiro 470
Segers, Eliane 317
Seifert, Tami 164
Sherman, Milla 325
Shimada, Atsushi 340, 348
Singh, Anjali 92
Slater, Stefan 398
So, Hyo-Jeong 148
Solera-Monforte, Sergi 195
Sonkar, Shashank 203
Sonmez Unal, Deniz 276
Sreedharan, Sarath 446
Stec, Hayden 19
Stinar, Frank 356
Subramonyam, Hariharan 11
Sudarshan, Vidya 116
Swanson, Luke 398

T
Taneja, Karan 92
Tavakoli, Mohammadreza 187
Teffera, Leah 19
Thillen, Ben 36
Thomas, Danielle R. 220
Thoms, Lars-Jochen 124
Tian, Yu 309
Toth, Adrienn 68
Tzamir-Levi, Taly 164

V
Van Camp, Alyssa 228
Vitale, Jessica 228

W
Walker, Erin 276
Wang, Yun 268
Watts, Field M. 268
Wawer, Michal 332
Wei, Zhanlan 398
White, Kiyo 85
Wiedbusch, Megan 325
Williams, Betsy 301
Wu, Haolun 44
Wu, Xuansheng 268
Wu, Yuyan 195, 462

X
Xavier, Cleon 406
Xiao, Feiwen 285
Xiong, Jinjun 285
Xue, Xiaorui 19

Y
Yang, Tianyuan 236
Yu, Jichao 268
Yu, Yongan 414

Z
Zarestky, Jill 446
Zhai, Xiaoming 268
Zhang, Jiayi 19, 398
Zhang, Kun 148
Zhang, Shan 431
Zheng, Qingxiao 285
Zhong, Zirui 431
Zhou, Shang-Ming 116
Zhu, Gaoxia 116
Zimmermann, Verena 68
Zion-Golumbic, Elana 19
Zou, Xiaohan 285
Zufferey, Noé 68