

Blockchain Technologies


Paul Moon Sub Choi
Seth H. Huang *Editors*

Finance and Large Language Models


 Springer

Blockchain Technologies

Series Editors

Dhananjay Singh , Department of Electronics Engineering, Hankuk University of Foreign Studies, Yongin-si, Korea (Republic of)

Jong-Hoon Kim, Kent State University, Kent, USA

Madhusudan Singh , Endicott College of International Studies, Woosong University, Daejeon, Korea (Republic of)

This book series provides details of blockchain implementation in technology and interdisciplinary fields such as Medical Science, Applied Mathematics, Environmental Science, Business Management, and Computer Science. It covers an in-depth knowledge of blockchain technology for advance and emerging future technologies. It focuses on the Magnitude: scope, scale & frequency, Risk: security, reliability trust, and accuracy, Time: latency & timelines, utilization and implementation details of blockchain technologies. While Bitcoin and cryptocurrency might have been the first widely known uses of blockchain technology, but today, it has far many applications. In fact, blockchain is revolutionizing almost every industry. Blockchain has emerged as a disruptive technology, which has not only laid the foundation for all crypto-currencies, but also provides beneficial solutions in other fields of technologies. The features of blockchain technology include decentralized and distributed secure ledgers, recording transactions across a peer-to-peer network, creating the potential to remove unintended errors by providing transparency as well as accountability. This could affect not only the finance technology (crypto-currencies) sector, but also other fields such as:

Crypto-economics Blockchain
Enterprise Blockchain
Blockchain Travel Industry
Embedded Privacy Blockchain
Blockchain Industry 4.0
Blockchain Smart Cities,
Blockchain Future technologies,
Blockchain Fake news Detection,
Blockchain Technology and It's Future Applications
Implications of Blockchain technology
Blockchain Privacy
Blockchain Mining and Use cases
Blockchain Network Applications
Blockchain Smart Contract
Blockchain Architecture
Blockchain Business Models
Blockchain Consensus
Bitcoin and Crypto currencies, and related fields

The initiatives in which the technology is used to distribute and trace the communication start point, provide and manage privacy, and create trustworthy environment, are just a few examples of the utility of blockchain technology, which also highlight the risks, such as privacy protection. Opinion on the utility of blockchain technology has a mixed conception. Some are enthusiastic; others believe that it is merely hyped. Blockchain has also entered the sphere of humanitarian and development aids e.g. supply chain management, digital identity, smart contracts and many more. This book series provides clear concepts and applications of Blockchain technology and invites experts from research centers, academia, industry and government to contribute to it.

If you are interested in contributing to this series, please contact madhusudan.singh@oit.edu
OR loyola.dsilva@springer.com

Paul Moon Sub Choi · Seth H. Huang
Editors

Finance and Large Language Models

Editors

Paul Moon Sub Choi
College of Business Administration
Ewha Womans University
Seoul, Korea (Republic of)

Seth H. Huang
Business School
The Hong Kong University of Science
and Technology
Clear Water Bay, Hong Kong

Ewha Womans University
Seoul, Korea (Republic of)

ISSN 2661-8338

ISSN 2661-8346 (electronic)

Blockchain Technologies

ISBN 978-981-96-5832-9

ISBN 978-981-96-5833-6 (eBook)

<https://doi.org/10.1007/978-981-96-5833-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 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 Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

If disposing of this product, please recycle the paper.

*Dear
Joung
Hwa and
Emma*

Preface

The integration of artificial intelligence (AI) agents and large language models (LLMs) is transforming the finance and trading sectors. These technologies enhance data analysis and decision-making by processing vast datasets with unparalleled speed and accuracy. AI agents identify patterns and predict market trends, while LLMs interpret unstructured data, providing deeper insights for trading strategies. This convergence improves trading efficiency and profitability, reshapes risk management and compliance, and personalizes financial services. As AI and LLMs evolve, they democratize access to advanced trading tools, benefiting individual traders and smaller institutions while driving innovation across the financial ecosystem.

This book delves into the foundational principles and recent advancements of LLMs and their integration into financial systems and managerial environments. It explores how these models enhance decision-making, improve predictive accuracy, and streamline operations. Each chapter focuses on a specific application of LLMs in finance, offering insights, methodologies, and case studies that illustrate their transformative potential.

LLMs are revolutionizing the financial industry by enhancing decision-making, predictive accuracy, and operational efficiency. Their capabilities include processing vast amounts of data, understanding complex financial concepts, and providing actionable insights. However, their integration also presents challenges, such as data privacy concerns, the need for significant computational resources, and ensuring model interpretability.

One notable application of LLMs is LLM-based time series analysis and regime detection, enhanced by Retrieval-Augmented Generation (RAG), contributing to adaptive trading strategies by enabling machines to better understand market contexts, conditions, and the implications of political and economic events. Fine-tuned, open-source LLMs can also enhance quantitative trading strategies by integrating numerical and textual data through techniques such as Low-Rank Adaptation (LoRA) and RAG. Another important application is in housing price appraisal, where models like ChatGPT demonstrate impressive reasoning capabilities and accuracy in real estate valuation. These advancements enable sophisticated and adaptive trading strategies, optimizing portfolio management.

LLMs also play a vital role in analyzing voluntary sustainability disclosures, assessing the impact of third-party assurance on corporate transparency, and examining the relationship between verbal masculinity in corporate communications and CEO compensation. Empirical evidence from India highlights factors influencing AI adoption in finance, while the intersection of federated learning and blockchain technology offers innovative solutions for collaborative AI model training. Finally, AI agents and deep learning algorithms are revolutionizing automated trading, driving the development of efficient market strategies.

This book is tailored for researchers, financial professionals, and enthusiasts eager to understand the transformative impact of LLMs on the financial industry and managerial decision-making. Through detailed explanations, practical examples, and forward-looking insights, readers will gain the knowledge and tools to harness the power of LLMs in their financial pursuits.

Ithaca, New York
Kowloon, Hong Kong

Paul Moon Sub Choi
Seth H. Huang

Contents

Large Language Models in Finance: An Overview	1
Paul Moon Sub Choi, Seth H. Huang, and Qishu Wang	
Housing Price Estimation and Reasoning Based on a Large Language Model	27
Seongeun Bae, Leehyun Jung, Sukyung Nam, Sihyun An, and Kwangwon Ahn	
Advancing Quantitative Trading Strategies Using Fine-Tuned Open-Source Large Language Models: A Hybrid Approach with Numerical and Textual Data Integration Using RAG and LoRA Techniques	43
Seth H. Huang, Jimin Kim, and Ka Lok Kellogg Wong	
Foundations of LLMs and Financial Applications	59
Yoonseo Chung, Jeonghyun Kim, MiYeon Kim, Minsuh Joo, and Hyunsoo Cho	
Voluntary Sustainability Disclosure and Third-Party Assurance: A Large Language Model Perspective	91
SoHyeon Kang and Sewon Kwon	
Verbal Femininity and CEOs Compensation	111
Sang-Joon Kim and Juil Lee	
Integrating LLM-Based Time Series and Regime Detection with RAG for Adaptive Trading Strategies and Portfolio Management	129
Chenkai Li, Chi Ho Roger Chan, Seth H. Huang, and Paul Moon Sub Choi	
Empirical Factor Identification for Artificial Intelligence in Finance: Indian Evidence	147
Rohit Kaushik	

Federated and Decentralized Finance: Decentralized Reward Mechanisms for Advanced AI Learning 157
Hyoseok Jang, Sangchul Lee, Haneol Cho, and Chansoo Kim

AI-Driven Financial Chart Analysis with Benchmarks: A Domain-Specific Large Language Model Approach 173
Hyoseok Jang, Sangchul Lee, Haneol Cho, and Chansoo Kim

Large Language Models in Finance: An Overview



Paul Moon Sub Choi , Seth H. Huang , and Qishu Wang 

Abstract The rapid advancement of large language models (LLMs) is revolutionizing industries, with finance emerging as one of the most promising beneficiaries. Financial LLMs (FinLLMs), developed on their foundations, can leverage advanced natural language processing techniques to process and generate insights from vast volumes of unstructured financial data. Particularly in specialized areas like quantitative investing, FinLLMs are poised to redefine the landscape by emulating the decision-making process of top traders. They can more efficiently capture market expectations, evaluate the impacts of market events, and aid investors across a wide range of investment practices. However, systematic research in the field of FinLLMs remains in its early stages. This paper provides a comprehensive overview of FinLLMs, aiming to encourage broader exploration of their mature applications in finance. The main content is summarized as follows: Firstly, we present a chronological overview tracing the evolution from general-domain pre-trained language models (PLMs) to specialized FinLLMs, highlighting critical advancements such as FinBERT, BloombergGPT, and FinMA. Secondly, we compare major FinLLMs by examining their training methods, datasets, and corresponding fine-tuning strategies. Thirdly, we summarize the characteristics and performance evaluations of seven benchmark financial NLP tasks. In addition, we explore the practical applications of FinLLMs in traditional finance and behavioral finance. Finally, we discuss the

P. M. S. Choi

Cornell SC Johnson College of Business, Cornell University, Ithaca, NY, USA

College of Business Administration, Ewha Womans University, Seodaemun-gu, Seoul, South Korea

P. M. S. Choi

e-mail: mc369@cornell.edu; paul.choi@ewha.ac.kr

S. H. Huang

Business School, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

e-mail: sethuang@ust.hk

Q. Wang (✉)

Business School, Seoul National University, Seoul, South Korea

e-mail: qishu_wang@snu.ac.kr

challenges and opportunities in the adoption of FinLLMs, including issues like data privacy and ethical considerations, while proposing directions for future research.

Keywords Large language models · Artificial intelligence · Fintech · Banking · Investing

1 Introduction

The rapid advancement of Artificial Intelligence (AI), particularly in Natural Language Processing (NLP), has catalyzed the development of LLMs, which excel at understanding and generating human language. Built upon transformative architectures like the Transformer [1], LLMs have revolutionized applications across various domains, including finance. In the financial sector, the growing complexity of data sources, including regulatory filings, earnings reports, and market news, has necessitated the adoption of sophisticated NLP solutions. This trend has spurred the application of LLMs in diverse financial tasks, such as predictive analytics, question answering, and risk modeling [2]. While general-purpose LLMs like GPT and BERT laid the groundwork [3, 4], domain-specific models, such as FinBERT [5] and BloombergGPT [6], have emerged to meet the unique demands of the financial sector, such as numerical reasoning and specialized financial vocabulary.

Recent innovations in FinLLMs have introduced various strategies for domain-specific specialization. These advancements include instruction fine-tuning with financial datasets such as FiQA-SA [7] and AnalystTone [8], as well as leveraging extensive corpora like Bloomberg’s FinPile [6]. These models address a wide array of financial applications, from regulatory compliance to ESG impact analysis and financial document summarization. Models like FinMA [9] and InvestLM [10] exemplify task-specific adaptability, setting new benchmarks for performance in specialized financial tasks.

Given the immense potential of LLMs in finance, this overview explores the evolution from general-domain LMs to FinLLMs. It reviews foundational developments, highlights their transformative applications in finance, and evaluates the performance of key models across critical benchmarks. By synthesizing insights from state-of-the-art financial LLMs, our work highlights emerging challenges, including data privacy [11] and ethical considerations, while exploring opportunities for future research and applications. The key contributions of this overview are outlined as follows:

- We examine the progression from general-purpose language models to specialized FinLLMs, focusing on methods and datasets that enhance domain-specific capabilities.
- We provide a comparative analysis of techniques and performance across prominent FinLLMs, highlighting innovations in instruction fine-tuning and data integration.

- The key financial tasks enabled by FinLLMs, such as risk assessment, sentiment analysis, and compliance optimization, are comprehensively summarized and evaluated in this work.
- We discuss opportunities and challenges in the field, particularly in relation to data privacy, model interpretability, and ethical considerations, offering insights into future research directions.

2 Evolution Survey of LLM for Finance Sector

2.1 *Foundational Developments in Language Models*

Recent advances in artificial intelligence, particularly in natural language processing, have led to the development of powerful LLMs like ChatGPT. These foundational models, constructed on increasingly sophisticated architectures, have significantly advanced NLP, enabling broader and more efficient language applications. The finance industry can benefit from the deployment of LLMs, as effective language understanding and generation can inform trading, risk modeling, customer service, and more.

However, this transformation was gradual rather than instantaneous. Initially, earlier iterations of recurrent neural networks (RNNs) and long short-term memory (LSTM) units were effective for various tasks but were limited, especially in handling long sequences of text due to issues such as vanishing gradients and computational inefficiency [12]. In response to these challenges, the introduction of the transformer architecture by Vaswani et al. [1] marked a significant breakthrough, facilitating the development and scaling of more advanced LLMs. The transformer's self-attention mechanism enables the simultaneous processing of entire sentences, enhancing scalability and operational efficiency. This innovation laid the groundwork for models like BERT [3] and GPT [4] to extend the range of NLP tasks they could handle, from translation to text generation and question answering.

Building on this foundation, BERT and GPT introduced novel methods for managing language tasks. BERT excelled in understanding text through its bidirectional training, proving highly effective for tasks requiring deep contextual understanding, such as sentiment analysis and named entity recognition [13, 14]. Conversely, GPT was designed for generating text, excelling in tasks such as text generation and dialog systems [15]. As these models evolved, their capabilities expanded, with GPT-3 (175 billion parameters) demonstrating few-shot learning, which enables it to perform new tasks with minimal examples [16]. Beyond scaling, new techniques like retrieval-augmented generation (RAG) further enhanced LLMs by integrating external knowledge sources, thereby improving performance on knowledge-intensive tasks [17].

2.2.2 Mixed-Domain PLMs for Broader Financial Applications

As financial tasks grew more complex, financial PLMs were required to incorporate knowledge from both general and financial domains. FLANG [19] is the first model to integrate general-domain and financial-domain data during its pre-training phase. Unlike FinBERT models, which are limited to financial corpora, FLANG utilizes a combination of general English language datasets and finance-specific datasets. This hybrid approach enables it to excel in both financial tasks and general NLP tasks.

Specifically, FLANG adopts the training methodology of ELECTRA [13], including a span boundary task. This task involves predicting masked tokens using a language model initially and then employing a discriminator to determine whether the tokens are original or substituted. The generator and discriminator are trained end to end, with masking using both words and financial terminology. The discriminator is then fine-tuned on specific tasks within the contributed benchmark suite, Financial Language Understanding Evaluation (FLUE) [19]. Shah et al. [19] conducted experiments that demonstrated FLANG's superior performance over FinBERT across all benchmarks, while maintaining the same number of parameters. This superiority was especially notable in tasks such as sentiment analysis, text classification, and question answering.

2.2.3 Transitioning to Large Financial LLMs

The further breakthrough in financial NLP came with the transition from financial PLMs to large-scale financial LLMs, exemplified by models like BloombergGPT [6]. Unlike the earlier FinBERT and FLANG models, which relied on relatively small corpora and task-specific datasets, BloombergGPT is trained on a vast general corpus (345 billion tokens) and a large financial corpus (363 billion tokens), including data from the web, news, filings, press releases, and Bloomberg's proprietary data. This extensive training enabled it to tackle more sophisticated financial tasks, such as deeper sentiment analysis, question answering, and even the generation of financial documents.

To further verify performance, Wu et al. [6] used GPT-NeoX, OPT, and BLOOM as control groups. Among all models evaluated, BloombergGPT demonstrated superior performance in four of the five tasks, specifically in ConvFinQA, FiQA-SA, FPB, and Headline tasks, and ranked second in named entity recognition (NER).

2.2.4 Instruction Fine-Tuning for Task-Specific Adaptability

The continuous refinement of LLMs for specialized financial tasks has been made possible through the introduction of instruction fine-tuning. This process involves additional training with explicit textual instructions aimed at enhancing the capabilities and controllability of LLMs. Models such as FinMA [9] and InvestLM [10] have been at the forefront of transforming existing financial datasets into instructional

datasets, which are then utilized to fine-tune LLMs. By leveraging task-specific instruction datasets, these models are better equipped to adapt to evolving financial scenarios.

FinMA optimizes the LLaMA model by fine-tuning with 136,000 instructional data samples. It includes an evaluation framework with five tasks and nine datasets, allowing it to accurately follow instructions for various financial tasks. In comparative tests on financial NLP tasks, FinMA consistently outperforms other LLMs, such as BLOOM, ChatGPT, and BloombergGPT, particularly in the FPB, FiQA-SA, and Headline datasets. This underscores the effectiveness of domain-specific instruction tuning in boosting LLM performance within specific sectors [9].

InvestLM is based on the LLaMA-65B model, leveraging a specialized financial-domain instruction dataset [20]. It employs the low-rank adaptation (LoRa) method [21] to enhance parameter tuning to enable more efficient training. Additionally, it uses linear rope scaling at a scale of 4, allowing a context length of 8,192 tokens. This expansion helps InvestLM better manage extensive financial texts, such as SEC filings, corporate disclosures, and analyst reports, optimizing it for regulatory compliance and financial document summarization.

2.3 *Horizontal Comparison of Financial-Domain LLMs*

Building on the evolutionary analysis in Sect. 2, this section offers a horizontal comparison, examining how updates in datasets and methods across FinLLMs have expanded their practical applications (Table 1). FinBERT-19, which primarily relied on the Financial PhraseBank (FPB) and FiQA-SA datasets, was initially designed for sentiment analysis [18]. This gave it a narrow scope, as it mainly focused on interpreting the sentiment of financial texts. However, with the introduction of FinBERT-20, the model expanded its capabilities by incorporating the AnalystTone dataset [8]. This dataset allowed for more detailed sentiment analysis, particularly by extracting sentiment embedded within analyst reports, which enhanced the model’s utility in tracking real-time market sentiment. FinBERT-21 further advanced these capabilities by integrating FiQA-QA and FinSBD19 datasets, enabling it to excel in question answering tasks and business document summarization. These improvements, supported by a larger corpus of 12 billion financial tokens, allowed FinBERT-21 to effectively handle more complex financial data [2, 5].

FLANG leverages datasets such as FiQA-QA and Headline to specialize in real-time financial event extraction, text classification, and named entity recognition. These datasets have expanded FLANG’s capabilities beyond mere static sentiment analysis, enabling it to adeptly process dynamic and time-sensitive financial information, such as breaking news or market developments [19]. BloombergGPT is based on the comprehensive dataset “FinPile,” which encompasses a variety of English financial documents, including news, filings, press releases, web-scraped financial content, and social media posts drawn from Bloomberg’s archives. This significantly enhances its responsiveness to the foundational tasks [6].

Table 1 Summary of FinPLMs and FinLLMs

Model family	Model name	Core framework	Parameters	Techniques	Data size	Datasets utilized	Specializations
Financial PLM	FinBERT-19 [18]	BERT	0.11B	Post-Pre-training, fine-tuning	General: 3.3B words, Financial: 29 M	FPB, FiQA-SA	Sentiment Analysis (SA)
	FinBERT-20 [8]	BERT	0.11B	Pre-training, fine-tuning	Financial: 4.9B tokens	FPB, FiQA-SA, AnalystTone	SA
	FinBERT-21 [5]	BERT	0.11B	Pre-training, fine-tuning	Financial: 12B words	FPB, FiQA-SA, FiQA-QA, FinSBD19	SA, Question Answering (QA), Structure Boundary Detection (SBD)
	FLANG [19]	ELECTRA	0.11B	Pre-training, fine-tuning	Financial: 696 k documents	FPB, FiQA-SA, Headline, FIN, FiQA-QA, FinSBD21	SA, QA, SBD, Named Entity Recognition (NER), Text Classification (TC)
Financial LLM	BloombergGPT [6]	BLOOM	50B	Pre-training, prompt engineering	General: 345B tokens, Financial: 363B tokens	FPB, FiQA-SA, Headline, FIN, ConvFinQA	SA, TC, NBR, QA
	FinMA [9]	LLaMA	7B, 30B	Instruction Fine-tuning, prompt engineering	General: 1 T tokens	FPB, FiQA-SA, Headline, FIN, FinQA, ConvFinQA, StockNet, CIKM18, BigData22	SA, TC, NBR, QA, Stock Market Prediction (SMP)

(continued)

Table 1 (continued)

Model family	Model name	Core framework	Parameters	Techniques	Data size	Datasets utilized	Specializations
	InvestLM [60]	LLaMA	65B	Instruction Fine-tuning, prompt engineering	General: 1.4 T tokens	FPB, FiQA-SA, FOMC, FinQA, ECTSum	SA, TC, QA, Summarization
	FinGPT [75]	Multi LLMs	7B	Instruction Fine-tuning, prompt engineering	General and Financial: 2 T tokens	FPB, FiQA-SA, Headline, FIN, FinRED	SA, TC, NBR, Relation Extraction

FinMA extends further to include datasets such as StockNet, CIKM18, and BigData22. This expansion enables it to perform more complex stock movement prediction tasks. Unlike financial NLP tasks, financial prediction tasks align more closely with real-world scenarios and present greater challenges.

InvestLM leverages an instructional dataset covering a broad spectrum of financially related topics, from Chartered Financial Analyst (CFA) exam questions to SEC filings and stock exchange discussions on quantitative finance. Consequently, InvestLM demonstrates enhanced capabilities in understanding financial texts and offers more insightful responses to investment-related queries than previous models [10]. And the notable advancement for FinGPT lies in its use of the FinRED dataset [22], a dataset specifically designed for relation extraction in the financial domain. This allows FinGPT to further broaden the specific scope of tasks based on previous models, providing a foundation for extracting and understanding financial relations within textual data.

2.4 Traditional Versus AI-Driven Finance

In summary, the transition from traditional financial models to AI-driven LLMs marks a significant transformation in financial data processing and analysis. Traditional models, while effective for structured and historical data, struggle with the vast amounts of unstructured data generated today. These models are limited by static data sources and rigid rules, which hinder their ability to process real-time information and nuanced textual data. In contrast, AI-driven models excel at integrating real-time market data, news feeds, social media sentiment, and regulatory documentation, providing critical insights for modern financial decision-making. AI models are particularly effective at processing large-scale, real-time data across multiple modalities, including text, numbers, and even multimedia. This gives them an advantage over traditional models, which struggle with unstructured data. Traditional models, often grounded in statistical methods, cannot match the speed and adaptability of AI models when responding to rapidly shifting market conditions.

Additionally, unlike rule-based models that require manual updates, models like InvestLM and FinMA can quickly adapt to new financial environments through fine-tuning and prompt engineering. This adaptability is particularly valuable for regulatory compliance, where AI models can analyze complex legal texts faster than traditional systems, ensuring institutions stay up to date with evolving regulations [9, 10]. A notable example is its application in areas like anti-money laundering (AML) practices [23]. In these contexts, the ability of LLMs to rapidly process and interpret complex regulatory documents enables financial institutions to identify suspicious activities and ensure compliance with evolving legal standards. However, this widespread use of LLMs in compliance also brings concerns regarding data privacy and security. For the sake of information security, many institutions prefer to process sensitive internal documents using in-house LLMs rather than public

models. For example, the use of public LLMs—whether free or subscription based—is often restricted to non-sensitive tasks like translation or grammatical verification of publicly available information. Even with non-confidential, internal information, there is a risk that using public models could allow the inference of sensitive financial insights. As such, companies prioritize internal LLMs for handling data that, while not strictly confidential, could indirectly reveal key financial details. These considerations highlight that while AI-driven finance offers significant advantages, it also brings various risks that must be carefully managed.

Lastly, AI-driven models also excel in predictive analysis, such as forecasting stock market trends and merger outcomes, integrating qualitative data like sentiment and policy interpretations for more comprehensive insights than traditional models can offer.

3 Cognitive Processes of LLMs: How Are They Transforming Financial Services?

3.1 *Financial Data Extraction*

Financial data extraction is the critical foundational stage in integrating LLMs into financial services. The abundance of financial texts provides valuable resources for analysts and investors, but how to efficiently extract useful information inside is even more important. As it is estimated that 80–90% of financial data is unstructured [24], manual extraction and filtering could be infeasible and unsustainable. NLP technology offers the pathway to automate this process.

Named Entity Recognition (NER) serves as a foundational task in this context, extracting critical structured information such as company names, transaction dates, stock prices, currencies, and events from unstructured financial text. This structured data enables downstream financial NLP tasks to operate more effectively. For instance, in question answering, NER can extract target entities from a query, allowing the model to pinpoint accurate answers. As in response to the question “What was Tesla’s revenue last year?” NER identifies “Tesla” and “last year” as key entities, driving the system to efficiently locate the relevant information.

It is worth noting that the original NER task is the extraction of information such as locations, organizations, and persons for general domain. But its application in finance requires specialized knowledge. This expertise is essential for extracting relevant data from complex documents like regulatory filings, news articles, and financial reports. For instance, publicly traded companies must file reports in extensible Business Reporting Language (XBRL) format. Manually annotating these documents is both labor intensive and expensive. In response, Loukas et al. [25] have adapted XBRL tagging for financial entity extraction and developed the FiNER-139 dataset, containing over 1.1 million sentences labeled with gold-standard XBRL tags. This

dataset has an extensive array of 139 different entity types, which has significantly enhanced the performance of BERT-based models in this domain.

Additionally, the integration of NER with relation extraction in the KPI-BERT [26] model facilitates the analysis of key performance indicators (KPIs) within financial reports. This integrated approach helps in identifying and contextualizing critical metrics within dense datasets. Their final experimental results also show significantly improved predictive performance on a practical dataset of German financial reports, surpassing multiple robust baselines, including a competing state-of-the-art span-based entity tagging approach.

However, once key information is recognized, managing this information remains challenging. Financial participants often seek to locate relevant portions of vast datasets through clear categorization. **Financial Text Classification (FTC)** is an NLP task designed to achieve this goal, enabling models to classify financial documents into predefined categories. Common applications include sentiment analysis of financial news and the classification of reports based on their relevance to specific market sectors. For instance, the FOMC dataset [27] categorizes FOMC documents as Dovish, Hawkish, or Neutral, reflecting the sentiment conveyed in the materials.

Moreover, multiple dimensions of information, such as price direction, interest rate trends, and even customer service queries in banking, are often classified. FLUE dataset includes the Gold News Headline collection [28], featuring 11,412 news headlines categorized into binary classes across nine different labels such as “price up” and “price down,” aimed at text classification tasks. Meanwhile, the FedNLP dataset [29] is derived from diverse FOMC documents. It is annotated with labels like Up, Maintain, or Down, reflecting decisions on the Federal Reserve’s Federal Funds Rate for upcoming periods. The Banking77 dataset [30] contains 13,083 entries pertaining to 77 distinct banking-related customer service intents, including situations like “card loss” or “linking to an existing card,” and is specifically designed for intent recognition and developing conversational systems.

Additionally, the FTC task underscores the considerable potential for progress in handling incomplete datasets. Zhao et al. [31] conducted a study focusing on how text classification models manage incomplete data in financial news classification. They developed a character-level vocabulary from the financial dataset and mapped text segments to high-dimensional spatial vectors. These vectors undergo processing with spatial convolution to pinpoint local features and are integrated with gated recurrent units that capture temporal dynamics. The classification is then executed using a SoftMax function that leverages these spatial and temporal features, managing to maintain a certain level of accuracy despite the challenges posed by fragmented real-time data.

After achieving functions like targeting broad objectives and general classifications, research has increasingly focused on directly extracting financial summaries to further expedite information processing. **Financial Text Summarization** task thus emerged, creating summaries to help decision-makers quickly access key insights. InvestLM includes summarization tasks for the first time, conducting experiments on the ECTSum dataset [32]. ECTSUM is a dataset with transcripts of earnings

calls (ECTs), hosted by publicly traded companies, as documents, and short expert-written telegram-style bullet-point summaries derived from corresponding Reuters articles. Their work demonstrated how models trained on this dataset could produce bullet-point summaries that effectively capture key aspects of discussions on financial performance and strategic plans, thus facilitating more informed decision-making by analysts.

Further, by combining ECTSUM with other text summarization datasets in the financial domain for the instruction-tuning step, Lee et al. [33] developed a method to convert the basic model, Llama3 8B, into a FinLLM, a specialized tool for summarizing financial texts. This transformation allows the model to become an “expert” in distilling key information from complex financial documents. They explored how LLMs could be fine-tuned to generate coherent summaries from regulatory filings and market analysis reports. Utilizing datasets like ECTSUM, summaries of earnings calls, and financial news, these models are trained to pinpoint essential details in intricate documents. This capability is especially valuable in scenarios where stakeholders need to review substantial information within a limited timeframe, such as during quarterly earnings announcements or in response to regulatory changes.

Beyond the general summarization and extraction of key information, identifying and classifying relationships between entities mentioned in text is also crucial in the financial domain. **Financial Relation Extraction (FRE)** has emerged as a widely used NLP task addressing this need. Financial news is often filled with noise, and not every sentence is useful for decision-making. FRE directly extracts critical relationships, such as identifying which entities (companies, individuals, events) are connected and how they interact, something that general information extraction often cannot clarify. For example, in the news headline “Microsoft plans to acquire Activision Blizzard for \$69 billion to strengthen its position in the gaming market,” FRE task can extract the structured relationship as: [Microsoft]—[acquires for \$69 billion]—[Activision Blizzard]. This structured data captures the transaction amount, target company, and strategic purpose, enabling analysts to rapidly evaluate and make informed decisions.

Beyond mergers, acquisitions, partnerships, and organizational structures, FRE is also crucial for regulatory compliance. For example, most previous AML systems relied on manually crafted rules and structured databases, which were ineffective at identifying complex and hidden money laundering activities, especially those with dynamic or time-varying characteristics, resulting in high false-positive rates. By using FRE and NER together, systems can more effectively build the relationship network of a target entity [34], analyzing complex hidden connections with suspicious entities, such as circular fund transfers or the use of shell companies—common indicators of money laundering activities.

To achieve optimal training outcomes, large-scale FRE datasets are also essential. Kaur et al. [35] utilized the REFinD dataset, currently the most extensive collection of relationships within financial documents, containing around 29,000 cases across 22 types of relationships among 8 categories of entity pairs. This detailed, structured dataset is crucial for capturing typical relationship types found in corporate filings and financial statements. It significantly improves FinLLMs’ capabilities to

discern complex interrelations, for example, linking specific financial indicators to overall financial health, thereby enhancing the precision of automated relationship extraction methods. Another significant dataset, FinRED [22], which is compiled from financial news and earnings call transcripts, encompasses a broader spectrum of financial relationships, including ownership ties, investment links, and contractual commitments. This dataset can theoretically provide a comprehensive benchmark for training models to detect nuanced relationships in financial documents.

3.2 *LLM-Based Financial Decision-Making*

Following data extraction, one of the most transformative advantages of FinLLMs over traditional deep learning methods lies in their reasoning capabilities. Unlike earlier models that primarily focus on data fitting, FinLLMs emulate human-like reasoning, enabling more effective interpretation of financial information. This is particularly evident in tasks such as sentiment analysis and market expectations, where the complexity and volume of related data often surpass human analytical capacity.

Market Sentiment Analysis is a vital NLP task here for decision-making, especially in stock investment. Despite the apparent efficiency of markets, stock prices often fluctuate independently of the discounted value of assets, or the technologies held by companies. Market sentiment, heavily influenced by social media, news, and trends, has emerged as a key driver of these fluctuations. However, discerning the hidden patterns inside is highly challenging for most individual investors. Effectively identifying sentiments in related financial texts is therefore essential for making quick and informed decisions.

For example, Bloomberg reported that portfolios based on trading sentiment significantly outperformed benchmark indices [36]. Previous studies in financial economics also indicate that news articles and social media sentiment can help predict market returns and company performance [37, 38]. The unique nature of sentiment, however, makes accurate analysis or application highly domain dependent. Araci [18] was the first to effectively implement this idea with FinBERT, which is specifically fine-tuned for financial text analysis. FinBERT was trained on a large corpus of financial documents, including earnings calls and market reports, allowing it to detect nuanced sentiments—positive, negative, or neutral—with high accuracy.

Based on FinBERT-19, FinBERT-20 and FinBERT-21 utilized datasets such as AnalystTone, FiQA-SA, and FinSBD19 to further refine sentiment analysis [5, 8, 18]. The AnalystTone dataset consists of labeled sentiment data from analyst reports, offering insights into the sentiment expressed by financial analysts regarding company performance and market conditions. The FiQA-SA dataset [7] focuses on extracting opinions and sentiment from financial question-answer pairs, providing a valuable resource for training models to understand investor and market participant reactions to various financial events and news. Additionally, the FinSBD19 dataset,

which contains labeled financial news and social media data, enables the model to quickly adapt to new trends and news that could influence market sentiment.

Building on a similar foundation as market sentiment analysis, the core goal of **Question Answering (QA)** task in the financial domain is to provide users of the extracted financial information with a faster pathway. For instance, addressing questions like “How did the company perform last year?” from financial statements requires capturing specific profitability or capital indicators as references. However, when users are uncertain about specific indicators, an intuitive Q&A output format aligns more closely with their general needs.

Maia et al. [7] first demonstrated this potential through the Financial Opinion Mining and Question Answering (FiQA) challenge, where QA models successfully extracted both factual data and nuanced opinions from financial texts, providing a practical foundation for analysts to gain insights into market dynamics and company-specific sentiment. Compared to QA tasks in general domains, the financial sector requires more complex numerical reasoning and the ability to interpret diverse data formats. To bridge this gap, Chen et al. [39] developed FinQA, a large-scale dataset of question–answer pairs from financial reports authored by financial experts. These answer pairs are primarily designed to address complex questions within financial data, theoretically offering more effective means to automate the analysis of large financial document corpora. ConvFinQA further advances this area by enabling conversational interactions, allowing a model to engage in multi-turn dialogs with users to extract information across rounds of questioning, thus enhancing contextual understanding of financial documents [40]. These datasets are critical for training models, as they provide structured resources that enable FinLLMs to adeptly extract and interpret detailed financial data, significantly reducing the time and effort required for manual analysis.

Finally, as an NLP task oriented toward future trends, **Stock Market Prediction** focuses on directly forecasting stock prices or trends. Compared to basic tasks like information extraction or sentiment analysis, stock market prediction further integrates the extracted information, leveraging time-series modeling and causal relationship analysis to deliver clear quantitative or qualitative predictions for financial decision-making recommendations.

For example, Wu et al. [41] proposed a novel Cross-modal attention-based Hybrid Recurrent Neural Network (CH-RNN), inspired by the developed DA-RNN model [42]. CH-RNN consists of two essential modules: one uses DA-RNN to capture stock trend representations across different stocks, and the other employs recurrent neural networks to model daily aggregated social texts. This integration allows the model to incorporate both textual sentiment and historical trends into its predictions.

To train such models more effectively, datasets that combine sentiment data with stock price movements are essential. The StockEmotions dataset [43], which detects emotions in the stock market, includes 10,000 English comments from StockTwits, a financial social media platform. Inspired by behavioral finance, this dataset proposes 12 fine-grained emotion classes that capture the spectrum of investor emotions with more detailed features, such as investor sentiment classes, nuanced emotions, emojis, and time-series data. This granularity allows models to respond more sensitively to

sudden shifts in investor sentiment, offering predictions that better reflect real-time market dynamics.

4 Practical Performance of FinLLMs: How They Operate in Financial Market?

The practical application and performance of FinLLMs in current financial markets merit further exploration, particularly regarding the improvement and updates of baseline models, which will influence the future development of FinLLMs. This section will delve into the practical performance of these models across various tasks, including market trading, portfolio management, regulatory compliance, and behavioral finance applications.

4.1 *FinLLMs for Trading Enhancements*

FinLLM trading agents focus on leveraging LLMs to analyze vast amounts of external data, such as news, financial reports, and stock prices, extracting insights to generate buy or sell signals. However, the practical directions and effectiveness of these trading agents vary depending on the characteristics and analytical processes of the models.

Firstly, news-driven architectures are the most fundamental types, integrating individual stock news and macroeconomic updates into the prompt context, then directing the LLM to predict stock price movements in the next trading period [44–47]. For instance, Lopez-Lira and Tang [44] demonstrated that ChatGPT can significantly predict out-of-sample daily stock returns, with stronger predictability observed for smaller stocks and those following negative news. Subsequently, studies such as Zhang et al. [48], Kirtac and Germano [49], and Delgadillo et al. [50] fine-tuned LLMs with financial datasets, showing enhanced performance by aligning the models with domain-specific knowledge.

More advanced architectures involve summarizing and refining news data as well as reasoning about the relationships between news and stock price movements. This process often requires effective utilization of financial text summarization modules. In some studies, these summaries are managed through a memorization module. During trading, relevant “memory” is retrieved as a “recommendation” context to inform the final trading decision [51]. Specifically, Fatouros et al. [51] developed MarketSenseAI, a framework that leverages GPT-4’s advanced reasoning capabilities for stock selection in financial markets by integrating Chain of Thought and In-Context Learning. MarketSenseAI analyzes diverse data sources and, through empirical testing on the competitive S&P 100 stocks over a 15-month period, achieved an excess return ranging from 10% to 30% and a cumulative return of up to 72% over the sample period.

Building on this foundation, subsequent research introduced trading agents, such as those incorporating extracted memory, further enhanced with reflection. Reflection, as described by Park et al. [52], is built on extracted memory using LLM-based summarization. This architecture extends an LLM to store a comprehensive record of the agent’s experiences, synthesize these memories into higher level reflections over time, and dynamically retrieve them for planning behavior [52]. In the context of financial market trading, this could represent higher level knowledge and insights progressively aggregated from raw memories and observations.

FinMem [53] and FinAgent [48] are representative models in this domain. FinMem introduces a trading agent with layered memorization and characteristic profiling. Raw inputs, such as daily news and financial reports, are summarized into structured memories. Its architecture consists of three core modules: profiling, to customize the agent’s traits; memory, which processes layered messages by retrieving relevant memories and integrating them with new observations to produce reflections stored in a layered memory bucket; and decision-making, where these memories and reflections are retrieved to generate final trading decisions. Using multi-source financial data from platforms like Yahoo Finance and Alpaca News API, FinMem’s performance was benchmarked against various algorithmic agents, including FINGPT and the general-purpose generative agent from Park et al. [52]. Experimental results validated its superior trading performance and can significantly boost cumulative investment returns.

FinAgent [48], building on a similar layered memory and reflection design, further incorporates a multimodal module that processes numeric, text, and image data to capture market dynamics and historical trading patterns. Additionally, it introduces a diversified memory retrieval system for market intelligence and reflection modules, effectively separating trading and retrieval tasks to enhance functionality and reduce noise. Comprehensive experiments on six financial datasets, including stocks and cryptocurrencies, demonstrated FinAgent’s superior performance across six financial metrics, outperforming nine state-of-the-art baseline models, including FinMem, with an average profit increase of over 36%.

4.2 LLM-Driven Portfolio Management Innovations

When managing portfolios comprising multiple stocks, FinLLM requires relatively more precise and rigorous strategies [48, 49, 54]. A common approach is to use ranking-based strategies. These strategies assign numerical scores to rank stocks and allocate funds based on the scores. Konstantinidis et al. [54] employed a fine-tuned Llama 2 7B model, named FinLlama, to perform sentiment analysis on all stocks in the S&P 500 index. Companies were ranked daily according to their sentiment signals, omitting those without sentiment data on a given day. With daily sentiment scores ranging from -1 to 1, the top 35% were assigned long positions, while the bottom 35% were allocated short positions. Experimental results demonstrated that

FinLlama outperformed the leading method in the field, FinBERT, achieving cumulative returns that exceeded FinBERT by 44.7%. Moreover, FinLlama delivered a significantly higher Sharpe ratio of 2.4 and exhibited lower annualized volatility, indicating its potential to provide investors with valuable guidance while simultaneously reducing portfolio risk.

Kirtac and Germano [49] also explored sentiment analysis using models such as GPT-3-based OPT, BERT, and FinBERT on 965,375 U.S. financial news articles from 2010 to 2023. Their results showed that the GPT-3-based OPT model significantly outperformed the others, achieving a prediction accuracy of 74.4% for stock market returns. Employing an OPT-based long-short strategy, and accounting for transaction costs of 10 basis points (bps), the model achieved a Sharpe ratio of 3.05. Between August 2021 and July 2023, this strategy generated a return of 355%, surpassing other strategies and traditional market portfolios.

FinLLMs can be further applied to ESG and impact investing in portfolio management. With the growing importance of environmental concerns and the rise of sustainable development, public attention to corporate non-financial information has intensified, prompting investors to integrate ESG ratings into their portfolio decisions [55–57]. Since most ESG information is presented in textual formats such as reports, disclosures, press releases, and 10-Q filings, FinLLMs can efficiently integrate ESG factors into portfolio management strategies through advanced frameworks and methodologies. Specifically, machine learning algorithms can classify companies based on ESG performance, identify ESG leaders and laggards, and construct diversified portfolios prioritizing sustainability.

The most updated models in this domain include ESGBERT [58] and GPT4ESG [59]. ESGBERT builds on a domain-specific adaptation of BERT by fine-tuning pre-trained BERT weights using a Masked Language Model (MLM) task on an ESG corpus, followed by further fine-tuning for Sequence Classification to predict two types of intelligence: (1) whether there is a change in environmental scores and (2) whether the change in environmental scores is positive or negative based on ESG-related text in 10-Q filings [58]. Experimental results demonstrated that ESGBERT outperforms the original BERT and baseline models in environment-specific classification tasks, confirming the potential of LLMs for further research and application in ESG investing.

Building on this foundation, Lin et al. [59] developed GPT4ESG, a system architecture combining BERT and GPT to rapidly analyze ESG investments and impacts of companies. They curated a dataset comprising ESG reports of prominent publicly listed U.S. technology companies, which was cleaned and preprocessed for compatibility with the model. A customized GPT assistant was then used to provide scoring, with results validated by experts. Additionally, a classification layer was integrated into the model's output, and fine-tuning was conducted with sector-specific ESG expertise. Experimental results revealed that the customized GPT4ESG model outperformed ESGBERT in ESG data classification, further simplifying ESG reporting and enabling stakeholders to make responsible portfolio decisions.

4.3 *LLM and Banking Industry*

In an increasingly digital world, customer experience has become a key differentiator for banks and financial institutions, alongside interest rates and fees [60]. Providing excellent customer support is a critical component of enhancing the overall experience, and advancements in technology are particularly vital in this area. The emergence of LLMs has driven significant transformations in the industry. A prominent application is conversational banking, offering 24/7 support and timely solutions for customer inquiries [61], while also enabling automation of back-office operations. To some extent, LLMs may even replace human financial advisors. Although they currently cannot provide tax advice, the prospect of AI fundamentally reshaping the role of traditional financial advisors in the future is not far-fetched.

For instance, in March 2023, Morgan Stanley Wealth Management (MSWM) collaborated with OpenAI to develop customized solutions using GPT-4 [62]. This tool allows financial advisors to quickly access Morgan Stanley's knowledge base and generate research reports or investment insights, enhancing client services. By June 2023, JPMorgan Chase began developing an AI investment advisor named IndexGPT, designed to analyze and select securities tailored to client needs. Similar to OpenAI's ChatGPT, IndexGPT is specialized for financial services and aims to provide personalized investment advice. Other firms, including BlackRock, Vanguard, and Fidelity, are also exploring similar AI-driven advisory products, underscoring the growing application of such tools in banking and financial services [63].

Additionally, LLMs can assist with credit scoring and loan processing. Banks often struggle with information asymmetry, lacking enough data to evaluate borrowers' creditworthiness. Traditionally, this process is manual, slow, and prone to errors. With LLMs like ChatGPT, financial institutions can automate the initial stages of loan processing, extracting and analyzing data from loan applications. However, despite generally high accuracy, variations in application formats or data representations may require human validation, especially for handling outlier cases. Similarly, for credit scoring, LLMs can efficiently combine structured and unstructured data to analyze borrowers' credit histories and quickly compute credit scores, accelerating the assessment process. Sanz-Guerrero and Arroyo [64], for example, fine-tuned BERT on loan descriptions from the Lending Club dataset to distinguish between default and non-default loans. Their findings demonstrate that integrating BERT-generated risk scores into traditional credit-granting models can significantly improve default predictive performance.

Lastly, fraud, particularly credit card fraud, remains a critical concern for the banking sector. With advancements in artificial intelligence, the scale and complexity of fraud may increase further. In response, AI shows potential for faster and more efficient detection and analysis of online fraud [65–67]. Korkanti [66] utilized the UCI Credit Card Fraud Detection Dataset, which contains transactions made by European cardholders in September 2013. The public dataset covers 2 days of transactions, with 492 fraud cases out of 284,807 transactions. They initially employed isolation forests and autoencoders for anomaly detection, followed by logistic regression and

Gradient Boosting Machines (GBMs) to validate transactions, achieving promising results in fraud detection. Building on this foundation, fine-tuned GPT-3 models can extract subtle features often overlooked by traditional models, particularly from unstructured data in transaction narratives and customer communications.

The role of LLMs in financial markets resembles a double-edged sword, simultaneously managing risks while posing challenges. The banking industry inherently involves significant risks, including data security, privacy concerns, and regulatory compliance. Handling sensitive financial and personal information under strict data governance requirements complicates the development and deployment of such models. Current research focuses primarily on ChatGPT's practical applications in this context, with fewer fine-tuned models tailored specifically for banking. Moreover, the fragmentation and proprietary nature of banking datasets limits the accessibility and scalability of fine-tuning efforts for domain-specific LLMs, hindering broader adoption and development. For these reasons, major banks such as JPMorgan Chase, Bank of America, Citigroup, and Goldman Sachs have restricted employees from using ChatGPT [68]. Further research and development are needed to explore compliant and secure implementations of LLMs in the banking sector.

4.4 Behavioral Analysis with FinLLMs

In previous sections, we discussed how FinLLMs can process unstructured texts from various sources, such as social media, news, and investor communications, to extract insights into market sentiment, investor preferences, and the dynamics of collective behavior. These capabilities significantly contribute to behavioral finance research, which, unlike the traditional rational market behavior hypothesis, examines how psychological factors and biases influence financial decision-making. FinLLMs excel at sentiment extraction by analyzing nuanced textual data and capturing subtle emotional undertones, such as optimism, fear, or uncertainty. Unlike traditional methods that rely on simple word counts or sentiment lexicons, FinLLMs embed these sentiments within a broader linguistic framework, enabling deeper and more precise analysis [44–50].

Moreover, we posit that beyond sentiment analysis, FinLLMs hold significant potential for identifying investor risk preferences and behavioral biases. By analyzing written decision justifications, trading logs, or survey responses, these models can uncover systematic trends such as risk aversion, overconfidence, or the disposition effect. Investors are also subject to various financial behavioral biases, such as recency bias and authority bias [69]. Xiao et al. [70] evaluated these biases using a curated multimodal dataset, DynoStock, which integrates stock histories of S&P 500 companies with their quarterly Earnings Per Share (EPS) reports. Their findings indicate that recent open-sourced Large Vision-Language Models (LVLMs), such as LLaVA-NeXT, MobileVLM-V2, Mini-Gemini, MiniCPM-Llama3-V2.5, and Phi-3-vision-128 k, are significantly affected by these biases. In contrast, proprietary models like GPT-4o showed negligible influence. GPT-4o's superior performance

likely stems from its larger scale and meticulously curated training data, which helps reduce susceptibility to human-like biases. This underscores the critical importance of sufficiently long historical datasets in mitigating recency bias.

Zhou et al. [71] proposed the “Financial Bias Indicator” (FBI) framework, consisting of components like Bias Unveiler, Bias Detective, Bias Tracker, and Bias Antidote, to identify, analyze, and mitigate irrational biases in LLMs. An evaluation of 23 leading models revealed that nearly all demonstrated financial irrationality. Notably, FinLLMs exhibited more pronounced biases compared to smaller, general-purpose models. This phenomenon is attributed to the amplification of biases during continued pre-training or fine-tuning with financial data.

To address this, they created the FinCausal dataset, comprising 200,000 financial causal text entries extracted from research reports, and utilized a Retrieval-Augmented Generation (RAG) approach to recall relevant causal knowledge. Experimental results showed that recalling four causal knowledge entries per test point significantly improved reasoning capabilities and reduced biases in the models.

By implementing and refining these measures, FinLLMs are expected to achieve greater financial acuity and reduced biases. For example, models akin to cognitive bias alerts could be advanced to filter massive transaction data, identify client trading biases, and mitigate them. However, opportunities and challenges often coexist, particularly in finance-related fields where information security and privacy concerns are paramount, as discussed in detail in the following section.

5 Challenges and Limitations for LLMs in Financial Market

The training of LLMs requires substantial data, often including sensitive information, leading to significant privacy and security challenges [11, 72–77]. Yao et al. [11] reviewed LLM security and privacy issues, discussing beneficial applications (e.g., vulnerability detection, secure code generation), negative effects (e.g., phishing, social engineering), vulnerabilities (e.g., jailbreak attacks), and defense measures. Li et al. [72] categorized privacy attacks in LLMs, outlined defense strategies, and explored future directions for enhancing privacy. Similarly, Neel et al. [73] examined privacy risks, such as memory retention of sensitive data, and reviewed mitigation techniques, focusing on red-teaming efforts to reveal vulnerabilities. Derner et al. [74] investigated ChatGPT-specific risks, but the focus on ChatGPT alone limited its scope. Qammar et al. [75] traced chatbot evolution, identifying vulnerabilities but offering limited depth on specific solutions. Schwinn et al. [76] analyzed threats and defenses, emphasizing evolving adversarial attacks while lacking details on specific methodologies. Yan et al. [77] emphasized addressing privacy leakage and attacks at different stages of the LLM lifecycle, such as federated learning, differential privacy, knowledge unlearning, and hardware-assisted privacy protection. They highlighted the need to advance pivotal technologies for privacy protection while

acknowledging that each stage still has significant room for improvement or further research, indicating a long journey ahead.

Furthermore, AML applications underscore both the potential and risks of LLMs in financial compliance [23]. By analyzing complex transaction patterns and regulatory texts, LLMs enhance AML efforts, enabling institutions to detect suspicious activities and adhere to evolving standards [78, 79]. However, the widespread application of LLMs in compliance raises serious concerns about data privacy and security. To ensure information security, many organizations prefer deploying on-premises LLMs rather than public models for processing sensitive internal documents. For instance, both free and subscription-based public models are typically limited to non-sensitive tasks, such as translation or grammar checks for publicly available content. Nevertheless, even non-confidential internal data processed by public models carries the potential risk of inferring sensitive financial insights. Therefore, organizations prioritize internal LLMs for handling data that, while not strictly confidential, may indirectly expose critical financial information. These considerations highlight that while AI-driven financial technology offers substantial benefits, its potential risks must be carefully addressed.

Regulatory compliance poses another major challenge, particularly given the black-box nature of LLMs. A black box refers to a system whose internal workings are opaque or invisible to users. Users can input data and receive outputs, but the logic or code generating those outputs remains hidden. This characteristic is common in many AI systems, where any of the three components—algorithms, training data, or models—may function as a black box. While algorithms are often public, developers may choose to keep models or training data confidential to safeguard intellectual property. However, regulatory frameworks require explainable and traceable decision-making processes, especially in critical areas such as fraud detection or loan approval. For instance, if an LLM flags a transaction as suspicious, regulators may require clear evidence for this decision, which can be difficult to extract from opaque models. Explainable AI (XAI) techniques are being developed to address these limitations [80, 81], offering mechanisms to make LLM-driven decisions interpretable without compromising their computational efficiency. But most current XAI techniques focus on technical aspects, neglecting the user-developer interactions needed for trust. Interactive systems that provide explanations and feedback are also essential to empirically demonstrate the trustworthiness of AI systems to users and decision-makers [82].

Lastly, ethical concerns further complicate the adoption of LLMs in finance. Biased training datasets may perpetuate discriminatory practices, such as unfair loan denials or incorrectly flagging legitimate transactions as fraudulent. These biases often stem from insufficient diversity in training data or inherent flaws in human-labeled datasets, necessitating financial institutions to address such issues to ensure fairness. Moreover, the misuse of LLMs to manipulate markets or evade regulatory scrutiny raises accountability challenges [83]. Assigning responsibility for errors or harmful decisions—whether to developers, data providers, or end-users—remains contentious. Governance frameworks must evolve to establish clear accountability standards, such as tailored ethical guidelines for specific domains and

dynamic auditing systems adapted to varying environments to mitigate ambiguity in accountability.

6 Conclusion

This overview has traced the transformative journey of LLMs from their inception as general-purpose models to their specialization as FinLLMs. We began by outlining the pivotal role of general models such as GPT and BERT in laying the foundation for FinLLMs, emphasizing innovations like transformer-based architectures and in-context learning. Then, we highlighted the evolution toward domain-specific adaptations like FinBERT, BloombergGPT, and InvestLM, which leverage fine-tuning on financial datasets and instruction tuning to address unique demands in the financial sector.

A comparative analysis of FinLLM technologies revealed their strengths across various tasks. For example, BloombergGPT demonstrated remarkable breadth, excelling in regulatory compliance and sentiment analysis through its extensive training corpus, while FinMA and InvestLM showcased task-specific adaptability via instruction fine-tuning and the integration of structured and unstructured financial datasets. The evaluation of key tasks such as risk assessment, compliance optimization, and market sentiment analysis affirmed FinLLMs' ability to outperform traditional models in speed, accuracy, and context sensitivity. These advancements signify a paradigm shift in how financial data is analyzed, enabling better decision-making for investors and regulators alike.

Despite their promise, FinLLMs also present critical challenges and opportunities, as discussed throughout the overview. Data privacy remains a major concern, particularly given the reliance of these models on sensitive financial data. Many institutions prefer in-house models to mitigate risks of data leakage or inference of confidential insights. Similarly, model interpretability poses challenges in high-stakes environments like compliance and fraud detection, where decisions must be transparent and justifiable. Ethical considerations, such as bias in training data leading to unfair financial decisions or potential misuse for market manipulation, further underscore the need for robust safeguards. Future research should then focus on developing more explainable FinLLMs, integrating XAI methodologies to ensure transparency and accountability in decision-making. Additionally, innovative techniques for secure and federated learning can enhance data privacy without compromising model performance. Ethical guidelines tailored to FinLLMs should also be prioritized to navigate their dual potential for beneficial and harmful applications.

Acknowledgements Young Jin Kim provided excellent research assistance.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. ArXiv.org. <https://doi.org/10.48550/arXiv.1706.03762>
2. Lee J, Stevens N, Han SC, Song M (2024) A survey of large language models in finance (FinLLMs). arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2402.02315>
3. Devlin J, Chang M.-W, Lee K, Toutanova K (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT
4. Radford A, Narasimhan K, Salimans T, Sutskever I et al (2018) Improving language understanding by generative pre-training. OpenAI
5. Liu Z, Huang D, Huang K, Li Z, Zhao J (2021) FinBERT: A pre-trained financial language representation model for financial text mining. In: Proceedings of IJCAI, pp 4513–4519
6. Wu S, Irsoy O, Lu S, Dabrovolski V, Dredze M, Gehrmann S, Kambadur P, Rosenberg D, Mann G (2023) BloombergGPT: a large language model for finance. ArXiv. <https://doi.org/10.48550/arxiv.2303.17564>
7. Maia M, Handschuh S, Freitas A (2018) WWW'18 open challenge: financial opinion mining and question answering. In: Companion proceedings of WWW, pp 1941–1942
8. Yang Y, Uy MCS, Huang A (2020) FinBERT: a pretrained language model for financial communications
9. Xie Q, Han W, Zhang X, Lai Y, Peng M, Lopez-Lira A, Huang J (2023) PIXIU: a large language model, instruction data and evaluation benchmark for finance. ArXiv.org. <https://doi.org/10.48550/arXiv.2306.05443>
10. Yang Y, Tang Y, KY Tam (2023) InvestLM: a large language model for investment using financial domain instruction tuning. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2309.13064>
11. Yao Y, Duan J, Xu K, Cai Y, Sun E, Zhang Y (2023) A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2312.02003>
12. Summers TX, Huang SH, Choi PMS (2021) Can we predict tropical storms? Springer, Evidence from artificial intelligence. In Blockchain Technologies for IoT Applications
13. Clark K, Le QV, Manning CD (2020) Pre-training text encoders as discriminators rather than generators. arXiv preprint [arXiv:2003.10555](https://arxiv.org/abs/2003.10555)
14. Cortis K, Freitas A, Daudert T et al (2017) Semeval-2017 task 5: fine-grained sentiment analysis on financial microblogs and news. In: Proceedings of SemEval, pp 519–535
15. Radford A, Wu J, Child R, Luan D et al (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
16. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P et al (2020) Language models are few-shot learners. NeurIPS 33:1877–1901
17. Lewis P, Perez E, Piktus A, Petroni F et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS 33:9459–9474
18. Araci D (2019) FinBERT: financial sentiment analysis with pre-trained language models. arXiv preprint [arXiv:1908.10063](https://arxiv.org/abs/1908.10063)
19. Shah R, Chawla K, Eidnani D et al (2022) When flue meets FLANG: benchmarks and large pre-trained language model for financial domain. In: Proceedings of EMNLP, pp 2322–2335
20. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G (2023) LLaMA: open and efficient foundation language models. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2302.13971>
21. Hu EJ, Wallis P, Allen-Zhu Z, Li Y et al (2021) LoRA: low-rank adaptation of large language models. In ICLR
22. Kershaw D, Koeling R (2022) FinRED: a dataset for relation extraction in financial domain. <https://doi.org/10.1145/3487553.3524637>

23. Han J, Huang Y, Liu S, Towey K (2020) Artificial intelligence for anti-money laundering: a review and extension. *Digit Financ* 2(3–4):211–239. <https://doi.org/10.1007/s42521-020-00023-1>
24. Harbert T (2021) Tapping the power of unstructured data. MIT Sloan. <https://mitsloan.mit.edu/ideas-made-to-matter/tapping-power-unstructured-data>
25. Loukas L, Fergadiotis M, Chalkidis I, Spyropoulou E et al (2022) FINER: financial numeric entity recognition for XBRL tagging. In: *Proceedings of ACL*, pp 4419–4431
26. Hillebrand L, Deuser T, Dilmaghani T, Kliem B, Loitz R, Bauckhage C, R Sifa (2022) KPI-BERT: a joint named entity recognition and relation extraction model for financial reports. In: 2022 26th international conference on pattern recognition (ICPR). <https://doi.org/10.1109/icpr56361.2022.9956191>
27. Shah A, Paturi S, Chava S (2023) Trillion dollar words: a new financial dataset, task and market analysis. *ArXiv.org*. <https://doi.org/10.48550/arXiv.2305.07972>
28. Sinha A, Khandait T (2021) Impact of news on the commodity market: dataset and results. In: *Proceedings of FICC*, Springer, pp 589–601
29. Lee J, Youn HL, Stevens, N., Poon, J., & Soyeon Caren Han. (2021). FedNLP: an interpretable NLP system to decode federal reserve communications. *ArXiv (Cornell University)*. <https://doi.org/10.1145/3404835.3462785>
30. Casanueva I, Temčinas T, Gerz D, Henderson M, Vulic I (2020) Efficient intent detection with dual sentence encoders. In: *Proceedings of NLP4ConvAI Workshop*, pp 38–45
31. Zhao W, Zhang G, Yuan G, Liu J, Shan H, Zhang S (2020) The study on the text classification for financial news based on partial information. *IEEE Access* 8:100426–100437. <https://doi.org/10.1109/ACCESS.2020.2997969>
32. Mukherjee R, Bohra A, Banerjee A, Sharma S et al (2022) ECTSum: a new benchmark dataset for bullet point summarization of long earnings call transcripts. In: *Proceedings of EMNLP*, pp 10893–10906
33. Lee M, Lay-Ki S (2024) Finance wizard at the FinLLM challenge task: financial text summarization. *ArXiv.org*. <https://arxiv.org/abs/2408.03762>
34. Han J, Barman U, Hayes J, Du J, Burgin E, Wan D (2018) NextGen AML: distributed deep learning based language technologies to augment anti money laundering investigation, pp 37–42. <https://aclanthology.org/P18-4007.pdf>
35. Kaur S, Smiley C, Gupta A, Sain J, Wang D, Siddagangappa S, Aguda T, Shah S (2023) REFinD: relation extraction financial dataset. In: *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*. <https://doi.org/10.1145/3539618.3591911>
36. Cui X, Lam D, Verma A (2016) Embedded value in bloomberg news and social sentiment data. *Bloomberg LP*
37. Tetlock PC (2007) Giving content to investor sentiment: the role of media in the stock market. *J Financ* 62(3):1139–1168
38. Tetlock PC, Saar-Tsechansky M, Macskassy S (2008) More than words: quantifying language to measure firms' fundamentals. *J Financ* 63(3):1437–1467
39. Chen Z, Chen W, Smiley C, Shah S, Borova I et al (2021) FinQA: a dataset of numerical reasoning over financial data. In: *Proceedings of EMNLP*, pp 3697–3711
40. Chen Z, Li S, Smiley C, Ma Z, Shah S, Wang WY (2022) ConvFinQA: exploring the chain of numerical reasoning in conversational finance question answering. In: *Proceedings of EMNLP*, pp 6279–6292
41. Wu H, Zhang W, Shen W, Wang J (2018). Hybrid Deep Sequential Modeling for Social Text-Driven Stock Prediction. <https://doi.org/10.1145/3269206.3269290>
42. Qin Y, Song D, Chen H-F, Cheng W, Jiang G, Cottrell GW (2017) A dual-stage attention-based recurrent neural network for time series prediction. *ArXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1704.02971>
43. Lee J, Youn HL, Poon J, Han SC (2023) Stockemotions: discover investor emotions for financial sentiment analysis and multivariate time series. *AAAI-24 Bridge*

44. Lopez-Lira A, Tang Y (2023) Can ChatGPT forecast stock price movements? Return predictability and large language models. <https://doi.org/10.48550/arxiv.2304.07619>
45. Wu R (2024) Portfolio performance based on LLM news scores and related economical analysis. SSRN Electron J. <https://doi.org/10.2139/ssrn.4709617>
46. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, Fan Y, Ge W, Han Y, Huang F, Hui B, Ji L, Li M, Lin J, Lin R, Liu D, Liu G, Lu C, Lu K, Ma J (2023) Qwen technical report. ArXiv.org. <https://doi.org/10.48550/arXiv.2309.16609>
47. Yang A, Xiao B, Wang B, Zhang B, Bian C, Yin C, Lv C, Pan D, Wang D, Yan D, Yang F, Deng F, Wang F, Liu F, Ai G, Dong G, Zhao H, Xu H, Sun H, Zhang H (2023) Baichuan 2: open large-scale language models. ArXiv.org. <https://arxiv.org/abs/2309.10305>
48. Zhang W, Zhao L, Xia H, Sun S, Sun J, Qin M, Li X, Zhao Y, Zhao Y, Cai X, Zheng L, Wang X, An B (2024) A multimodal foundation agent for financial trading: tool-augmented, diversified, and generalist. ArXiv.org. <https://doi.org/10.48550/arXiv.2402.18485>
49. Kirtac K, Germano G (2024) Sentiment trading with large language models. *Financ Res Lett* 62:105227–105227. <https://doi.org/10.1016/j.frl.2024.105227>
50. Delgadillo J, Kinyua J, Mutigwe C (2024) FinSoSent: advancing financial market sentiment analysis through pretrained large language models. *Big Data Cogn Comput* 8(8):87–87. <https://doi.org/10.3390/bdcc8080087>
51. Fatourous G, Metaxas K, Soldatos J, Kyriazis D (2024) Can large language models beat wall street? Unveiling the potential of AI in stock selection. ArXiv.org. <https://doi.org/10.48550/arXiv.2401.03737>
52. Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS (2023) Generative agents: interactive simulacra of human behavior. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2304.03442>
53. Yu Y, Li H, Chen Z, Jiang Y, Li Y, Zhang D, Liu R, Suchow JW, Khashanah K (2023) FinMem: a performance-enhanced LLM trading agent with layered memory and character design. ArXiv.org. <https://doi.org/10.48550/arXiv.2311.13743>
54. Konstantinidis T, Iacovides G, Xu M, Constantinides TG, Mandic D (2024) FinLlama: financial sentiment classification for algorithmic trading applications. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2403.12285>
55. Van Duuren E, Plantinga A, Scholtens B (2016) ESG integration and the investment management process: fundamental investing reinvented. *J Bus Ethics* 138(3):525–533. <https://doi.org/10.1007/s10551-015-2610-8>
56. Vannoni V, Ciotti E (2020) Esg or not Esg? A benchmarking analysis. *Int J Bus Manag* 15(8):152. <https://doi.org/10.5539/ijbm.v15n8p152>
57. Eccles RG, Kastropeli MD, Potter SJ (2017) How to integrate ESG into investment decision-making: results of a global survey of institutional investors. *J Appl Corp Financ* 29(4):125–133
58. Mehra S, Louka R, Zhang Y (2022) ESGBERT: language model to help with classification tasks related to companies' environmental, social, and governance practices. *Embed Syst Appl*. <https://doi.org/10.5121/csit.2022.120616>
59. Lin LH-M, Ting F-K, Chang T-J, Wu J-W, Tsai RT-H (2024) GPT4ESG: streamlining environment, society, and governance analysis with custom AI models. <https://doi.org/10.1109/iceib61477.2024.10602567>
60. Huang K, Chen X, Yang Y, Ponnappalli J, Huang G (2023) ChatGPT in Finance and Banking. *Future of business and finance*, pp 187–218. https://doi.org/10.1007/978-3-031-45282-6_7
61. Johnson M (2023) A brave new world: ChatGPT's potential to reshape the financial services landscape. *Forbes* <https://www.forbes.com/sites/meaghanjohnson/2023/03/20/a-brave-new-world-chatgpts-potential-to-reshape-the-financial-landscape>
62. Stanley M (2023) Key milestone in innovation journey with OpenAI. <https://www.morganstanley.com/press-releases/key-milestone-in-innovation-journey-with-openai>
63. Top investment firms using AI for asset management (2023) U.S. News and World Report. <https://money.usnews.com/investing/articles/7-top-investment-firms-using-ai-for-asset-management>

64. Sanz-Guerrero M, Arroyo J (2024) Credit risk meets large language models: building a risk indicator from loan descriptions in peer-to-peer lending. <https://doi.org/10.2139/ssrn.4979155>
65. Papasavva A, Johnson S, Lowther E, Lundrigan S, Mariconti E, Markovska A, Tuptuk N (2024) Application of AI-based models for online fraud detection and analysis. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2409.19022>
66. Korkanti S (2024) Enhancing financial fraud detection using LLMs and advanced data analytics. In: 2024 2nd international conference on self sustainable artificial intelligence systems (ICSSAS). IEEE, pp 1328–1334
67. Simran T, Geetha J (2024) Enhancing graph database interaction through generative AI-driven natural language interface for financial fraud detection. In: 2024 15th international conference on computing communication and networking technologies (ICCCNT). IEEE, pp 1–8
68. Jackson R (2023) Understanding (and using) ChatGPT in banking. American bankers association. ABA Bank J 115(3):16–17. <https://www.proquest.com/scholarly-journals/understanding-using-chatgpt-banking/docview/2816135975/se-2>
69. Ricciardi V, Simon HK (2009) What is behavioral finance? Bus Educ Technol J 2(2):1–9. <https://ssrn.com/abstract=256754>
70. Xiao Y, Lin Y, Chiu M-C (2024) Behavioral bias of vision-language models: a behavioral finance view. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2409.15256>
71. Zhou Y, Ni Y, Liu X, Zhang J, Liu S, Ye G, Chai H (2024) Are large language models rational investors? ArXiv.org. <https://doi.org/10.48550/arXiv.2402.12713>
72. Li H, Chen Y, Luo J, Kang Y, Zhang X, Hu Q, Chan C, Song Y (2023) Privacy in large language models: attacks, defenses and future directions. ArXiv.org. <https://doi.org/10.48550/arXiv.2310.10383>
73. Neel S, Chang P (2023) Privacy issues in large language models: a survey. ArXiv.org. <https://doi.org/10.48550/arXiv.2312.06717>
74. Derner E, Batistić K (2023) Beyond the safeguards: exploring the security risks of ChatGPT. ArXiv.org. <https://doi.org/10.48550/arXiv.2305.08005>
75. Qammar A, Wang H, Ding J, Naouri A, Daneshmand M, Ning H (2023) Chatbots to ChatGPT in a cybersecurity space: evolution, vulnerabilities, attacks, challenges, and future recommendations. J Latex CI Files 14(8). <https://doi.org/10.48550/arxiv.2306.09255>
76. Schwinn L, Dobre D, Günnemann S, Gidel G (2023) Adversarial attacks and defenses in large language models: old and new threats. ArXiv.org. <https://doi.org/10.48550/arXiv.2310.19737>
77. Yan B, Li K, Xu M, Dong Y, Zhang Y, Ren Z, Cheng X (2024) On protecting the data privacy of large language models (LLMs): a survey. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2403.05156>
78. Alkhalili M, Qutqut MH, Almasalha F (2021) Investigation of applying machine learning for watch-list filtering in anti-money laundering. IEEE Access 9:18481–18496
79. Jiao M (2023) Big data analytics for anti-money laundering compliance in the banking industry 49:302–309. <https://doi.org/10.54097/hset.v49i.8522>
80. Mohseni S, Zarei N, Ragan ED (2018) A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.1811.11839>
81. Alicioglu G, Sun B (2021) A survey of visual analytics for explainable artificial intelligence methods. Comput Graph. <https://doi.org/10.1016/j.cag.2021.09.002>
82. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, Guidotti R, Ser JD, Díaz-Rodríguez N, Herrera F (2023) Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fus 99:101805. Sciencedirect. <https://doi.org/10.1016/j.inffus.2023.101805>
83. Jiao J, Afroogh S, Xu Y, Phillips C (2024) Navigating LLM ethics: advancements, challenges, and future directions. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2406.18841>

Housing Price Estimation and Reasoning Based on a Large Language Model



Seongeun Bae , Leehyun Jung , Sukyung Nam, Sihyun An ,
and Kwangwon Ahn

Abstract This study investigates the applicability of a large language model (LLM) to housing price appraisal in terms of predictive power and explainability. We first transform a hedonic dataset into a convertible format to construct the LLM-based appraisal framework using the established prompt engineering. We then compare the results to those obtained using a traditional hedonic pricing model. Our findings reveal that LLM outperforms the traditional benchmark model concerning two accuracy measures (i.e., root mean square error and R^2 value) in appraising housing prices. This outcome indicates the substantial capability of LLM for seizing nonlinearity in the hedonic dataset. Furthermore, the LLM-based appraisal framework provides three-dimensional interpretations, including (1) the directional impacts, (2) the qualitative importance of the hedonic variables concerning housing prices, and (3) narrative reasoning for the appraised prices. These findings reinforce that the proposed LLM-based valuation model is a potential tool for understanding the mechanism of housing prices. Investors can implement our framework to estimate properties and support decision-making through explainable LLM results. Moreover, policymakers can benchmark our results when developing monitoring systems and designing transparent real estate markets.

Keywords ChatGPT · Hedonic pricing model · Housing price · Large language model

S. Bae · L. Jung · K. Ahn (✉)

Department of Industrial Engineering, Yonsei University, Seoul, Korea
e-mail: k.ahn@yonsei.ac.kr

S. Bae · L. Jung · S. An · K. Ahn

Center for Finance and Technology, Yonsei University, Seoul, Korea

S. Nam

Ehyun Asset Management, Seoul, Korea

S. An

Department of Statistics and Data Science, Yonsei University, Seoul, Korea

1 Introduction

Real estate assets, which comprise substantial funds, have been appreciated by various valuation frameworks as the fluctuations in housing prices can affect the national economic system and financial stability [1–3]. The significance of accurate housing price estimation has garnered the interest of academic circles and practitioners in scrutinizing the way of constructing precise models for housing price appraisal. Furthermore, transparency has been fundamental in advancing housing price appraisal frameworks because unveiling how each housing factor affects property prices can foster the reliability of valuation models [4] and facilitate insightful interpretations derived from the results [5].¹

Hedonic price models (HPMs) have been intensively applied in appraising housing prices [8, 9] because they are well-versed approaches in interpreting the impacts of hedonic variables on housing prices. However, the embedded linearity in HPMs calls for more advanced valuation frameworks for precisely appraising housing prices [10, 11], including machine learning and deep learning models; however, such sophisticated models have been likened to a “black box” [12]. Their outstanding capability of seizing nonlinearity is boosted by the complex nature of model structures, resulting in an opaque valuation process. In this sense, housing price appraisals necessitate a stand-alone model framework alongside flexibility (to guarantee precise estimations) and transparency (to reinforce the reliability of the results).

As an alternative yet nascent approach, the existing literature has investigated how large language models (LLMs) can be used to explore market price movements [13, 14] and investment sentiment analysis [15, 16]. In a concurrent paper, Gloria et al. [17] demonstrated that an investment scenario using ChatGPT (a chat generative pre-trained transformer) can provide significantly profitable investment decisions. The studies above contended that LLMs could be state-of-the-art instruments for forecasting market returns and extracting significant information from textual resources; however, LLMs’ reasoning capability has been largely overlooked in finance studies. Moreover, the applicability of LLMs in appraising housing prices has yet to be explored.

We found three notable findings. (1) The LLM-based housing price appraisal model can value housing prices more precisely than the HPM by injecting contextual information into input prompts. (2) The estimated results can be interpreted for each hedonic variable’s qualitative importance and directional impact on housing prices. (3) The LLM-based valuation model can provide narrative reasoning for the appraised housing prices, where the rationale aligns with existing empirical evidence. Investors can utilize our research framework to evaluate real estate assets precisely and facilitate decision-making processes by supporting explainable results derived from LLM. Furthermore, policymakers can benchmark the results when developing monitoring systems and constructing transparent real estate markets.

¹ Identifying hedonic variables’ effects on housing prices can be a pivotal means in allocating public resources and services when constructing urban systems [6] and scheming timely policies and regulations [7].

The remainder of this study is as follows: Section 2 describes the dataset used and methodology conducted in the current study, Section 3 presents the results and discusses our findings, and Section 4 concludes the study.

2 Data and Methodology

2.1 Data

This study utilizes an aggregated hedonic dataset of 30,698 apartment² transaction records from Busan, South Korea, in 2018 and 2019. Busan is considered the survey area since this site is well urbanized. It has a large population and various environmental infrastructures, including natural parks, green spaces, and seafronts [8], leading to rich interpretations of how hedonic variables impact housing prices.

We refined our compiled dataset by eliminating outliers and duplicate records. From the whole housing factor group, 16 hedonic variables are used in appraising housing prices; the variable structure aligns with existing literature [5, 19–21]. The variables whose distributions show non-Gaussian features are transformed into a logarithmic scale. Table 1 summarizes the description of variables. The Variables column indicates the name of each hedonic variable denoted in this study and the Detail column delineates each hedonic variable's characteristics.

2.2 Experimental Design

Figure 1 graphically illustrates the analytical procedures. First, we transform the entire hedonic dataset into a convertible format to interact with LLM (i.e., the prompt form of numerical values). We then conduct prompt engineering to improve the capability of the LLM-based housing price appraisal model in terms of predictive power and reasoning tasks. Two prompt formats are used in the training procedure, followed by the hyperparameter optimization. We validate the predictive power by comparing the LLM-based model to HPM regarding root mean square error (RMSE) and R^2 based on randomly sampled sub-datasets from the whole dataset. Each sub-dataset has 2,000 observations, divided into a train set with a fraction of 0.6 and a test set with a fraction of 0.4, respectively. Finally, we compare the reasoning capability of the LLM-based appraisal model to the signs of HPM's regression coefficients and empirical evidence from existing literature regarding directional impacts and qualitative importance.

² Apartments are the dominant housing type in South Korea [18].

Table 1 Description of hedonic variables

Variables	Detail
Property prices*	Log-transformed Korean Won (KRW) per square meter (KRW/m ²)
<i>Property characteristics</i>	
Size	Unit size aggregated in square meters (m ²)
Floor	Floor level of transacted property
Year	The year of each apartment complex was built
Units	Number of households in an apartment complex
Parking	Number of parking spaces divided by the number of units
<i>Environmental amenities</i>	
Dist. green*	Log-transformed network distance to the nearest park, hill, or mountain in meters
Dist. water*	Log-transformed network distance to the nearest river, stream, pond, or seashore in meters
<i>Local built environment</i>	
Dist. subway*	Log-transformed network distance to the nearest subway station in meters
<i>Local demographics</i>	
Top univ.	Number of Seoul National University entrants from high schools within a 5-km radius of properties
Sex ratio	Percentage of the number of men divided by the number of women
Median age	The age that divides the population equally
Pop. density*	Log-transformed number of people per square kilometer (km ²)
Higher degree	Percentage of the number of people with a higher degree divided by the number of people aged 15 years or older
<i>Seasonality controls</i>	
Spring	Seasonal dummy indicating that a transaction occurred in March, April, or May
Fall	Seasonal dummy indicating that a transaction occurred in September, October, or November
Winter	Seasonal dummy indicating that a transaction occurred in December, January, or February

Note Variables with an asterisk have been transformed into a logarithmic scale

2.3 HPM

HPM has long been utilized in appraising housing prices [8–10, 22], and HPM can be a representative approach for capturing each hedonic variable’s marginal contribution to housing prices [23, 24]. The log-linear form of HPM can be expressed as follows:

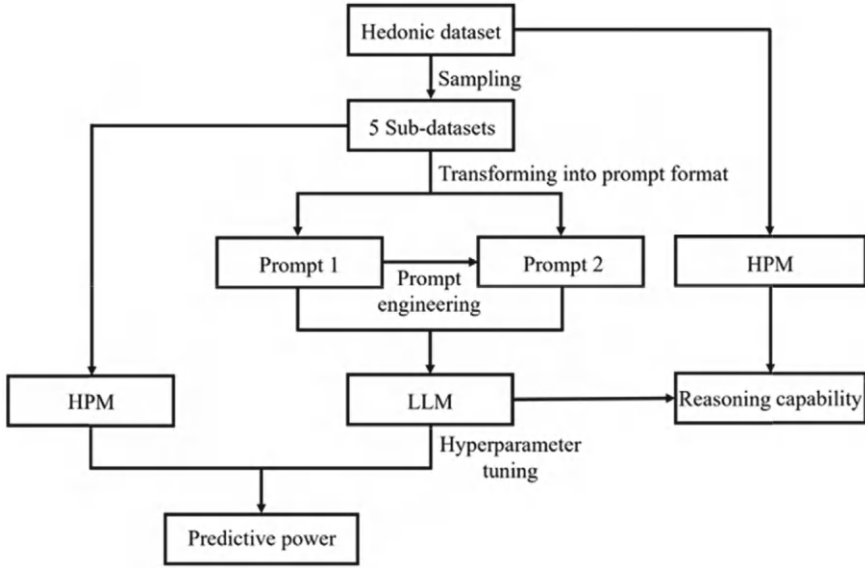


Fig. 1 Flowchart of experimental design

$$\ln p_i = \alpha + \sum_{j=1}^J \beta_j x_{ij} + \sum_{k=1}^K \theta_k Q_{ik} + e_i, \quad (1)$$

where p_i represents the price per square meter of each transacted housing unit i ; α is the intercept; J is the number of hedonic variables, including property characteristics, local demographics, and environmental amenities; $K (= 3)$ refers to seasonal dummy variables; the coefficients β_j and θ_k are regression coefficients using the ordinary least squares method; and e_i is the residual error.

2.4 GPT-4o-Based Housing Price Appraisal Model

Generative pre-trained transformers (GPT), such as GPT-3.5 (ChatGPT) and GPT-4, have shown their superiority in finance studies, including financial text analysis [25] and short/long-term times series forecasting [26]. Specifically, the GPT-based LLMs take the input form of a “token” and mainly comprise autoregressive decoder architecture. This architecture sequentially predicts the next token by leveraging information from the previous token. This intrinsic nature can facilitate the learning of contextual relationships between tokens [27], showcasing that GPT-based LLMs (GPT-4) are proficient on financial literacy tests [28].

The GPT-based LLMs can be adapted to new tasks alongside a few training steps [29]; hence, we conjecture that the GPT-based housing price appraisal model

has a competitive predictive power compared to HPM. Specifically, we expect that LLM’s capability of appraising housing prices can effectively persist even when it is trained on a small portion of the dataset. Furthermore, GPT models have substantial reasoning capability [30] via the input-output context window [31], contextually explaining and rationalizing the appraised housing prices. This study employs the GPT-4o model, the advanced version of the GPT-4 model,³ to develop the LLM-based housing price appraisal model.

2.5 Fine-Tuning

We tune the GPT-4o model utilizing the application programming interface. LLMs have increased their sizes with tremendous parameters to address various disciplines; however, a stand-alone scaling up may require additional training tactics to significantly improve understanding of the given tasks [33], such as logical and mathematical reasoning [34]. Therefore, we configure two fine-tuning strategies: prompt engineering and optimizing hyperparameters.

Previous studies show that prompt engineering can be a practical approach to optimize and improve the LLMs’ applicability to downstream tasks [35] in terms of precision [36] and accuracy of reasoning [37]. Accordingly, we introduce the prompt engineering approach to enhance the predictive power and transparency of the GPT-4o-based housing price appraisal model. First, the role-play framing is conducted when configuring each prompt to improve LLM’s understandability and capability [38] regarding appraising housing prices and reasoning the results. That is, we impose the model’s role when initiating the prompting.

Referencing Ding et al. [39] and Hegselmann et al. [40], we first construct a straightforward input prompt (Prompt 1) and an engineered prompt (Prompt 2) by injecting contextual information corresponding to each hedonic variable. Prompt 1 configures the input prompt by naively arraying the numerical values of hedonic variables and Prompt 2 further includes the contextual description for hedonic variables, as summarized in Table 1. Thus, we designed two GPT-4o-based housing price appraisal models for each prompt in the comparative analysis.

In the optimization process, OpenAI primarily suggests the tunable hyperparameter set of epochs, batch size, and scaling factor for learning rate [41]. Accordingly, we iterate the training steps to find an optimal hyperparameter set to achieve the highest training accuracy. This approach results in the triplet of (3, 2, 2) for (*epochs*, *batch size*, and *scaling factor for the learning rate*), respectively. Figure 2 graphically describes our fine-tuning strategies, including role-play framing, and exemplifies some prompts used in this study. Figure 3 presents the output of LLMs derived from the constructed prompts and fine-tuning strategies.

³ The GPT-4 model is architected based on the transformer style and pre-trained on a variety of documents [32].

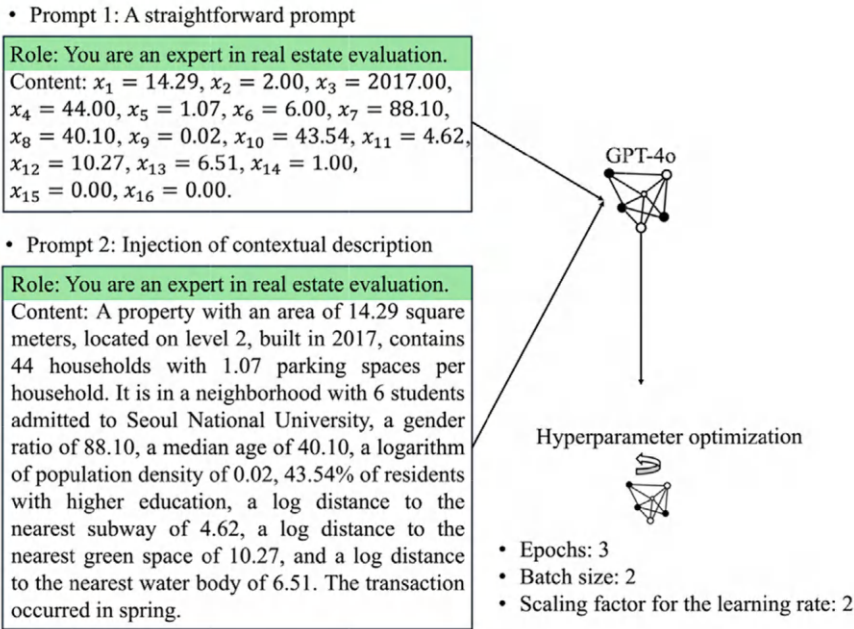


Table 2 Comparison of predictive power

	HPM		Prompt 1		Prompt 2	
	RMSE	R ²	RMSE	R ²	RMSE	R ²
Dataset 1	0.3028	0.6394	0.2922	0.6709	0.2678	0.7236
Dataset 2	0.3190	0.6476	0.2667	0.7585	0.2792	0.7354
Dataset 3	0.2975	0.7026	0.2596	0.7781	0.2476	0.7980
Dataset 4	0.2824	0.7218	0.2581	0.7723	0.2463	0.7925
Dataset 5	0.3033	0.6862	0.2538	0.7848	0.2779	0.7418
Average	0.3010	0.6795	0.2661	0.7529	0.2638	0.7583

3 Results and Discussion

3.1 Predictive Power of Appraising Housing Prices

We first compare the predictive power of GPT-4o-based housing price appraisal models to that of HPM regarding RMSE and R². Table 2 shows that the GPT-4o-based housing price appraisal models surpass the HPM across all sub-datasets. On average, the LLM trained on Prompt 2 shows a higher predictive power than that trained on Prompt 1. This finding implies that our prompt engineering strategy can facilitate the capability of LLM to appraise housing prices.

Figure 4 highlights that injecting contextual description into the prompting (i.e., Prompt 2) can improve the precision of LLM’s housing price appraisal. When considering the standard deviation ($\pm 1\sigma$) from the average of R², LLM’s accuracy boundary with Prompt 1 overlaps with that of HPM, as shown in Panel (a) of Fig. 4. In contrast, the LLM trained on Prompt 2 has a distinct accuracy boundary against the HPM, as shown in Panel (b) of Fig. 4. This outcome reinforces that prompt engineering can enhance the applicability of LLM in terms of predictive power [36], leading to the precise capture of nonlinearity embedded in the housing price dataset.

3.2 Reasoning Capability of LLM Retrieved from Each Prompt

After training two LLMs on each prompt to appraise housing prices, we investigate their interpretability and reasoning capability through additional prompting. Figure 5 presents the question prompt and the reasoning results for each LLM. Prompt 1, configured with naively arrayed numerical values, misinterprets the appraised housing prices along with irrelevant housing factors. In contrast, Prompt 2 desirably explains how each hedonic variable used in this study affects housing prices. These results align with existing evidence and show that prompt engineering can

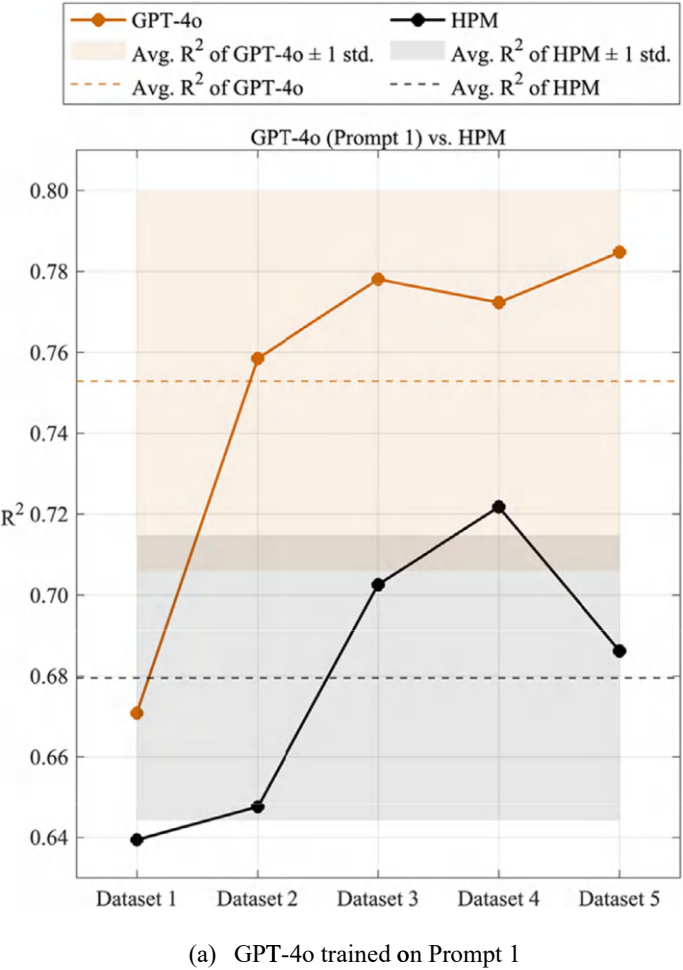
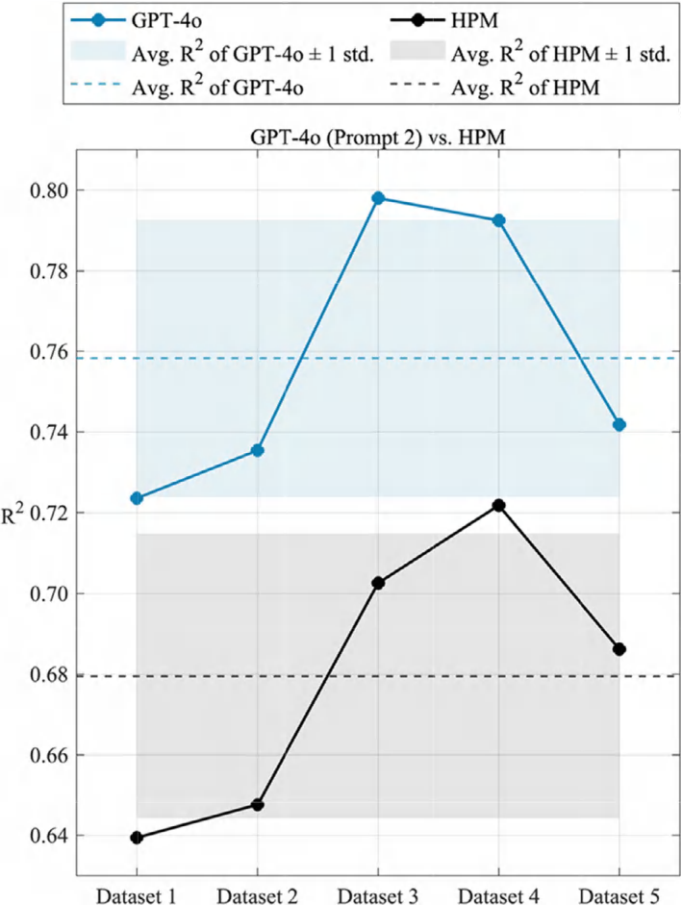


Fig. 4 Model results of R^2 for sub-datasets (a) GPT-4o trained on prompt 1, (b) GPT-4o trained on prompt 2

improve the understandability of LLM to given tasks [33], enhancing the accuracy of reasoning capability [37].

3.3 Validation of Reasoning Capability

We validate the reasoning capability of LLM with Prompt 2, which shows more plausible interpretation results against the LLM with Prompt 1. Accordingly, we compare the reasoning results of LLM with Prompt 2 to the signs of regression



(b) GPT 4o trained on Prompt 2

Fig. 4 (continued)

coefficients in the HPM concerning directional impacts and to the existing empirical evidence regarding the qualitative importance in appraising housing prices.

While the LLM is tailored to Prompt 2 and a sub-dataset, the sign of each regression coefficient in HPM is derived from the whole dataset. This setup enables us to identify whether the fine-tuned LLM with little information can have compatible interpretability in the comparative analysis concerning directional impacts on housing prices.

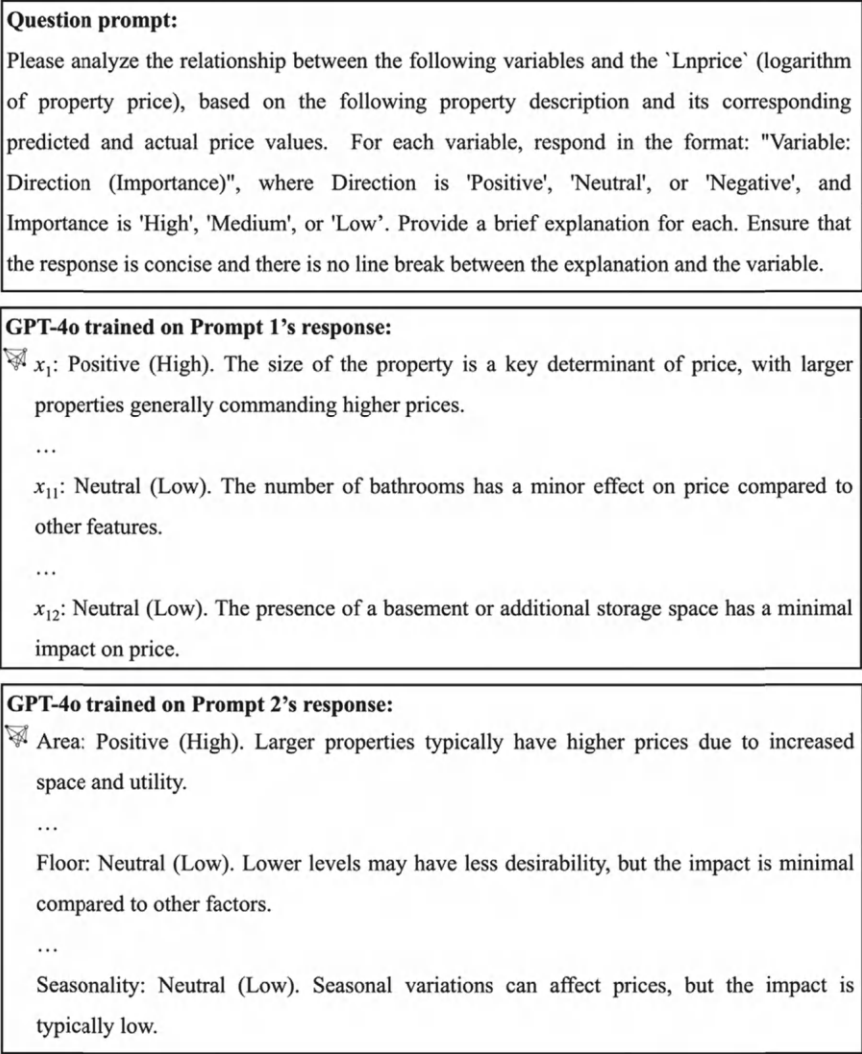


Fig. 5 Question prompt and reasoning outputs for each LLM

3.3.1 Directional Impacts and Qualitative Importance of Hedonic Variables

Table 3 summarizes the directional impact and qualitative importance derived from the two models. The interpretations regarding the qualitative importance of the GPT-4o model are layered as (1) high- or mid-level significance for housing characteristics, urban infrastructures, and local demographics and (2) low-level significance for seasonality controls.

Table 3 Signs of regression coefficients, directional impacts, and qualitative importance

Variables	HPM	GPT				
		Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Size	Positive	Positive (High)	Positive (High)	Positive (High)	Positive (High)	Positive (High)
Floor	Positive	Neutral (Low)	Neutral (Low)	Neutral (Low)	Neutral (Low)	Neutral (Low)
Year	Positive	Positive (High)	Positive (Medium)	Positive (Medium)	Positive (High)	Positive (Medium)
Units	Positive	Neutral (Low)	Neutral (Low)	Neutral (Low)	Neutral (Low)	Neutral (Low)
Parking	Positive	Positive (Medium)	Positive (Medium)	Positive (Medium)	Positive (Medium)	Positive (Medium)
Dist. green	Negative	Negative (High)	Negative (Medium)	Negative (Medium)	Negative (Medium)	Negative (Medium)
Dist. water	Negative	Neutral (Low)	Neutral (Low)	Neutral (Low)	Neutral (Low)	Neutral (Low)
Dist. subway	Negative	Negative (High)	Negative (High)	Negative (High)	Negative (High)	Negative (High)
Top univ.	Positive	Positive (Medium)	Positive (Medium)	Positive (Medium)	Positive (Medium)	Positive (Medium)
Sex ratio	Negative	Neutral (Low)	Negative (Low)	Negative (Low)	Negative (Low)	Negative (Low)
Median age	Positive	Negative (Medium)	Negative (Medium)	Negative (Medium)	Negative (Medium)	Negative (Medium)
Pop. density	Positive	Neutral (Low)	Neutral (Low)	Neutral (Low)	Neutral (Low)	Neutral (Low)
Higher degree	Positive	Positive (High)	Positive (Medium)	Positive (Medium)	Positive (High)	Positive (Medium)
Seasonality	Positive	Neutral (Low)	Neutral (Low)	Neutral (Low)	Neutral (Low)	Neutral (Low)

Note Positivity in the *Seasonality* variable indicates that all seasonal dummy variables (spring, fall, and winter) positively impact housing prices; this variable is derived from GPT-4o based on Prompt 2. The qualitative importance is described in each parenthesis. The shaded areas denote the variables of high- or mid-level qualitative importance

In sub-dataset cases, the LLM concludes that the directional impacts on housing prices are consistent for the group of hedonic variables that hold high- or mid-level quantitative importance. The same group’s directional impacts are accorded with the signs of regression coefficients in HPM, except for the case of *Medium Age*. This outcome implies that the LLM with a small portion of the information set can be compatible with the traditional econometric model with a complete information set in terms of identifying the directional impact of each hedonic variable on housing price.

Table 4 Reasoning results of GPT-4o-based prompt 2

Variables	Reasoning description
Size	Larger properties typically have higher prices due to increased space and utility
Floor	Lower levels may have less desirability, but the impact is minimal compared to other factors
Year	Newer properties are generally more desirable and command higher prices
Units	The number of households can indicate community size but has a lesser impact on price
Parking	More parking spaces per household increase convenience and desirability
Dist. green	Closer proximity to green spaces is generally preferred
Dist. water	While proximity to water can be desirable, its impact here is minimal
Dist. subway	Proximity to subway stations is a key factor in property desirability
Top univ.	Proximity to prestigious schools often increases property values
Sex ratio	The gender ratio has minimal direct impact on property prices
Median age	Older neighborhoods might be less desirable, impacting prices negatively
Pop. density	While density can affect desirability, its impact is less significant
Higher degree	Areas with more educated residents often have higher property values
Seasonality	Seasonal variations can affect prices, but the impact is typically low

Note The shaded areas denote the variables of high- or mid-level qualitative importance

3.3.2 Rationale for the Interpretation Results of LLM

Table 4 represents the results responding to the prompt, “...Provide a brief explanation for each.” We highlight the results of Dataset 4 in which the GPT-4o based on Prompt 2 shows the highest predictive power in RMSE.⁴

Each description shows the directional impacts and/or qualitative importance of each hedonic variable in which information is narratively reasoned. For example, the delineation concerning the *Sex Ratio* variable mainly portrays the qualitative importance in relation to housing prices. In the case of explaining the *Year* variable, the description articulates, “*Newer properties are generally more desirable and command higher prices.*” This describes both the qualitative importance and directional impacts on housing prices.

Finally, we compare the results of narrative reasoning to existing empirical evidence. The GPT-4o-based reasoning results indicate that residents pay more premiums for spacious dwelling areas [42], newly constructed buildings [43], and affordable parking spaces [44]. In short, housing characteristics like *Size*, *Year*, and *Parking* have high- and mid-level qualitative importance with positivity concerning housing prices [5]. Moreover, the LLM acknowledges the impacts of proximity to

⁴ The LLM with Prompt 2 on Dataset 3 also has the most precise accuracy in terms of R^2 ; thus, we checked the case of Dataset 3 and confirmed the symmetric findings with those in the case of Dataset 4.

transportation systems [19] and green spaces [8, 45] on housing prices. Overall, these findings support the applicability of LLM in analyzing the relationship between hedonic variables and housing prices. Through interactive prompt inputs, the LLM-based housing price appraisal model can provide three-dimensional interpretation, including directional impacts, qualitative importance, and narrative explanations.

4 Conclusion

This study investigated the potential of GPT-4o-based LLM in appraising housing prices and delineated the estimated results. We improved the predictive power and reasoning capability of LLM using fine-tuning strategies, including prompt engineering and hyperparameter optimization. We compared the precision of housing price appraisals and the validity of interpretations to the results of HPM and previous studies. Our findings demonstrated that incorporating contextual descriptions for each hedonic variable into the prompting process can foster precise housing price appraisals and clear interpretations.

Findings suggest the implications focusing on LLM's competitive predictive power and reasoning capability concerning housing prices. First, investors can consider engaging LLMs as their valuation frameworks to estimate real estate assets and benefit decision-making processes from a triplet of reasoning results. Policy-makers can use LLM's scalability to construct real-time monitoring systems and scheme timely regulations for transparent real estate markets.

Acknowledgment This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C1004258, Kwangwon Ahn) and the NRF, Korea, under project BK21 FOUR (Big Data-Based Interdisciplinary Education and Research for Data Science).

References


1. Carrasco-Gallego JA (2020) Real estate, economic stability and the new macro-financial policies. *Sustainability* 13(1):236
2. Gertler M, Gilchrist S (2018) What happened: financial factors in the great recession. *J Econ Perspect* 32(3):3–30
3. Jordà Ò, Schularick M, Taylor AM (2015) Betting the house. *J Int Econ* 96:S2–S18
4. Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat MAA, Dwivedi YK (2023) A systematic literature review of artificial intelligence in the healthcare sector: benefits, challenges, methodologies, and functionalities. *J Innov Knowl* 8(1):100333
5. An S, Ahn K, Bae J, Song Y (2024) Economic impacts of a subway system: exploring local contexts in a metropolitan area. *Res Transp Bus Manag* 56:101188
6. Hu N, Legara EF, Lee KK, Hung GG, Monterola C (2016) Impacts of land use and amenities on public transport use, urban planning and design. *Land Use Pol* 57:356–367

7. Iban MC (2022) An explainable model for the mass appraisal of residences: the application of tree-based machine learning algorithms and interpretation of value determinants. *Habitat Int* 128:102660
8. An S, Jang H, Kim H, Song Y, Ahn K (2023) Assessment of street-level greenness and its association with housing prices in a metropolitan area. *Sci Rep* 13(1):22577
9. Chau K, Chin TL (2003) A critical review of literature on the hedonic price model. *Int J Hous Sci Appl* 27(2):145–165
10. Hong J, Choi H, Kim WS (2020) A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. *Int J Strateg Prop Manag* 24(3):140–152
11. Li S, Jiang Y, Ke S, Nie K, Wu C (2021) Understanding the effects of influential factors on housing prices by combining extreme gradient boosting and a hedonic price model (XGBoost-HPM). *Land* 10(5):533
12. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1(5):206–215
13. Lopez-Lira A, Tang Y (2023) Can ChatGPT forecast stock price movements? Return predictability and large language models. Preprint at [arXiv:2304.07619](https://arxiv.org/abs/2304.07619)
14. Menéndez Medina A, Heredia Álvaro JA (2024) Using generative pre-trained transformers (GPT) for electricity price trend forecasting in the Spanish market. *Energies* 17(10):2338
15. Li M, Chen L, Zhao J, Li Q (2021) Sentiment analysis of Chinese stock reviews based on BERT model. *Appl Intell* 51:5016–5024
16. Li M, Li W, Wang F, Jia X, Rui G (2021) Applying BERT to analyze investor sentiment in stock market. *Neural Comput Appl* 33:4663–4676
17. Gloria B, Melsbach J, Bienert S, Schoder D (2024) Real-GPT: efficiently tailoring LLMs for informed decision-making in the real estate industry. *J Real Estate Portf Manag* 1–17
18. Jang H, Ahn K, Kim D, Song Y (2018) Detection and prediction of house price bubbles: evidence from a new city. In: *Computational science–ICCS 2018: 18th international conference, Wuxi, China*, pp 782–795
19. Ahn K, Jang H, Song Y (2020) Economic impacts of being close to subway networks: a case study of Korean metropolitan areas. *Res Transp Econ* 83:100900
20. Dahal RP, Grala RK, Gordon JS, Munn IA, Petrolia DR, Cummings JR (2019) A hedonic pricing method to estimate the value of waterfronts in the Gulf of Mexico. *Urban Urban Green* 41:185–194
21. Dai X, Felsenstein D, Grinberger AY (2023) Viewshed effects and house prices: identifying the visibility value of the natural landscape. *Landsc Urban Plan* 238:104818
22. Yang J, Rong H, Kang Y, Zhang F, Chegut A (2021) The financial impact of street-level greenery on New York commercial buildings. *Landsc Urban Plan* 214:104162
23. Boyle M, Kiel K (2001) A survey of house price hedonic studies of the impact of environmental externalities. *J Real Estate Lit* 9(2):117–144
24. Sirmans S, Macpherson D, Zietz E (2005) The composition of hedonic pricing models. *J Real Estate Lit* 13(1):1–44
25. Yang H, Liu XY, Wang CD (2023) FinGPT: open-source financial large language models. Preprint at [arXiv:2306.0603](https://arxiv.org/abs/2306.0603)
26. Zhou T, Niu P, Sun L, Jin R (2023) One fits all: power general time series analysis by pretrained LM. In: *37th conference on neural information processing systems, Los Angeles, USA*, pp 43322–43355
27. Shrestha P, Kandel J, Tayara H, Chong KT (2024) Post-translational modification prediction via prompt-based fine-tuning of a GPT-2 model. *Nat Commun* 15(1):6699
28. Niszczota P, Abbas S (2023) GPT has become financially literate: insights from financial literacy tests of GPT and a preliminary test of how people use it as a source of advice. *Finance Res Lett* 58:104333
29. Yang K, Tao J, Lyu J, Ge C, Chen J, Shen W, Zhu X, Li X (2024) Using human feedback to fine-tune diffusion models without any reward model. In: *IEEE/CVF conference on computer vision and pattern recognition, Seattle, USA*, pp 8941–8951

30. Huang LY, Zhang X, Wang Q, Chen ZS, Liu Y (2024) Evaluating media knowledge capabilities of intelligent search dialogue systems: a case study of ChatGPT and new Bing. *J Knowl Econ* 1–24
31. Hagendorff T, Fabi S, Kosinski M (2023) Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nat Comput Sci* 3(10):833–838
32. OpenAI (2023) GPT-4 technical report. Preprint at [arXiv:2303.08774](https://arxiv.org/abs/2303.08774)
33. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. In: 36th conference on neural information processing systems, Los Angeles, USA, pp 24824–24837
34. Rae J, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, Aslanides J., Henderson S, Ring R, Young S, Rutherford E (2021) Scaling language models: methods, analysis & insights from training gopher. Preprint at [arXiv:2112.11446](https://arxiv.org/abs/2112.11446)
35. Fang T, Zhang Y, Yang Y, Wang C, Chen L (2024) Universal prompt tuning for graph neural networks. In: 37th conference on neural information processing systems, Vancouver, Canada, pp 52464–52489
36. Maharjan J, Garikipati A, Singh NP, Cyrus L, Sharma M, Ciobanu M, Barnes G, Thapa R, Mao Q, Das R (2024) OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Sci Rep* 14(1):14156
37. Polak MP, Morgan D (2024) Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat Commun* 15(1):1569
38. Shanahan M, McDonell K, Reynolds L (2023) Role play with large language models. *Nature* 623(7987):493–498
39. Ding JE, Thao PNM, Peng WC, Wang JZ, Chung CC, Hsieh MC, Tseng YC, Chen L, Luo D, Wu C, Wang CT (2024) Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records. *Sci Rep* 14(1):20774
40. Hagselmann S, Buendia A, Lang H, Agrawal M, Jiang X, Sontag D (2023) TabLLM: Few-shot classification of tabular data with large language models. In: 26th international conference on artificial intelligence and statistics, Valencia, Spain, pp 5549–5581
41. Deng Y, Xing Y, Quach J, Chen X, Wu X, Zhang Y, Moureaud C, Yu M, Zhao Y, Wang L, Zhong S (2024) Developing large language models to detect adverse drug events in posts on X. *J Biopharm Stat* 1–12
42. Li LH, Cheung D, Sun H (2015) Does size matter? The dynamics of housing sizes and prices in Hong Kong. *J Hous Built Environ* 30(1):109–124
43. Liu N, Strobl J (2023) Impact of neighborhood features on housing resale prices in Zhuhai (China) based on an (M) GWR model. *Big Earth Data* 7(1):146–169
44. Vargas-Calderón V, Camargo JE (2022) Towards robust and speculation-reduction real estate pricing models based on a data-driven strategy. *J Oper Res Soc* 73(12):2794–2807
45. Wu C, Du Y, Li S, Liu P, Ye X (2022) Does visual contact with green space impact housing prices? An integrated approach of machine learning and hedonic modeling based on the perception of green space. *Land Use Pol* 115:106048

Advancing Quantitative Trading Strategies Using Fine-Tuned Open-Source Large Language Models: A Hybrid Approach with Numerical and Textual Data Integration Using RAG and LoRA Techniques



Seth H. Huang , Jimin Kim, and Ka Lok Kellogg Wong

Abstract This paper explores the latest methodologies for fine-tuning open-source Large Language Models (LLMs) in enhancing quantitative trading strategies by integrating numerical data (e.g., historical prices, technical indicators) with textual data (e.g., news, earnings reports, social media sentiment). We employ Retrieval-Augmented Generation (RAG) with a vector database to efficiently handle and contextualize textual data, alongside Low-Rank Adaptation (LoRA) techniques for cost-effective and scalable model fine-tuning. The proposed approach aims to create a hybrid trading model that combines the predictive power of LLMs with traditional quantitative methods, improving accuracy and adaptability in financial markets. This study details the implementation process, highlighting practical innovations such as the integration of real-time data pipelines and adaptive model tuning. Experimental results show significant improvements in predictive accuracy and risk-adjusted returns, demonstrating the practical value of these advanced fine-tuning methodologies in finance.

Keywords Large Language Models (LLMs) · Quantitative trading strategies · Retrieval-Augmented Generation (RAG) · Low-Rank Adaptation (LoRA) · Hybrid trading models

S. H. Huang (✉) · K. L. K. Wong
Business School, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
e-mail: sethuang@ust.hk

K. L. K. Wong
e-mail: klkwongac@connect.ust.hk

J. Kim
College of Arts and Sciences, Cornell University, Ithaca, New York, USA
e-mail: jk2756@cornell.edu

1 Introduction

This paper explores the latest methodologies for fine-tuning open-source Large Language Models (LLMs) in enhancing quantitative trading strategies by integrating numerical data (e.g., historical prices, technical indicators) with textual data (e.g., news, earnings reports, social media sentiment).

We employ Retrieval-Augmented Generation (RAG) with a vector database to efficiently handle and contextualize textual data, alongside Low-Rank Adaptation (LoRA) techniques for cost-effective and scalable model fine-tuning [1–3]. The proposed approach aims to create a hybrid trading model that combines the predictive power of LLMs with traditional quantitative methods, improving accuracy and adaptability in financial markets.

This study details the implementation process, highlighting practical innovations such as the integration of real-time data pipelines and adaptive model tuning. Experimental results show significant improvements in predictive accuracy and risk-adjusted returns, demonstrating the practical value of these advanced fine-tuning methodologies in finance.

Section 2 details the experimental design, including data collection, pre-processing, and fine-tuning for numeric and textual data. It also covers the use of RAG and LoRA to enhance model performance. Section 3 presents results, evaluating the model with metrics like the Sharpe ratio, and discusses limitations, the penetration effect on the price movement of NVDA, and future improvements such as expanded datasets and advanced trading strategies.

2 Experimental Design and Methodology

This study focuses on trading stocks based on news and earnings announcements while also investigating the potential penetration effect. It analyzes 50 stocks from industries closely connected to NVIDIA (NVDA) to evaluate whether their price movements influence NVDA's performance. Covering the period from July 31, 2023, to October 31, 2024, this research serves as a proof of concept.

2.1 Data Collection

Numeric Data

Historical financial data, including stock prices, trading volumes, and market indicators, was sourced from reputable platforms such as Bloomberg, Reuters, and various financial data APIs. Minute-level closing price data and volatility metrics for the selected stocks were retrieved through the Polygon API, providing in-depth insights

into market dynamics and enabling a comprehensive analysis of price movements and volatility trends.

Earnings surprise data were collected for all 50 stocks based on quarterly announcements and compared against Bloomberg analysts' consensus estimates. Key metrics analyzed include revenue surprise (%), EPS surprise (%), EBIT and EBITDA surprise (%), gross margin surprise (%), pretax income (loss) surprise (%), and net income surprise (%). The data were meticulously mapped to Refinitiv's announcement timelines to ensure accurate temporal alignment, enabling a precise assessment of how earnings surprises impacted stock performance during the study period.

Textual Data

Social media sentiment was analyzed using the Reddit API (PRAW) to extract posts related to the 50 selected stocks within the specified time window. The API provided metadata such as the number of likes, dislikes, and comments for each post. These additional metrics enable feature engineering, which will be explored further in Sect. 2.2.

Earnings call transcripts were scraped from Capital IQ using the Beautiful Soup library, extracting both timestamps and full transcript content. This data provides insights into the timing and context of key statements, enhancing the understanding of company performance and management commentary.

News articles with accurate timestamps were sourced from Refinitiv, primarily focusing on Reuters as a reliable provider of real-time financial news. This ensures the data captures market-relevant events in real time, enabling analysis of stock price changes directly influenced by news releases. To maintain accuracy, GPT was not used to gather news data, as its summaries may lack real-time detail and fail to reflect immediate market impacts.

2.2 Data Pre-processing

Numeric Data Pre-processing

To ensure a reliable dataset, numerical data was standardized, missing values were addressed, and key features were engineered.

Trend indicators, such as the Simple Moving Average (SMA) and Exponential Moving Average (EMA), were derived from the engineered minute-level price data collected. These indicators facilitate the identification of price trends by smoothing fluctuations. Volatility metrics, including Standard Deviation and Bollinger Bands, measure price variability and assist in detecting overbought or oversold conditions. Momentum indicators, such as the Relative Strength Index (RSI) and Momentum, assess the rate and magnitude of price movements. Additionally, tools like the Moving Average Convergence Divergence (MACD) and Williams %R are employed to identify market extremes and provide buy or sell signals.

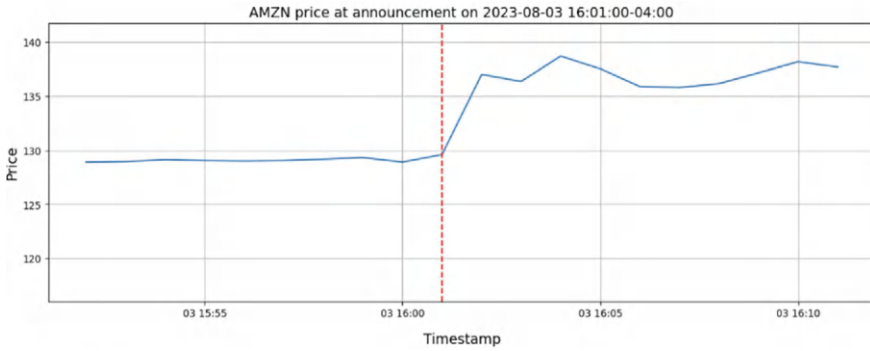


Fig. 1 AMZN price reaction to earnings announcement on August 3, 2023. *Source* Authors

10-min pre-and post-announcement stock price changes are extracted from the existing price data, to examine the market reactions to earnings surprises. Accurate data is sourced and validated using Refinitiv timestamps, enhancing the model’s ability to predict the impact of financial events on stock prices (Fig. 1).

Textual Data Pre-processing

Clean and pre-process textual data by removing noise, performing tokenization, and applying NER to identify relevant financial entities. Use sentence embeddings to convert texts into vectors suitable for further analysis.

The news dataset was refined through a structured cleaning process to ensure relevance and consistency for modeling. Starting with 71,059 articles, duplicates and irrelevant content, such as stock price targets, general summaries, and opinion pieces, were removed, reducing the dataset to 31,667 articles. Articles without corresponding price data at the specified timestamp were excluded, narrowing the dataset to 18,717 entries. These were converted into 3,072-dimensional word embeddings using OpenAI’s “text-embedding-3” model [4] and stored in a vector database.

2.3 Fine-Tuning Methodologies

Retrieval-Augmented Generation (RAG)

The integration with a vector database involves implementing Retrieval-Augmented Generation (RAG) using a tool like Pinecone to efficiently retrieve relevant textual information based on the contextual queries provided by the LLM. This approach allows the model to ground its predictions in real-time, relevant data, enabling it to produce more accurate and context-aware outputs [1].

The implementation involves setting up a pipeline where incoming financial news and reports are indexed into a vector database. The LLM is fine-tuned to dynamically query this database, retrieving the most relevant information, such as similar past

news events, their associated technical indicators, and the respective price changes. This process enhances the model's ability to generate informed and accurate outputs for trading signal generation.

Low-Rank Adaptation (LoRA)

- **Advanced LoRA Techniques:** Utilize LoRA to fine-tune large open-source models (such as Llama 3.2, GPT-4o-mini) by injecting task-specific knowledge through low-rank matrices, reducing computational costs and memory requirements [2, 3]. Explore recent advancements in LoRA, such as dynamic adaptation layers, to further optimize model performance for financial tasks.
- **Implementation:** Fine-tune the LLM on a domain-specific corpus of financial texts using LoRA. Apply task-specific prompts and leverage prompt tuning techniques to enhance model relevance and accuracy without needing extensive retraining of the full model.
- **Model details:** The model employs Unsloth pre-quantized 4-bit models for memory and computational efficiency. LoRA updates, with a rank of 16 and scaling factor (LoRA Alpha) of 16, target key transformer components, including attention projections and feed-forward layers. Gradient checkpointing reduces memory usage, enabling longer context handling. Training utilizes Hugging Face's transformers and trl libraries, with a batch size of 2 (gradient accumulation over 4 steps), a learning rate of $2e-4$, and 10 epochs. The adamw_8bit optimizer ensures memory efficiency with a weight decay of 0.01 to prevent overfitting, balancing resource efficiency and performance.

2.4 Hybrid Model Integration

- **Sentiment and Contextual Analysis:** Use the fine-tuned LLM to generate sentiment scores and contextual insights from textual data. In the Capital IQ transcripts and Reddit social media posts, we apply sentiment analysis (using the NLTK VADER API) to textual data, such as titles and content bodies, to calculate standardized atmospheric metrics for trading specific stocks at a given time. DistilBERT was also deployed to determine the relevance of each media post. For numeric data, such as net upvotes and the number of comments, we standardize values by dividing them by their 20-day exponential moving averages to prioritize recent posts while retaining historical trends. We further adjust these values using a decay factor of 1.1^{-t} , assuming a half-life of 7 days (Figs. 2 and 3).
- **Model Architecture:** Develop a hybrid model that incorporates both numerical features and LLM-derived insights. Use an ensemble approach where predictions from both data types are combined to generate a final trading signal.

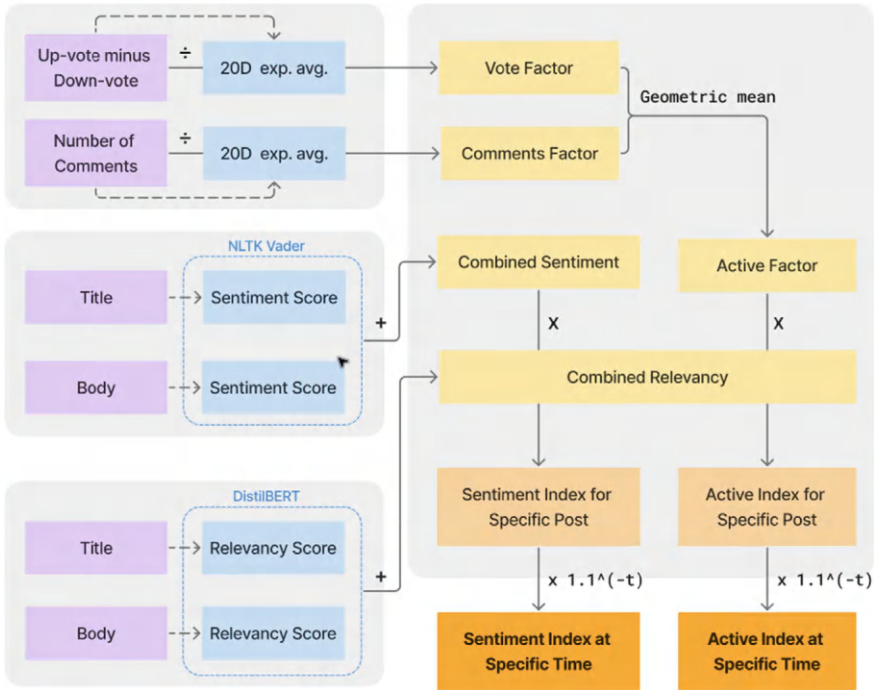


Fig. 2 Sentiment and activity-based relevancy scoring system. *Source* Authors

2.5 Experimental Setup

Triggering events include real-time earnings surprises, earnings call transcripts, and financial news articles related to specific stocks or sectors. These inputs are supported by a vector database containing past news articles and earnings transcripts, annotated with historical price impacts and sentiment metrics. This database retrieves contextually similar historical events to enhance predictions and provide insights into the market impact of real-time events.

A robust backtesting framework is implemented to simulate trades based on the outputs of the hybrid model. A trading signal is generated when the LLM predicts a price movement exceeding a 2% threshold, factoring in a 0.5% transaction cost per trade (1% per buy and sell trading signal) to reflect realistic trading conditions. To ensure the validity of results and avoid forward-looking bias, the framework strictly separates the train and test datasets, ensuring the training data is always prior to the test set in time. This approach enables an accurate assessment of the model's predictive performance and trading strategy efficiency.

Model performance is assessed using both technical and financial metrics to ensure a comprehensive evaluation. For backtesting, standard classification metrics are used, including the confusion matrix, accuracy, precision, recall, and F1-score. Financial

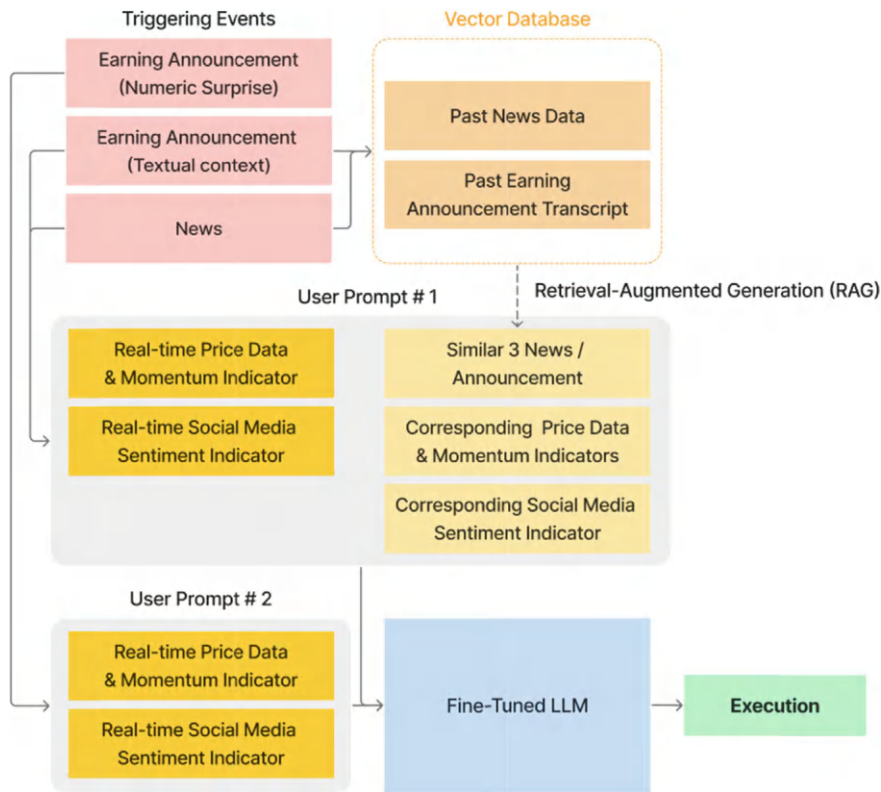


Fig. 3 Retrieval-augmented generation for financial event analysis. *Source* Authors

performance is measured through key metrics such as the Sharpe ratio and Sortino ratio to evaluate risk-adjusted returns, along with profit and loss (P&L) analysis and drawdown statistics to capture portfolio performance and risk exposure.

2.6 Results and Analysis

- **Performance Evaluation:** Compare the hybrid model’s performance against baselines. Highlight improvements in predictive accuracy, risk-adjusted returns, and adaptability to different market conditions.
- **Sensitivity and Robustness Analysis:** Conduct sensitivity analysis to understand how variations in textual data quality or numerical inputs affect the model’s performance. Test the model’s robustness by simulating scenarios with noisy or incomplete data.

3 Results and Findings

3.1 Results and Analysis

From the confusion matrix, it is evident that while the model’s precision is not very high, the trading strategy remains effective due to the 2% threshold for generating signals. Even in cases of false positives, if the magnitude of the stock’s movement exceeds the transaction cost of 1%, the trade can still yield a profit. This highlights the practical utility of the model despite its precision limitations, as it focuses on capturing significant market movements.

Accuracy	0.986	True positives	32	Sharpe ratio	0.623
Precision	0.533	True negatives	3659	Sortino ratio	1.787
Recall	0.571	False positives	28	Cum. profit and loss	+151.16%
F1-score	0.552	False negatives	24	Maximum drawdown	−3.85%

The simulated trading gains may not fully reflect the actual performance of the strategy due to the presence of certain forward-looking biases. In this proof of concept, a 20-min trading window is introduced to ensure that the fine-tuned model captures price changes driven solely by the news and associated metrics. However, implementing a buy/sell strategy that establishes positions 10 min before a triggering event is unrealistic in real-life scenarios. In practice, significant price jumps often result from short-term extreme events, which cannot be anticipated with perfect timing.

In Sect. 3.2, we will conduct a simulation using delayed entering of 1 min after the triggering event (Fig. 4).

3.2 Delayed Execution

In the previous sections, we set the trading window to 10 min before and after the triggering events to demonstrate that the fine-tuned LLM can capture the price impact of such events. However, this approach introduces a degree of forward-looking bias, as in real-world scenarios, trade cannot be executed prior to uncertain triggering events (Fig. 5).

From Fig. 6, we observe that the profit increases as the trading window is extended, likely due to the current limited adoption of LLM-driven trading, which requires human traders time to react.

Additionally, a lower signal generation threshold typically yields higher profits despite reduced precision, as evidenced by the highest return scenario of + 206.87%, which achieved a precision of only 0.785. This is likely because, while the fine-tuned

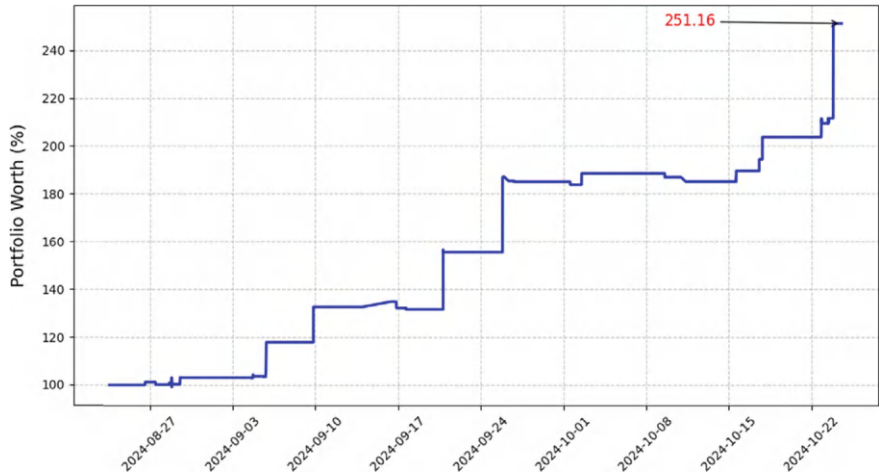


Fig. 4 Change in portfolio worth (%) driven by news and earnings call. *Source* Authors

model may not predict the exact magnitude of price movements, it accurately captures trends, leading to more frequent trade at lower thresholds.

3.3 Cross Comparison of Llama and GPT

Llama 3.2 excels in domain-specific tasks with optimized parameters, while GPT-4o-mini offers broader generalization and cost-efficiency, highlighting task-specific trade-offs [5].

Two models—LlaMA 3.2 3B and GPT-4o mini—were fine-tuned for 1,600 steps each. Figure 7 presents the comparative results.

The results show that even the lowest performing fine-tuned model from OpenAI, GPT-4o-mini, trained with 90% fewer steps, still outperforms the Llama 3.2 3B model in profitability under the scenario where the threshold is 0.5, trades start at + 1 min, and positions close at + 31 min.

Steps / Epochs	Llama 3.2 3B								
	Trade at -10 min Close at +10 min			Trade at +1 min Close at +3 min			Trade at +1 min Close at +6 min		
	10 epochs (16000+ steps)								
Signal Generating Threshold	0.5	1	1.5	0.5	1	1.5	0.5	1	1.5
Accuracy	0.859	0.961	0.978	0.981	0.991	0.995	0.956	0.988	0.991
Precision	0.403	0.478	0.483	0.087	0.083	0.000	0.141	0.238	0.154
Recall	0.435	0.458	0.531	0.047	0.056	0.000	0.113	0.156	0.095
True Positive	192	65	43	2	1	0	11	5	2
True Negative	3056	3568	3654	3149	3184	3198	3309	3425	3441
False Positive	284	71	46	21	11	8	67	16	11
False Negative	249	77	38	41	17	7	86	27	19
Mean of return (per execution)	0.52%	1.49%	2.02%	-0.11%	-0.03%	-0.23%	0.03%	0.01%	0.34%
Sharpe ratio (2 months)	0.116	0.107	0.098	-0.026	-0.004	-0.030	0.005	0.000	0.021
Sharpe ratio (annualized)	0.283	0.263	0.240	-0.063	-0.010	-0.073	0.011	0.001	0.052
Sortino ratio (2 months)	0.094	0.074	0.054	-0.011	-0.002	-0.011	0.002	0.000	0.010
Sortino ratio (annualized)	0.231	0.181	0.132	-0.028	-0.006	-0.026	0.005	0.000	0.024
Maximum Drawdown	-14.45%	-3.05%	-3.05%	-2.38%	-0.95%	-1.15%	-5.78%	-4.16%	-1.28%
Profit and Loss (2 months)	904.51%	576.76%	455.51%	-1.98%	-0.30%	-1.15%	1.99%	0.02%	4.09%

Steps / Epochs	Llama 3.2 3B								
	Trade at +1 min Close at +11 min			Trade at +1 min Close at +16 min			Trade at +1 min Close at +31 min		
	10 epochs (16000+ steps)								
Signal Generating Threshold	0.5	1	1.5	0.5	1	1.5	0.5	1	1.5
Accuracy	0.911	0.981	0.990	0.876	0.965	0.983	0.785	0.938	0.970
Precision	0.216	0.206	0.200	0.275	0.167	0.217	0.288	0.319	0.343
Recall	0.213	0.143	0.115	0.265	0.114	0.102	0.238	0.107	0.119
True Positive	44	7	3	84	10	5	140	22	12
True Negative	3248	3539	3577	3121	3521	3592	2765	3446	3575
False Positive	160	27	12	221	50	18	346	47	23
False Negative	163	42	23	233	78	44	448	184	89
Mean of return (per execution)	0.20%	0.58%	0.88%	0.28%	0.51%	0.86%	0.25%	0.92%	1.44%
Sharpe ratio (2 months)	0.050	0.037	0.028	0.081	0.043	0.030	0.089	0.073	0.061
Sharpe ratio (annualized)	0.122	0.090	0.068	0.197	0.104	0.074	0.219	0.179	0.149
Sortino ratio (2 months)	0.058	0.025	0.020	0.060	0.017	0.006	0.087	0.053	0.035
Sortino ratio (annualized)	0.143	0.061	0.049	0.148	0.041	0.014	0.212	0.130	0.085
Maximum Drawdown	-2.23%	-1.07%	-0.57%	-4.93%	-3.87%	-3.87%	-4.24%	-1.75%	-1.44%
Profit and Loss (2 months)	46.67%	21.45%	13.72%	123.11%	34.15%	19.33%	206.87%	81.90%	59.76%

Fig. 5 Horizontal comparison of different trading timings. Source Authors

3.4 Other Experiments

Pre-trained models (Llama 3.2 3B)

Pre-trained models exhibit limitations in accurately understanding content and adhering to specific instructions. Despite being explicitly tasked with generating a 2-decimal numerical output, these models often produce irrelevant or non-numerical responses, likely due to their tendency to overgeneralize and elaborate on the provided input (Figs. 8 and 9).

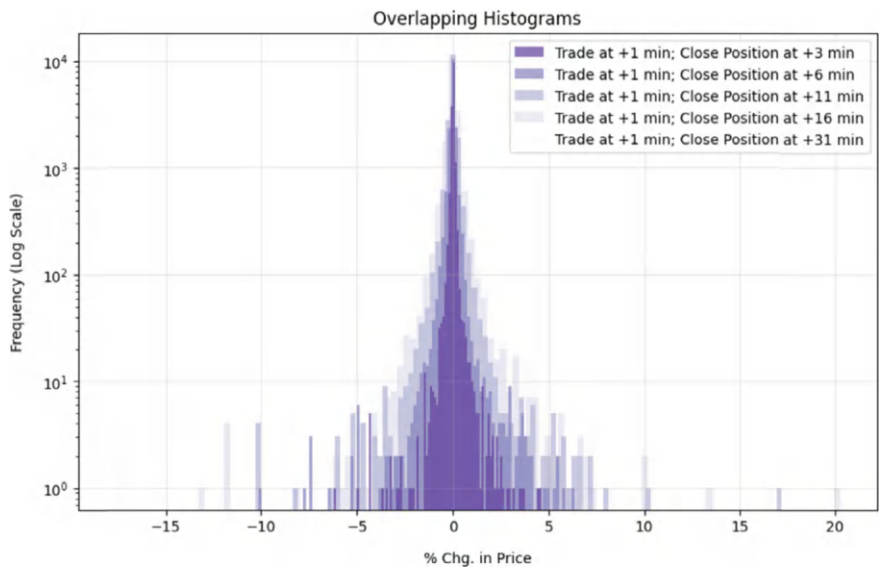


Fig. 6 Frequency distribution of percentage price changes across different trading windows. *Source* Authors

Excluding earnings call transcripts

Relying solely on news as the triggering event reduces accumulated returns, yielding + 126.13% compared to + 151.16% reported in Sect. 3.1.

Accuracy	0.987	True positives	29	Sharpe ratio	0.568
Precision	0.569	True negatives	3666	Sortino ratio	1.188
Recall	0.527	False positives	22	Cum. profit and loss	+126.13%
F1-score	0.547	False negatives	26	Maximum drawdown	−6.83%

3.5 Penetration Effects Driving Price Movement on NVDA

Since the selected companies are correlated with NVDA, we are investigating whether trading NVDA based on news announcements of individual stocks is feasible—specifically, if these announcements significantly impact NVDA’s price. However, our analysis of the frequency of 3% price movements in NVDA induced by these stocks revealed no strong alignment with their correlation coefficients. This

	Llama 3.2 3B				GPT 4o-mini			
	Trade at -10 min Close at +10 min							
Steps / Epochs	1600 steps							
Signal Generating Threshold	0.5	1	1.5	2	0.5	1	1.5	2
Accuracy	0.846	0.917	0.970	0.981	0.843	0.958	0.977	0.983
Precision	0.206	0.124	0.212	0.295	0.208	0.302	0.321	0.286
Recall	0.121	0.215	0.187	0.236	0.131	0.119	0.120	0.109
True Positive	52	29	14	13	56	16	9	6
True Negative	3114	3404	3617	3658	3102	3572	3650	3674
False Positive	201	205	52	31	213	37	19	15
False Negative	377	106	61	42	373	119	66	49
Mean of return (per execution)	0.26%	0.28%	0.65%	0.91%	0.19%	1.02%	1.34%	0.92%
Sharpe ratio (2 months)	0.041	0.042	0.032	0.030	0.036	0.043	0.034	0.030
Sharpe ratio (annualized)	0.101	0.102	0.077	0.073	0.087	0.105	0.083	0.072
Sortino ratio (2 months)	0.028	0.028	0.011	0.008	0.027	0.018	0.010	0.005
Sortino ratio (annualized)	0.068	0.067	0.027	0.021	0.066	0.045	0.025	0.013
Maximum Drawdown	-10.99%	-12.27%	-11.56%	-11.14%	-8.43%	-3.64%	-3.48%	-3.64%
Profit and Loss (2 months)	79.80%	81.15%	49.56%	44.83%	58.08%	65.33%	43.22%	20.63%

	Llama 3.2 3B				GPT 4o-mini			
	Trade at +1 min Close at +31 min							
Steps / Epochs	1600 steps							
Signal Generating Threshold	0.5	1	1.5	2	0.5	1	1.5	2
Accuracy	0.818	0.900	0.963	0.972	0.783	0.916	0.958	0.976
Precision	0.354	0.186	0.219	0.123	0.312	0.198	0.125	0.128
Recall	0.173	0.238	0.139	0.119	0.303	0.165	0.089	0.085
True Positive	102	49	14	7	178	34	9	5
True Negative	2925	3279	3548	3590	2719	3355	3535	3606
False Positive	186	214	50	50	392	138	63	34
False Negative	486	157	87	52	410	172	92	54
Mean of return (per execution)	0.36%	0.37%	0.69%	0.76%	0.26%	0.45%	0.51%	0.56%
Sharpe ratio (2 months)	0.098	0.094	0.073	0.072	0.095	0.076	0.048	0.036
Sharpe ratio (annualized)	0.240	0.231	0.178	0.176	0.232	0.186	0.118	0.089
Sortino ratio (2 months)	0.049	0.045	0.037	0.033	0.080	0.053	0.020	0.010
Sortino ratio (annualized)	0.121	0.109	0.091	0.082	0.197	0.131	0.049	0.025
Maximum Drawdown	-5.92%	-5.92%	-1.34%	-1.34%	-5.00%	-2.30%	-2.30%	-2.67%
Profit and Loss (2 months)	167.62%	154.92%	52.72%	51.41%	291.31%	106.11%	40.43%	22.62%

Fig. 7 Comparative performance of the LLaMA 3.2 3B and GPT-4o mini models after 1,600 fine-tuning steps *Source* Authors

is reasonable, as correlation coefficients measure long-term pairwise movements and do not account for sudden surges in individual stocks. Additionally, if a trader holds significant positions in both the individual stock and NVDA, capital shifts between them could trigger opposing price movements in NVDA (Fig. 10).

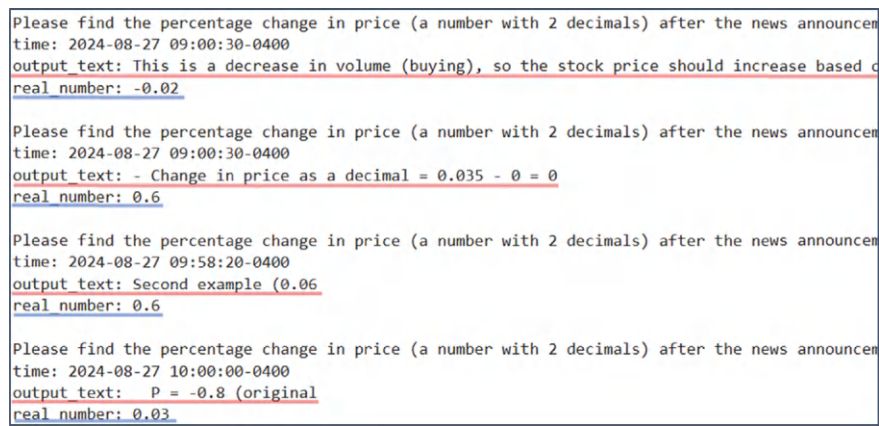


Fig. 8 Output of a pre-trained model compared to actual percentage change. *Source* Authors

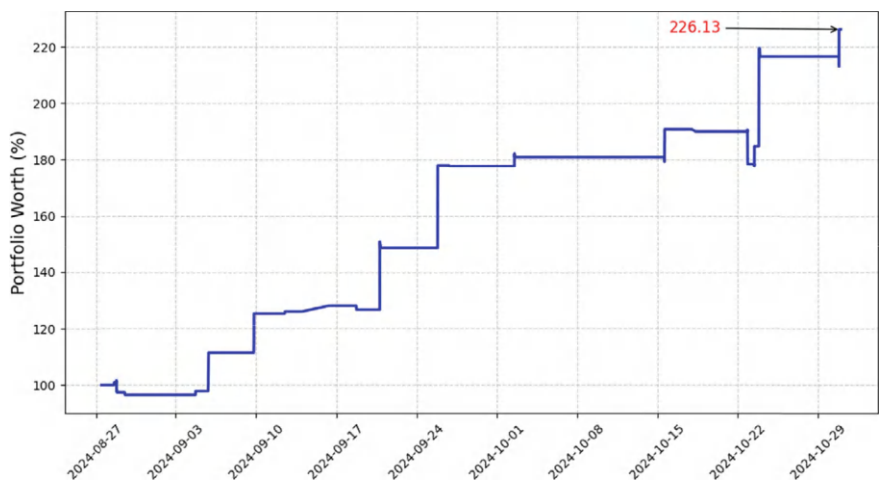


Fig. 9 Change in portfolio worth driven by news-triggered signals only. *Source* Authors

4 Future Improvement

This is a proof of concept, and the data depth and time horizon can be further adjusted. With these modifications, the streamlined approach can be implemented in real-life scenarios.

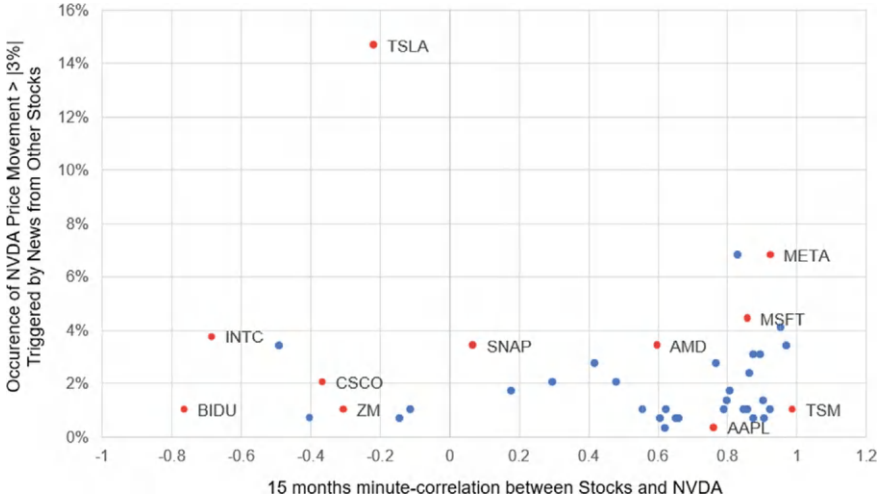


Fig. 10 Correlation coefficient with NVDA does not imply that news from other stocks directly affects NVDA price change in short term. *Source* Authors

4.1 Higher Quality Data

We are currently working with 1.5 years of data but plan to extend the time span and include broader, high-quality news sources for fine-tuning. Future improvements will also incorporate additional trading information as well as covering more social media (e.g., Twitter) to enhance signal accuracy.

4.2 Integration of Additional Quantitative Models

Surprise data can be quantified and incorporated into existing quantitative trading models, such as XGBoost, Random Forests, or neural networks. This integration can create a more comprehensive model and improve predictive accuracy.

4.3 Trading Strategy

We are currently implementing a simple buy or sell strategy within a certain time span. In the future, we can introduce the time span as well as the trading signal generating threshold as variables and incorporate derivatives, such as option prices, to enhance the model’s capability for generating real-time trading signals.

5 Conclusion

In this study, we explored the potential of integrating financial news and earnings transcripts into quantitative trading models using fine-tuned language models. By implementing a robust backtesting framework and incorporating key evaluation metrics, we assessed the effectiveness of our approach in generating accurate trading signals. While limitations such as forward-looking bias and data constraints remain, our framework demonstrates the feasibility of leveraging hybrid models to predict market movements within short time spans. Future enhancements, including broader datasets, advanced technical indicators, and the inclusion of derivatives, can further improve the model's accuracy and responsiveness, making it a valuable tool for real-time trading strategies.

References

1. Lewis P et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst* 33:9459–9474
2. Hu EJ et al (2021) LoRA: low-rank adaptation of large language models. arXiv preprint [arXiv:2106.09685](https://arxiv.org/abs/2106.09685)
3. Wang D, Kim D, Jin B, Zhao X, Fu T, Yang S, Liu X-Y (2024) FinLoRA: finetuning quantized financial large language models using low-rank adaptation. arXiv preprint [arXiv:2412.11378](https://arxiv.org/abs/2412.11378)
4. OpenAI (2024) New embedding models and API updates. <https://openai.com/index/new-embedding-models-and-api-updates/>
5. OpenAI (2024) GPT-4 technical report. <https://openai.com/research>

Foundations of LLMs and Financial Applications



Yoonseo Chung^{ID}, Jeonghyun Kim^{ID}, MiYeon Kim^{ID}, Minsuh Joo^{ID},
and Hyunsoo Cho^{ID}

Abstract The integration of Large Language Models (LLMs) into the financial industry represents a transformative advancement in artificial intelligence, addressing the complexities of data-driven finance. This chapter explores how cutting-edge LLMs can be aligned with financial practices to enhance efficiency and foster innovation in financial services. The discussion begins with an overview of LLM, including their architecture, training processes, and the datasets they leverage. It then examines finance-specific adaptations, such as FinBERT and BloombergGPT, which are tailored to address domain-specific challenges. The chapter also addresses key challenges in applying LLMs to the financial domain, such as real-time data integration, and evaluates potential solutions, including retrieval-augmented generation (RAG). By analyzing these innovations and challenges, the chapter envisions a future where LLMs redefine the landscape of financial technology.

Keywords Large language model (LLM) · Artificial intelligence (AI) · Data-centric finance · Financial data optimization · Automation in finance

Y. Chung · J. Kim · M. Kim · M. Joo · H. Cho (✉)
Department of AI, Ewha Womans University, Seoul, Korea
e-mail: chohyunsoo@ewha.ac.kr

Y. Chung
e-mail: yyss22@ewhain.net

J. Kim
e-mail: wjdgus630@ewha.ac.kr

M. Kim
e-mail: kmy8228@gmail.com

M. Joo
e-mail: judyjoo21@ewha.ac.kr

1 Introduction

The rise of large language models (LLMs) has brought about a new era in artificial intelligence, transforming how we address complex challenges across various domains [1]. With its intricate decision-making processes, vast amounts of heterogeneous data, and high-stakes outcomes, the financial industry is especially ready for transformation through LLM-based solutions.

Finance has always been a data-centric domain. Professionals navigate through structured datasets like stock prices and economic indicators, as well as unstructured sources such as earnings reports, regulatory texts, and customer inquiries. This information's sheer scale and diversity make extracting meaningful insights an arduous task, traditionally reliant on human expertise and specialized tools. Imagine financial data as an orchestra with various instruments—text, numbers, charts, audio—each needing special attention from expert “musicians.” Historically, integrating these varied data forms into cohesive insights has been labor intensive and disjointed. LLMs can act as the conductor of this orchestra, harmonizing the entire ensemble into a cohesive performance and producing insights that were previously difficult to obtain.

Advancements in AI, exponential growth in data availability, and increasing demand for actionable insights have created an ideal environment for LLM adoption in finance. As companies face mounting regulations, unpredictable markets, and rising customer expectations, LLMs' ability to adapt, learn, and deliver value has become more critical than ever. For instance, in regulatory compliance—a constant challenge for financial institutions—regulations are often conveyed in lengthy, complex legal texts, necessitating hours of meticulous review by compliance teams. LLMs can parse and summarize these texts, flagging critical elements and inconsistencies, thereby accelerating human workflows and reducing the risk of oversight. Similarly, in the realm of investment strategies, LLMs assist by synthesizing market sentiment from vast datasets, identifying trends, and even generating actionable investment theses [2]. These examples underscore the transformational potential of LLMs in the financial sector. They automate repetitive tasks and augment human decision-making, bridging gaps in efficiency and capability. The confluence of technological advancements and industry needs positions LLMs as pivotal tools in reshaping the financial landscape.

This chapter invites readers to explore the synergy between finance and LLMs through detailed analyses and practical examples. We delve into LLMs' architecture, training methodologies, and data optimization techniques, focusing on those specialized for financial applications. By breaking down LLMs from multiple perspectives—including their training, architecture, datasets, and size—we aim to illuminate their working mechanisms and transformative potential in finance. Specifically, we begin by discussing the ongoing digital transformation within finance, the types of financial data that LLMs handle, and the specific roles LLMs play in the industry. We then explore LLMs specialized for financial usage, detailing their construction, fine-tuning methodologies and data optimization techniques. Finally, we examine practical applications of financial LLMs, such as retrieval-augmented gen-

eration (RAG) [3] for real-time insights and multimodal integrations that combine text with other financial data formats like time series and multimedia.

This chapter aims to foster innovation and redefine the future of financial services by bridging the gap between cutting-edge AI technology and financial practice. We invite you to engage with the material, consider the possibilities, and envision how LLMs can be leveraged to drive efficiency, insight, and competitive advantage in the financial industry.

2 The Financial Sector and LLMs

This chapter explores the increasing role of LLMs within the rapidly evolving financial sector. As the industry undergoes a sweeping digital transformation, financial institutions leverage unprecedented volumes of data to enhance decision-making, improve customer experiences, and identify new opportunities. Against this backdrop, LLMs have emerged as powerful tools capable of automating traditionally human-driven analytical tasks, thus reshaping operational paradigms across mobile banking, online payments, and digital assets, leading to substantial improvements in customer experience and operational efficiency [4, 5]. This chapter examines the digital shift in the financial industry, then outlines the nature of financial data and the tasks it enables, and finally discusses how LLMs are being integrated to tackle these tasks with growing sophistication.

2.1 Digital Transformation of the Financial Industry

The financial sector has historically been at the forefront of innovation, as its outcomes are closely tied to profitability, making it highly sensitive to even subtle changes in economic conditions. It has consistently introduced new products, services, and regulatory frameworks in response to shifting economic landscapes. Previously, financial services relied heavily on paper-based documentation, manual recordkeeping, and face-to-face client consultations. Over the last few decades, however, the industry has steadily adopted digital technologies—ranging from online banking portals and electronic trading platforms to real-time risk management systems—profoundly [6] changing how these services are delivered and consumed.

This digital transformation has accelerated in recent years, propelled by widespread high-speed Internet access, the proliferation of mobile devices, and the diversification of financial offerings. Complex computational models, algorithmic trading tools, and advanced risk assessment techniques now form integral parts of everyday financial operations. As a result, the financial ecosystem generates enormous volumes of structured and unstructured data, including transactional records, client communications, regulatory filings, social media commentary, and detailed market analytics.

Until recently, making effective use of this varied and complex information for strategic decision-making had almost exclusively required extensive domain expertise in areas such as economics, corporate finance, and market analysis. Without advanced AI solutions, this expertise-driven approach was the primary method for extracting meaningful insights. As the industry continues to evolve, however, emerging AI tools, including LLMs, are poised to enhance and streamline these processes, amplifying the capabilities of financial experts and transforming how data-driven insights are generated.

2.2 Types and Characteristics of Financial Data and Tasks

As the financial industry's digital landscape expands, so does the variety and complexity of the data it generates. Understanding the types of information available and the tasks they enable is critical for recognizing how LLMs can be applied effectively. Historically, such data-driven activities relied on the skills and judgment of finance professionals, but with LLMs becoming more accessible and sophisticated, these processes are increasingly ripe for automation and augmentation.

Financial Data

Financial data today encompasses a broad spectrum, ranging from cleanly structured numerical feeds to more complex, unstructured textual sources. Common categories include [7]

1. **Market Data:** Prices, trading volumes, volatility measures, and yield curves are central to activities like trading, market making, and portfolio management. These streams are often updated at high frequencies, requiring robust real-time ingestion and analysis systems.
2. **Fundamental Data:** Corporate filings such as annual reports, earnings statements, and regulatory disclosures provide insights into a company's performance, capital structure, and risk profile. This data type underpins fundamental analysis, credit risk assessment, and valuations.
3. **Transactional and Behavioral Data:** Payment records, loan applications, credit card usage patterns, and insurance claims offer detailed perspectives on consumer behavior and financial health. Such data informs credit scoring, fraud detection, customer segmentation, and targeted marketing.
4. **Sentiment and News Data:** Unstructured sources—including news articles, analyst reports, social media posts, and central bank announcements—provide context for market events and investor sentiment. While notoriously challenging to process at scale, these sources can influence trading strategies, guide portfolio rebalancing, and help institutions anticipate regulatory changes.

Financial Tasks

The sheer volume and diversity of financial data have given rise to numerous tasks [8] aimed at extracting actionable insights. These tasks traditionally required domain specialists equipped with advanced quantitative skills and deep sector knowledge:

1. **Risk Management and Compliance:** Assessing creditworthiness, evaluating market and liquidity risks, and ensuring adherence to regulatory standards demand a nuanced understanding of complex data. Regulatory changes or newly disclosed information often require manual review by experts, who interpret and apply rules to maintain compliance.
2. **Portfolio Construction and Investment Analysis:** Analysts must integrate market data, economic indicators, and corporate fundamentals to identify promising investment opportunities. They then construct portfolios aligned with specific objectives, risk tolerances, and market conditions, adjusting allocations as circumstances evolve.
3. **Trading and Market Forecasting:** High-frequency traders and long-term investors rely on accurate market forecasts. Complex models draw from historical price patterns, macroeconomic indicators, and real-time sentiment data. Before advanced AI tools, these forecasts typically rested on human intuition supplemented by statistical models.
4. **Customer Engagement and Personalization:** Financial institutions strive to offer personalized advice, products, and services. This customization relies on analyzing individual transaction histories, credit behaviors, and communications, which is traditionally a labor-intensive process driven by human relationship managers and marketing teams.
5. **Fraud Detection and Cybersecurity:** Protecting consumers and institutions from fraudulent activities requires constant vigilance. Analysts or dedicated fraud teams have historically monitored transaction flows, user activities, and unusual behavioral patterns, identifying anomalies through extensive domain know-how.

2.3 *LLMs and Finance*

Until the advent of robust AI solutions, carrying out these tasks depended heavily on human expertise, lengthy documentation reviews, and manual cross-referencing of multiple data sources. Finance professionals often combine their understanding of economic theory and market structure with firm-specific knowledge and industry best practices [9]. This approach, while effective, is time-consuming, labor intensive, and prone to human error, prompting an ongoing search for methods to automate, scale, and refine these essential activities.

LLMs have entered this landscape as powerful tools capable of processing both structured and unstructured information at unprecedented scales. By interpreting text, summarizing lengthy documents, and extracting key insights, LLMs offer new avenues for efficiency and accuracy in tasks that were once the sole domain of human experts. In risk management, for example, LLMs can swiftly review thousands of

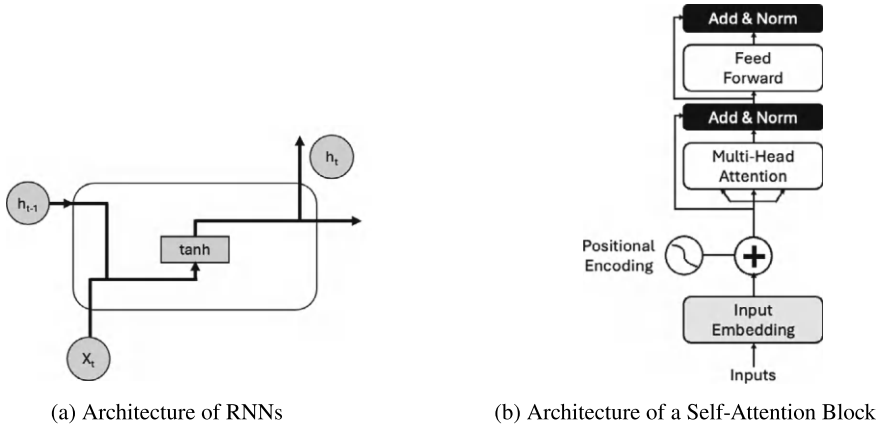


Fig. 1 Diagram illustrating the structure of an RNN (a) and the self-attention block of a Transformer (b) [11, 12]

regulatory filings, flagging relevant changes for compliance officers to consider. These models can synthesize market reports, earnings calls, and analyst opinions for investment analysis, providing a quick overview of complex scenarios. In customer engagement, LLMs enable more sophisticated chatbots and recommendation systems that help clients navigate products and services with intuitive, conversational interactions [10]. Similarly, LLM-driven fraud detection systems can detect subtle linguistic cues and anomalies in transaction logs, enhancing early warning capabilities.

By lowering the cost and time associated with data interpretation, LLMs do more than just replicate human-level analysis; they augment it. Financial experts are freed from routine tasks, allowing them to focus on strategic decision-making, critical interpretation, and scenario planning. As LLMs continue to evolve, they promise to make financial institutions more agile, informed, and customer-centric, redefining future financial professionals' skill sets and value propositions (Fig. 1).

3 Brief Introduction to LLM

LLMs are increasingly being adopted in the financial industry, where they assist with tasks like analyzing market sentiment, automating report generation, and enhancing customer interactions [13]. Their ability to process large volumes of textual data with accuracy and contextual understanding makes them valuable tools for navigating the complexities of financial language. While their success is often attributed to breakthroughs in architectures like the Transformer [14], the training methodologies also play a critical role in their performance. Key processes such as pre-training on large datasets and post-training on domain-specific data ensure these models can

handle the complexities of financial language. In this chapter, we will explore the general architecture of LLM (i.e., Transformer), the development of finance-specific Pre-trained Language Models (PLMs), and the importance of post-training methods.

3.1 Transformer

The Transformer architecture, introduced in 2017, revolutionized Natural Language Processing (NLP) by addressing limitations in previous models (e.g., RNNs [15] and LSTMs [16]). By processing all words in a text simultaneously rather than sequentially, the Transformer is faster and better suited to analyzing complex documents.

What Makes the Transformer Special?

Several key features set the Transformer apart from earlier models, making it particularly powerful:

1. **Self-Attention (Attention)**

Self-attention allows the Transformer to understand how different words in a sentence relate to each other, regardless of their position. For example, in “The weather today is sunny, and it’s perfect for a picnic,” the model understands that “it” refers to “the weather.” This capability ensures deeper comprehension of text by focusing on relevant relationships.

2. **Multi-Head Attention**

With multi-head attention, the model can analyze text from multiple perspectives simultaneously. This feature enables it to recognize both fine-grained details, like identifying named entities, and broader patterns, such as understanding the tone of a conversation.

3. **Positional Encoding**

Since the Transformer processes words in parallel rather than sequentially, positional encoding retains the order of words. It is crucial for understanding sentences where meaning depends on word sequence, such as “The cat chased the mouse” versus “The mouse chased the cat.”

These features allow the Transformer to excel in core NLP tasks like machine translation, text summarization, and sentiment analysis, offering unprecedented accuracy and efficiency (Fig. 2).

How These Advantages Translate to Finance?

The strengths of the Transformer in NLP also benefit financial applications by enabling better analysis of complex language and large datasets:

1. **Understanding Context:** Self-attention helps models identify relationships in financial texts, such as linking earnings results to market trends.

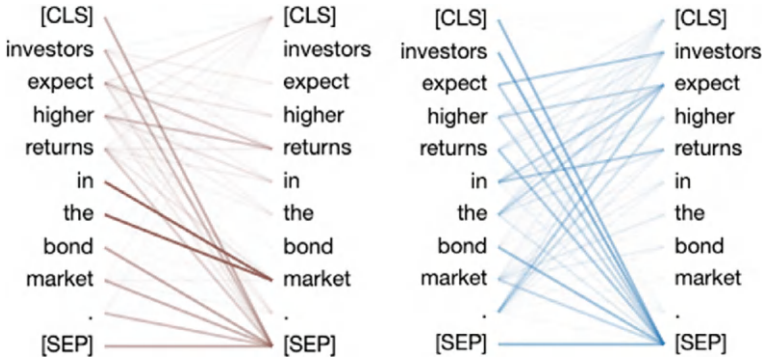


Fig. 2 Visualization of self-attention and multi-head attention results [14, 17]

2. **Identifying Key Patterns:** Multi-head attention allows models to detect trends or anomalies, such as correlating news sentiment with stock movements.
3. **Preserving Order:** Positional encoding ensures accurate interpretation of sequential data, like transaction histories or market updates.

While the Transformer’s design was not created specifically for finance, its general strengths in processing complex language and context make it a powerful tool. Models like FinBERT [18], which adapt the Transformer for financial tasks, have demonstrated its ability to effectively tackle the unique challenges of financial language.

3.2 Finance PLM (Pre-trained Language Model)

The Transformer architecture has proven transformative for NLP tasks, enabling models to capture complex patterns and contextual relationships in textual data. Finance-specific Pre-trained Language Models (PLMs) leverage the same architecture (i.e., Transformer), adapting it to meet the challenges of finance. By fine-tuning the Transformer’s capabilities, PLMs can address the nuanced demands of financial NLP tasks, including sentiment analysis, entity recognition, and text classification. These PLMs share an identical architecture and pre-training methodology with LLMs, differing primarily in scale. Despite their smaller size, PLMs have played a critical role in bridging the gap between general-purpose NLP tools and the specialized needs of the financial industry.

How Do We Pre-train LM?

Pre-training is a foundational step in building a language model, enabling it to learn patterns, semantics, and relationships from large volumes of text. The core idea behind pre-training is rooted in “distributed semantics,” which assumes that the meaning of a word can be inferred from its surrounding words. This principle allows

language models to predict missing or related words and develop a deep understanding of context. One of the most effective and widely used pre-training tasks is Masked Language Modeling (MLM), as introduced in models like BERT [19]. MLM involves masking random words in sentences and training the model to predict the masked words based on their context. For example, in the sentence “The company’s [MASK] rose by 15%,” the model learns to predict “profit.” MLM helps the model capture bidirectional context, a critical feature for tasks requiring nuanced understanding.

Models are typically trained on large text corpora. For general-purpose models, this includes diverse datasets such as books and encyclopedias. Finance-specific PLMs, like FinBERT, focus on financial texts, such as news articles, earnings reports, and regulatory documents, to adapt the model to domain-specific language and terminology. By leveraging BERT’s foundational methods, Finance PLMs overcome the challenges of domain-specific language, such as jargon and complex structures. The next section will explore Finance LLMs, building on these foundations to scale up capabilities for broader and more complex financial applications.

FinBERT-series

FinBERT is a model that applies these characteristics of BERT to the financial domain. Both **FinBERT-19** and **FinBERT-20** are specialized in finance sentiment analysis but differ in their pre-training approaches. FinBERT-19 initializes its weight with a BERT model pre-trained on a general corpus, followed by further pre-training on a financial-domain corpus. The model is then fine-tuned on a financial-domain-specific corpus to specialize in financial sentiment analysis [18]. In contrast, FinBERT-20 is pre-trained on a financial communication corpus from scratch and then undergoes fine-tuning. FinBERT outperforms typical BERT in financial sentiment analysis tasks [20]. FinBERT-21, on the other hand, is designed for financial text mining. It adopts a mixed domain pre-training approach, which sits between continuous pre-training and financial-domain-specific pre-training from scratch. The model gains expertise in financial knowledge by utilizing both general-domain- and financial-domain-specific corpora during pre-training, while retaining a broad understanding of general semantic information [21].

FLANG (Financial LANGUAGE)

Google’s model ELECTRA adopts Replaced Token Detection (RTD) [22] as its training objective instead of the typical Masked Language Modeling (MLM). RTD involves replacing certain tokens with plausible tokens generated by a generator (similar to masking) and then training a discriminator to predict whether a given token has been replaced or not. This process creates a competitive learning between the generator and the discriminator. FLANG applies the RTD training objective to the financial domain. FLANG trains the discriminator to predict whether these tokens have been replaced by using financial keywords and phrases as the replacement tokens. This process enables the model to learn finance-specific word representations. The study about FLANG also introduced five financial NLP benchmark tasks collectively named Financial Language Understanding Evaluation (FLUE). FLUE includes tasks such as Sentiment Analysis, Headline Text Classification, Named Entity Recogni-

tion, Structure Boundary Detection, and Question Answering. FLANG has demonstrated its utility by outperforming typical BERT and ELECTRA models on these benchmarks [23].

So far, we have explored Finance PLMs before the advent of Finance LLMs. Although smaller in scale compared to LLMs, Finance PLMs have leveraged the characteristics of NLP domain models to capture the trends and complexities of the vast and intricate financial domain, contributing to tasks such as financial sentiment analysis and financial text mining. Section 4 will delve into Finance LLMs, discussing their necessity and utility in greater detail.

3.3 Post-training

Post-training is the process of refining a language model after its initial training to make it better suited for specific tasks or more aligned with user needs. Imagine a language model as a well-educated person with a broad understanding of many topics. Post-training is like giving them extra training in a particular profession, such as finance, to ensure they excel in their job. This step is critical in turning a general-purpose model into a highly effective tool for specialized domains like financial analysis. In simple terms, post-training fine-tunes a pre-trained model by teaching it specific skills. The initial pre-training phase equips the model with general language understanding, but post-training ensures it becomes an expert in a particular field. This is achieved by showing the model examples of how to solve specific problems or follow particular instructions.

Supervised Fine-Tuning (SFT) is a technique that fine-tunes a model in a supervised fashion using input and ground-truth answer pairs for a specific task. **Instruction Fine-Tuning (IFT)** is similar to SFT but involves instructions in the dataset. For instance, consider the task of financial sentiment analysis. Labeled data structured in the format “input: [statement about the financial situation], label: [positive/neutral/negative]” is fed into the model and makes the model predict the label for a given input. In the case of IFT, instructions are given which can directly accelerate the performance and controllability of the response. For example:

- (In the case of IFT) Instruction: “*What is the sentiment of this news? Please choose an answer from positive/neutral/negative.*”
- Input: “*The company reported a significant increase in its quarterly profits.*” Label: *positive.*
- Input: “*The company’s revenue fell short of expectations, leading to a sharp decline in its stock price.*” Label: *negative.*

The model is fine-tuned to predict the corresponding label based on the given instruction and input.

Reinforcement Learning with Human Feedback (RLHF) [24, 25] trains a model to generate responses aligned with human preferences by utilizing reinforce-

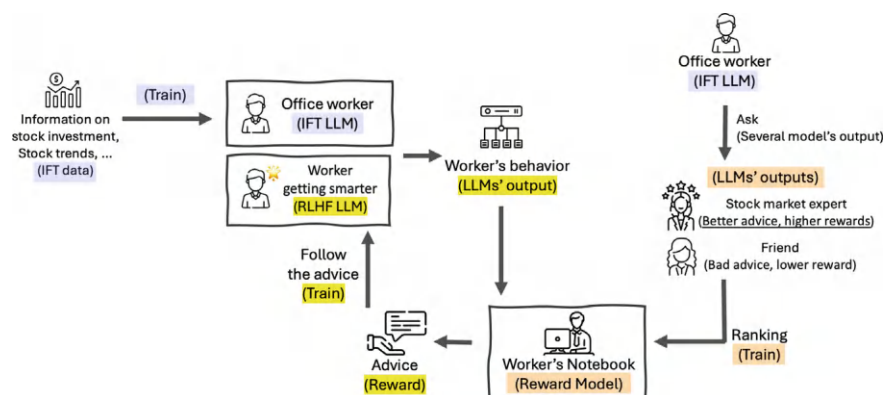


Fig. 3 Illustration of RLHF process

ment learning and human feedback. Unlike typical machine learning or deep learning algorithms, reinforcement learning involves an agent and an environment interacting during the learning process. The main goal of reinforcement learning is to enable the agent to learn a policy—a set of actions—that maximizes rewards in a given environment (Fig. 3).

For example, imagine an office worker starting to invest in stocks. The worker might have various questions about how to buy and sell stocks or how to invest successfully. To find answers, he consulted both a stock market expert and a friend who had been investing for just a year. Naturally, the expert's answers, which are more professional and effective regarding market trends and strategies, would take precedence over the friend's advice. The worker would likely highlight the expert's advice in his investment notebook, prioritizing it over the friend's.

From an RLHF perspective, the worker's notebook acts as a trained reward model that assigns higher rewards to the expert's responses compared to the friend's. When the worker starts buying and selling stocks, he would follow the expert's advice from the notebook, learning how to act can maximize the profit in various situations. Eventually, he could achieve significant profits. Here, let's consider the worker as a language model and worker's behavior as the language model's output. This process mirrors how a reward model assigns rewards to language model's outputs. If the reward is low, the model avoids those outputs, and if the reward is high, it generates similar outputs by adjusting its parameters accordingly. RLHF trains a model to align with human preferences by learning what responses maximize rewards in various financial situations. It ensures the model's outputs align with real human preferences and market dynamics as closely as possible.

To summarize, these post-training methods are valuable tools for significantly improving a model's alignment and performance within specific domains.

3.4 Emergent Abilities—The Secret Behind Scaling

Many people might have wondered, seeing the rise of recent Large Language Models (LLMs), why the size of these models continues to grow. The simple answer is that as the size of a model increases, its ability to understand and generate language improves significantly. In other words, as language models grow larger, they do more than just incrementally enhance in performance—they begin to exhibit emergent abilities [1] that were not present in smaller models. These abilities represent surprising skills that arise once the model surpasses a certain size threshold, transforming LLMs into invaluable tools across various domains.

One domain where the transformative power of LLMs is becoming increasingly evident is finance. In the financial sector, Natural Language Processing (NLP) techniques are actively being applied to tasks such as sentiment analysis, question answering, and stock market prediction. A representative example is BloombergGPT [26]. BloombergGPT, a type of LLM trained with 50 billion parameters, combines 345 billion tokens of general-domain data and 363 billion tokens of financial-domain data. Compared to traditional Language Models (LMs), it has achieved outstanding performance in financial data analysis and NLP tasks.

In addition, models such as FinMA [27], InvestLM [28], and FinGPT [29] have demonstrated excellent performance in the financial domain by leveraging LLM-based models that exhibit emergent abilities.

What Are Emergent Abilities?

Emergent abilities are capabilities that develop spontaneously in larger models, even though these models were not explicitly trained for those tasks. According to a study by Brown et al. [1], they are defined as the model's ability to perform tasks for which it was not explicitly trained. While such abilities are not observed in smaller models, they suddenly emerge in larger models.

This phenomenon can be compared to a person suddenly discovering a talent after mastering the basics of a skill. For instance, a larger language model may unexpectedly gain the ability to perform tasks like logical reasoning or following complex instructions—capabilities that smaller models cannot achieve effectively. This demonstrates how scaling up model size and training data leads to the spontaneous development of advanced skills, further enhancing the versatility and performance of large language models.

In-context Learning (ICL)

In-Context Learning (ICL) allows the model to learn and adapt to a task by observing examples provided within the prompt without requiring additional fine-tuning. For instance, if a user gives a few examples of how to classify financial news as “positive” or “negative,” the model can quickly generalize and apply this classification to new data within the same interaction.

The central concept of ICL is to enable learning by drawing analogies. ICL utilizes a prompt context consisting of a few examples written in natural language templates. Subsequently, the question is combined with the prompt context to form

the input, which is then fed into the language model to generate predictions. Unlike supervised learning, this process does not involve updating the model's parameters; instead, the model identifies patterns from the given examples and predicts the results accordingly.

ICL provides an intuitive and interpretable interface between humans and language models by leveraging example-based learning [1]. It has gained attention for effectively incorporating human knowledge into models simply by modifying prompts and templates [30, 31]. Additionally, ICL operates like human analogy-based learning processes [32] and functions as a training-free framework that adapts to new tasks without requiring additional training stages. This approach reduces computational costs and allows easy application to large-scale, real-world tasks [33].

In complex tasks such as Relation Extraction (RE) [34], ICL has demonstrated performance comparable to, or even better than, traditional supervised learning models with only a few examples. For instance, in the financial domain, ICL-based RE can accurately extract and classify relationships between entities from unstructured data, such as news articles, corporate earnings reports, and regulatory filings. This approach is highly efficient as it allows the rapid integration of domain-specific knowledge while reducing the computational cost of processing large-scale data. Moreover, providing more relevant examples in the prompt context can further improve the model's performance and help mitigate unnecessary hallucinations [35] during predictions. Therefore, in specialized fields such as the financial domain, ICL offers a flexible and intuitive solution with the potential for practical application in various real-world tasks.

Prompt Engineering

This refers to the model's ability to respond effectively to tailored instructions or queries. By crafting a well-structured prompt, users can guide the model to generate concrete outputs. For example, asking, "Summarize this earnings report in one paragraph for investors", can yield a concise and relevant summary.

This capability highlights the importance of prompt engineering. By utilizing task-specific instructions, known as prompts, prompt engineering enhances model efficiency without modifying its core parameters, similar to ICL. Instead of updating the model parameters, prompts guide the model's behavior solely based on the given input, enabling the flexible integration of pre-trained models into various downstream tasks. The following methodologies are representative examples:

1. **Zero-Shot Prompting:** Zero-shot prompting is a technique that represents a paradigm shift in utilizing LLMs, as it does not require extensive training data [36]. Instead, it uses prompts that only describe the task, guiding the model to perform new tasks. This approach provides no labeled input–output examples, but the model generates predictions for the new task by leveraging its pre-trained knowledge.
2. **Few-Shot Prompting:** Few-shot prompting involves providing a small number of input–output examples to help the model understand and perform the task, unlike zero-shot prompting, which includes no examples [1]. Even a few high-quality examples can significantly improve the model's performance on complex

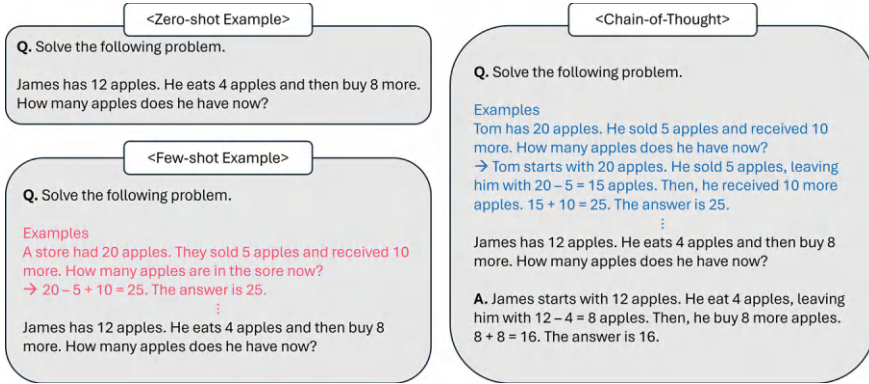


Fig. 4 Various prompting techniques: zero-shot prompting (without sample), in-context learning (with samples), and CoT prompting (with samples and corresponding reasoning) [1, 31]

tasks; however, including examples incurs additional token costs. Moreover, the selection and composition of examples in the prompt can significantly influence the model's behavior, and the model may exhibit biases, such as a preference for frequently occurring words.

3. **Chain-of-Thought (CoT) Prompting:** CoT prompting is a technique that guides large language models to follow a consistent and step-by-step process for solving complex reasoning problems [31]. This approach presents a logical reasoning chain within the prompt, helping the model break the problem into intermediate steps. For instance, when solving multi-step math problems, the reasoning process is divided into sequential steps to derive the final answer. This method effectively enables the model to generate more logical and structured responses.

Through these methods, prompt engineering plays a key role in various tasks such as financial data analysis, risk management, report summarization, and sentiment analysis. For example, it can be applied to automatically summarize structured financial data into reports or analyze customer reviews and news articles to evaluate market sentiment. Additionally, in risk management, carefully designed prompts enable the swift and accurate identification and assessment of risk factors for specific companies or industries (Fig. 4).

4 LLMs in Financial Analysis

4.1 Necessity for Building Financial LLMs

General LLMs are not optimized for finance, so retraining is required to incorporate the economic knowledge that these models lack. Even in the case of ChatGPT [37], it does provide plausible answers to finance questions but cannot provide answers at

the level of an expert. A general LLM can be likened to someone who has “studied finance” but lacks the ability to offer specialized answers or analysis. They have a basic finance understanding but cannot provide expert insights. However, if this person were to study finance diligently and for an extended period, they could become an expert, and we would then be able to trust the financial knowledge they possess. Similarly, if an LLM with a basic understanding of finance is trained with a large amount of economic knowledge, a highly reliable Financial LLM can be created. These Language Models in Finance are referred to as FinLLMs [38].

4.2 Major Finance LLMs

FinLLMs can be broadly divided into two categories [38]:

1. **Mixed-Domain LLM with Prompt Engineering:** This refers to cases where financial knowledge is imparted to the model by leveraging financial corpora during the pre-training phase.
2. **Instruction Fine-Tuned LLM with Prompt Engineering:** This involves imparting financial knowledge to the model through fine-tuning with instruction data.

An example of the former is BloombergGPT [26], while an example of the latter is FinGPT [29].

Instruction fine-tuning, as explained in Sect. 3.3, is a technique that became possible as language models grew larger, and so is prompt engineering. Prompt engineering refers to strategies for adjusting prompts to improve the accuracy of a model’s responses. One example is the Chain-of-Thought (CoT) [31] method, which allows models to break down multi-step problems into intermediate steps, allocating additional computation to problems that require more reasoning.

Using such tuning techniques, which can be applied to large models, FinLLMs optimized for the finance domain are created.

BloombergGPT

BloombergGPT [26] is a FinLLM that uses BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) [39] as its backbone model. It is pre-trained in financial corpora, particularly a proprietary dataset called FinPile, which is created from internal documents selected by Bloomberg analysts and external documents. This high-quality dataset enables BloombergGPT to achieve superior performance and handle challenging and specific tasks.

BloombergGPT excels in tasks such as the generation of Bloomberg Query Language (BQL), suggesting news headlines, and answering financial questions. The generation of BQL involves converting natural language queries into BQL, a language designed for querying and analyzing data on the Bloomberg platform. For instance, a query like “*Tell me the last stock price of Apple*” is translated into a BQL command like

```
SELECT PX_LAST FROM AAPL US Equity WHERE FIELDS = 'LAST PRICE'
```

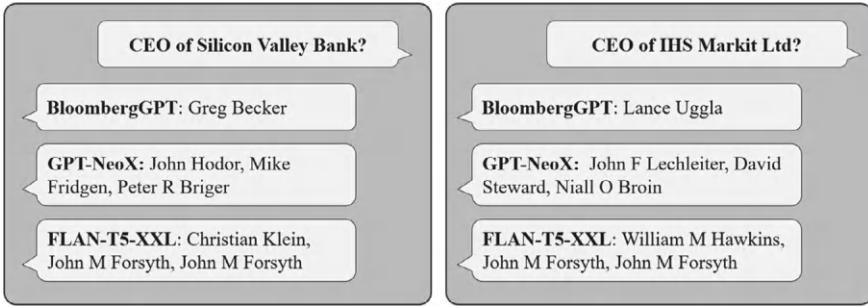


Fig. 5 Testing the abilities of various LLMs, including BloombergGPT [26]

allowing users to retrieve Apple’s latest stock price (PX_LAST) accurately and efficiently. Stock price fluctuations can also be easily obtained by converting them into BQL language, which can complement LLM’s weakness in mathematical reasoning. Another noteworthy task is Financial Question Answering. BloombergGPT, trained on vast financial-domain-specific data, delivers highly accurate answers to finance-related queries. For example, it performs exceptionally well in identifying a company’s CEO. In a comparison presented in Fig. 5, BloombergGPT correctly identified CEOs, whereas GPT-NeoX failed, and FLAN-T5-XXL consistently produced unrelated results by ignoring the given company in the query. This highlights BloombergGPT’s ability to provide more accurate responses than general-purpose LLMs due to its extensive training in high-quality financial corpora.

However, BloombergGPT has some limitations. With a model size of 50 billion parameters, the computational cost for training is extremely high. Additionally, restricted access to its dataset imposes barriers to further research and development, presenting a significant challenge for broader model advancements.

FinGPT

FinGPT [29] is an open-sourced and data-centric framework developed to address the limitations of BloombergGPT. Designed to integrate with various open-source LLMs of different sizes, FinGPT emphasizes data democratization in its research approach. The framework includes a Data Source Layer and a Data Engineering Layer in its pipeline (Fig. 6), ensuring that high-quality data is collected and curated for model training. This focus is essential because financial data is highly time-sensitive and exhibits a low Signal-to-Noise Ratio (SNR). For such data, filtering out irrelevant or noisy information is critical to selecting only the most valuable insights for training. FinGPT overcomes these challenges through techniques like prompt engineering and feature extraction, as illustrated in its architectural pipeline (Fig. 6).

As of November 2024, FinGPT implementations are available on Hugging Face [40], leveraging several open-source LLMs such as Llama-2 (7B, 13B) [41], Falcon (7B) [42], MPT (7B) [43], and BLOOM (7B1) [39]. Using open-source LLMs enhances trust in the model by providing full access to the codebase. This openness accelerates research and enables the development of personalized models through fine-tuning.

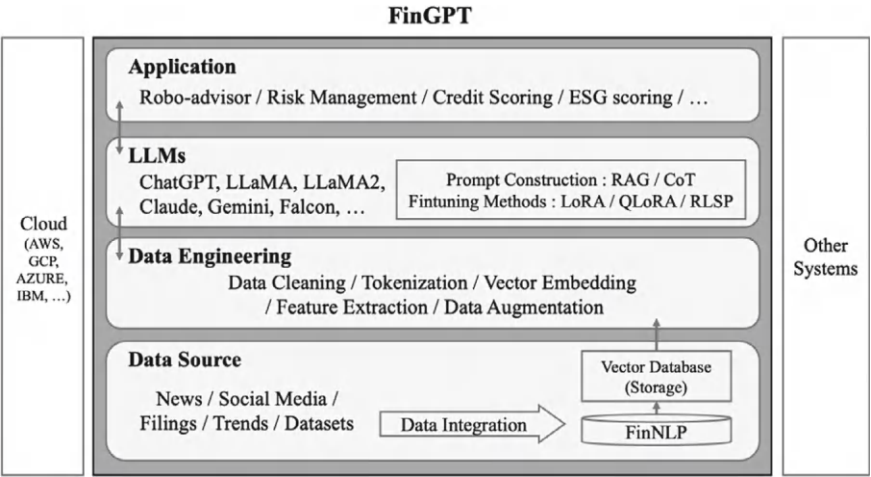


Fig. 6 FinGPT framework [29]

Two key fine-tuning methods used in FinGPT are Low-Rank Adaptation (LoRA) [44] and Reinforcement Learning on Stock Prices (RLSP). LoRA is a Parameter-Efficient Fine-Tuning (PEFT) technique that updates only a small subset of parameters instead of retraining the entire model, making it an efficient approach for model tuning [45]. Details about LoRA are provided in Sect. 5.1. RLSP trains FinGPT by rewarding accurate stock price predictions and penalizing incorrect ones, helping the model improve iteratively. This mechanism is analogous to Reinforcement Learning with Human Feedback (RLHF) used in ChatGPT but tailored for stock price prediction.

Another fine-tuning method suggests three stages of Instruction Fine-Tuning (IFT) [46] as shown in Fig. 7. The first is Task-Specific Instruction Tuning, where the model is fine-tuned with instruction data specific to a particular task, such as sentiment analysis or stock price prediction, to create a model specialized for that task. The second is Multi-Task Instruction Tuning [48], which involves using instruction data from multiple domains simultaneously to enable the model to handle various tasks. Finally, the third is fine-tuning the model to provide high-quality responses with zero-shot prompting without additional explanation. This paradigm enhances adaptability to various financial datasets while enabling cost-effective and systematic benchmarking in task-specific, multi-task, and zero-shot instruction tuning tasks.

Other Open-Source FinLLMs

Other open-source financial LLMs include FinMA (or PIXIU) [27] and InvestLM [28]. FinMA utilizes two fine-tuned LLaMA models [49] (7B, 30B) as its backbone and has also developed an evaluation benchmark called FLARE (Financial Language Understanding And Prediction Evaluation Benchmark). FLARE includes financial prediction tasks such as stock movement prediction and credit scoring. InvestLM, on the other hand, employs LLaMA (65B) as its backbone model. In its

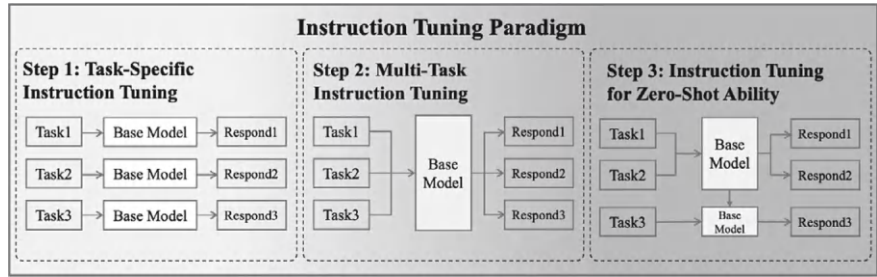


Fig. 7 Overview of the proposed instruction tuning paradigm in FinGPT [47]

research, InvestLM focuses on building its own instruction dataset, further enhancing the model’s ability to handle financial-specific tasks effectively (Table 1).

In this way, research and development on FinLLMs that can obtain high-quality and accurate answers by specializing LLMs in finance are actively underway. LLMs that have undergone the pre-train or fine-tune introduced above can perform the work required in the field of finance more accurately than general LLMs.

5 Challenges and Opportunities in Financial LLM Applications

In the digital era, financial institutions and banks leverage LLMs to streamline decision-making processes and enhance customer satisfaction. In the financial sector, providing users with the most up-to-date information to support informed decision-making is crucial. However, LLMs face certain limitations in this regard. Due to their size and complexity, LLMs need help to incorporate real-time information effectively. Additionally, as black-box models, their inner workings are not easily interpretable, raising concerns about explainability. Moreover, the high computational cost of running large models can make frequent use financially prohibitive for users. This chapter will explore how these challenges can be addressed and overcome across various sections.

5.1 Reflecting Real-Time Data

In finance analysis and finance-related AI services, incorporating the latest data and real-time updates (i.e., real-time fluctuations) is crucial. The financial environment constantly changes as news, social media posts, and other market-related information flow every minute and second. However, due to the nature of LLMs, which

Table 1 Summary of FinLLMs and their classification by use cases

Model	Backbone	Params	Techniques	Usage
BloombergGPT [26]	BLOOM [39]	50B	PT, PE	Generation of Bloomberg query language (BQL) Suggesting news headlines Answering financial questions
FinGPT [29]	Several open-source LLMs	7B	PT, IFT [46], LoRA [44], RLPE,	Stock movement prediction Robo-advisor Insolvency prediction
FinMA [27]	LLaMA [49]	7B / 30B	IFT, PE	News headline classification Stock movement prediction Fraud detection
InvestLM [28]	LLaMA	65B	IFT, PE, PEFT	Investment advice Financial text summarization

PT = Pre-training, PE = Prompt engineering

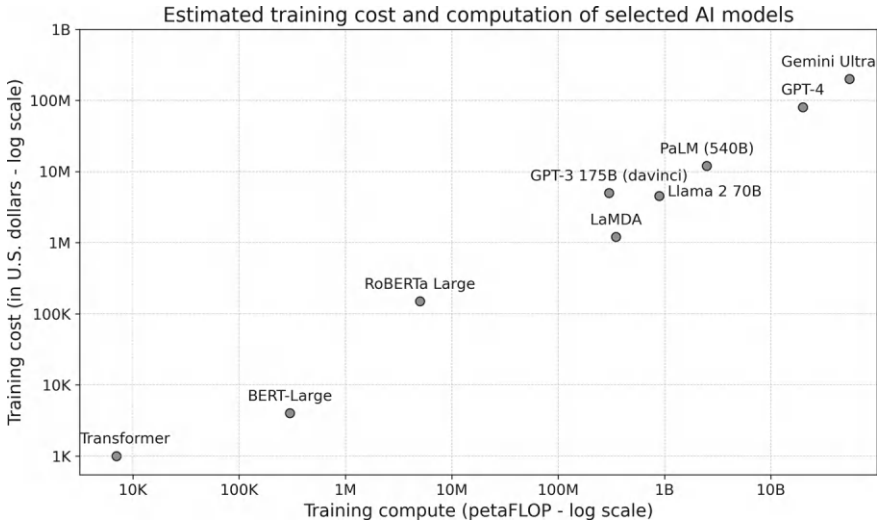


Fig. 8 Estimated training cost and computation of selected AI models [50]

generate responses based solely on pre-trained information, it is very challenging to incorporate the latest data in real time.

If fine-tuning a pre-trained model can be done quickly (i.e., reducing computational costs), it would be possible to inject the latest data into the model relatively quickly. This is achieved through efficient parameter tuning and model compression.

On the other hand, if LLMs could be directly connected to live social media feeds, news sites, or other frequently updated information sources, they would be able to provide users with the latest information. This technology is referred to as RAG [3].

Cost Issue

AI services in finance are often implemented on edge devices, such as ATMs, mobile banking applications, and branch-embedded systems, where AI models are deployed. However, these devices face significant constraints in computational power and memory. Additionally, since edge devices directly interact with customers, the services must operate quickly to provide a satisfactory user experience. Large language models (LLMs), with their billions of parameters, pose a challenge in resource-constrained environments due to the difficulty of delivering fast results. Furthermore, deploying AI services on a wide scale must consider operational costs, which are inevitably high for LLMs due to their substantial computational resource requirements (Fig. 8).

Reducing the model size while maintaining its performance can address these challenges. Smaller models can execute faster in resource-constrained environments and require less computational power for storage and processing, thereby lowering operational costs.

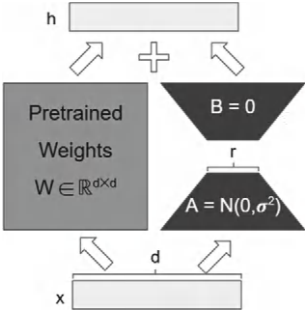
The effectiveness of this approach can be illustrated with an example. The BloombergGPT [26] model is a leading AI model for financial natural language processing, but its development entails substantial investment. Training this model requires approximately 50 billion parameters and a dataset comprising 708 billion tokens, leading to significant computational costs. Moreover, continually retraining the model to reflect the evolving financial market becomes costly.

However, a framework that addresses the challenges of BloombergGPT has been introduced: FINGPT [29]. This framework collects and utilizes financial data in real-time while employing **Low-Rank Adaptation (LoRA)** [44] to significantly reduce the number of trainable parameters. Specifically, FINGPT gathers diverse financial data sources such as Financial News, Company Filings, Social Media Discussions, and Company Announcements to update the model continuously.

By integrating LoRA for efficient updates, FINGPT reduces the number of trainable parameters from 6 billion to just 3 million. This dramatic reduction alleviates the computational burden of retraining while ensuring that the model remains updated with the latest financial information, making it a cost-effective and resource-efficient solution for real-world applications in finance.

To reduce computational costs in FINGPT, the method LoRA was applied. LoRA is a technique that learns downstream tasks by injecting low-rank decomposed matrices into a pre-trained model. Specifically, it freezes the parameters of the pre-trained model and introduces low-rank decomposed matrices into each layer, which are then trained on task-specific information. Because only the low-rank matrices are trained, the number of parameters involved in training and the memory requirements are significantly reduced. Compared to fine-tuning the GPT-3 model with 175 billion parameters, LoRA reduces the number of trainable parameters by a factor of 10,000 and decreases GPU memory requirements by threefold (Fig. 9).

Fig. 9 LoRA’s concept [44]



LoRA belongs to a broader class of techniques known as Efficient Parameter Tuning [45], which involves adding a small number of parameters to a language model and fine-tuning only these added parameters. Other methods in this category include Adapter Tuning [51] and Prefix Tuning [52]. These techniques enable tuning of a fraction of the parameters while achieving performance comparable to full fine-tuning. This means fewer parameters need to be trained and stored, but the model's overall performance remains intact.

In addition to efficient parameter tuning, there are methods for model compression that achieve similar effects. While efficient parameter tuning focuses on reducing the parameters added for downstream tasks, model compression techniques aim to minimize the size and memory requirements of the backbone model itself. Representative methods include **quantization**, **pruning**, and **knowledge distillation**.

Quantization [53] reduces the number of bits needed to represent the trained parameters of a model. Memory usage and latency can be significantly decreased by converting weights stored as 32-bit or 16-bit floating-point numbers to 8-bit integers (INT8).

Pruning reduces the model size by removing less important weights [54]. The general process involves [55]

1. Assigning importance scores to weights.
2. Comparing weights within groups.
3. Removing weights with the lowest importance scores in each group.

Early pruning methods required retraining or iterative processes, demanding substantial computational resources. However, recent advancements allow pruning in a single forward pass using calibration data, eliminating the need for repeated retraining. This streamlined approach effectively reduces model size while maintaining performance (Fig. 10).

Knowledge distillation [56] transfers knowledge from a large model (the teacher network) to a smaller model (the student network) to enable the student model to achieve performance similar to the teacher model. The training process involves [57]

- Soft labeling: The teacher network generates soft labels, representing probability distributions over classes. These labels help the student network learn from the nuances of the teacher's predictions.
- Loss function [58]: Metrics such as KL-divergence measure the output differences between the teacher and student, guiding the training process (Fig. 11).

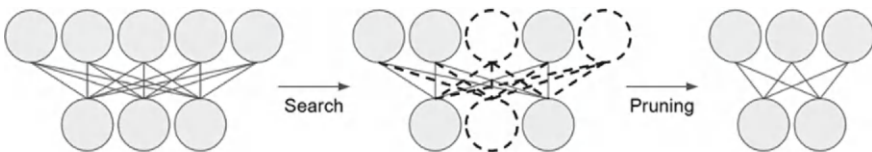


Fig. 10 Pruning's concept [54]

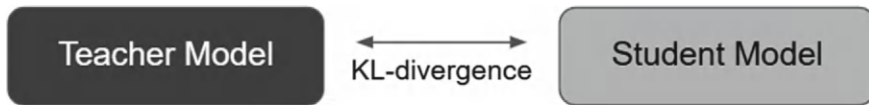


Fig. 11 Knowledge distillation's concept [56]

Let's examine a study that applied knowledge distillation to financial time series forecasting tasks [57]. This study addressed the issue of training stability caused by noisy data in financial time series forecasting by utilizing an online distillation approach.

In this method, the teacher network and student network are trained simultaneously in an online manner. The process involves the following steps:

1. **Teacher Network Ensemble:** The teacher network ensemble is trained using existing labels, which inherently contain noise.
2. **Label Refinement:** The ensemble extracts refined labels that mitigate the noise effect.
3. **Student Network Training:** These refined labels are then used to train the student network.

This approach effectively creates a smaller, noise-resilient model.

Both efficient parameter tuning and model compression methods like quantization, pruning, and distillation significantly reduce computational resource requirements. This reduction lowers operational costs and enhances the speed of AI services, delivering better user experiences. Continuous research in these areas is critical to developing advanced financial services. Applying novel methods to financial LLMs will remain essential for creating efficient and high-performing AI solutions.

Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) [3] introduces an information retrieval process, which enhances the generation process by retrieving relevant objects from available data stores, leading to higher accuracy and better robustness [59]. In the case of LLMs, since they generate answers based on large pre-trained datasets, they cannot know external domain-specific data, internal knowledge from an organization, or newly discovered knowledge. By using RAG, LLMs can access unseen data-external knowledge that was not parameterized during pre-training—allowing them to generate answers with the latest information. Furthermore, while LLMs are black-box models and cannot show which information was used or why a specific output was generated, RAG provides the advantage of checking which documents were retrieved and used. This approach helps address the black-box issue in the model's decision-making process, thereby increasing the reliability of the generated results. Compared to fine-tuning to incorporate new data, RAG requires less time and cost to return relevant data, making it a more cost-effective and efficient approach (Fig. 12).

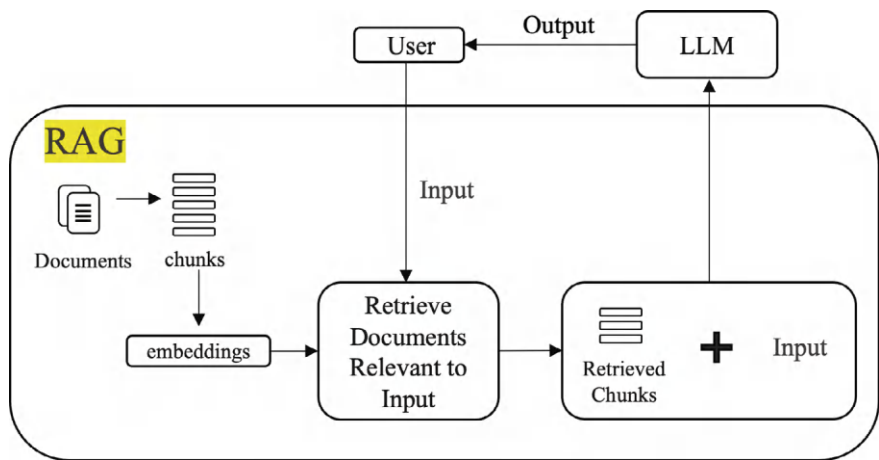


Fig. 12 A representative instance of the RAG process applied to question answering [60]

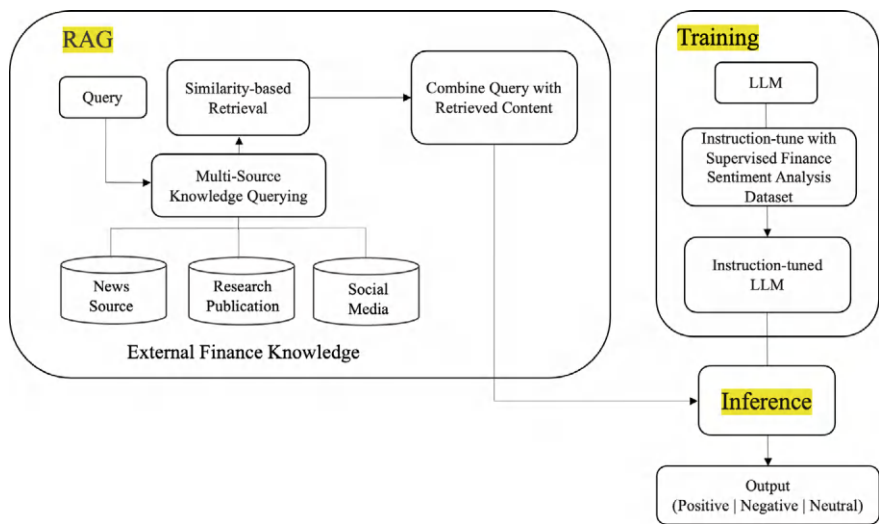


Fig. 13 Framework of retrieval-augmented large language model for financial sentiment analysis [61]

Let me introduce a framework that combines RAG with the previously discussed FinGPT [61]. They present a retrieval-augmented LLM framework specifically designed for financial sentiment analysis, optimizing information depth and context through external knowledge retrieval, thereby ensuring nuanced predictions [62]. The RAG structure in FinGPT follows a two-step knowledge retrieval process: 1. multi-source knowledge querying and 2. similarity-based retrieval (Fig. 13).

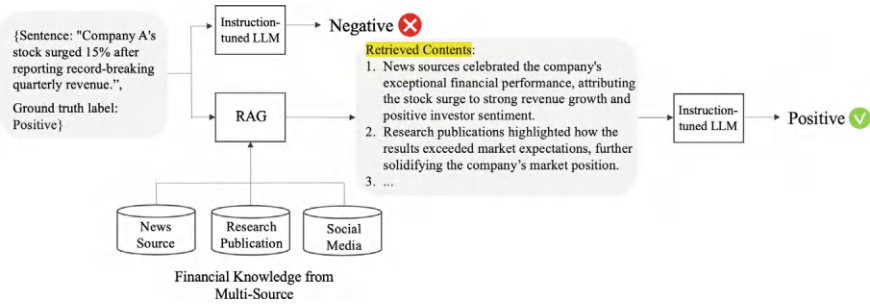


Fig. 14 A showcase of RAG-instruction-tuned LLM [64]

Before searching, it is crucial to distinguish the sources of the data, as the accuracy of the results greatly depends on where the data is retrieved from. The sources are categorized into news outlets like Bloomberg, Yahoo Finance, Reuters, CNBC, and Market Screener; renowned institutions like Goldman Sachs’ Marquee, Citi’s Velocity, and Seeking Alpha and social media platforms such as X (formerly Twitter) and Reddit. Now, the two-step knowledge retrieval process produces query results in two stages.

In the first step, the multi-source knowledge querying phase, unnecessary content such as financial article headlines or tweets is filtered out based on the query. If the query includes time-related information, the search is restricted to that specific time range to improve the quality of the search. The search then returns a list of relevant context snippets from the identified financial sources.

In the second step, the most relevant content from the results obtained in the first stage is filtered and extracted. Experimentally, only contexts with a similarity score of 0.8 or higher are selected. To accurately search for key financial terms and catch subtle differences in specific financial terms, the Szymkiewicz–Simpson coefficient [63] is used as the similarity measurement. This allows for a precise search, especially in financial news, where it is crucial to identify the correct stock tickers and financial terminology. As a result, the LLM can generate appropriate responses to the query using the retrieved context from the RAG mechanism (Fig. 14).

For example, when given the sentence “Company A’s stock surged 15% after reporting record-breaking quarterly revenue,” the structure for determining whether the statement is positive or negative using RAG is shown.

In this way, RAG allows for the incorporation of external knowledge to reflect the latest information. However, information retrieval is inherently flawed due to information loss in item representations and Approximate Nearest Neighbor (ANN) searches [59]. This is a fundamental limitation of RAG, meaning that irrelevant content or misleading information could be retrieved and returned. Therefore, even if RAG is used, it is not 100% reliable due to the intervention of various factors, so it should be kept in mind that some caution should be exercised when using it.

5.2 Explainability

As black-box models, LLMs lack transparency in how they generate results, making it difficult to understand the internal processes behind their outputs [4]. This characteristic gives rise to explainability issues, where explainability refers to the ability to articulate the model's outcomes in human-understandable terms [65].

In the financial sector, explainability issues raise significant concerns about using LLMs. Transparency in decision-making processes is crucial in this field, as incorrect decisions can lead to substantial financial losses. To address these concerns and facilitate the adoption of LLMs in finance, the integration of explainable AI (XAI) is essential.

The necessity of explainable AI (XAI) can be clearly identified through regulatory requirements [65]. The Monetary Authority of Singapore (MAS) mandates adherence to the principles of Fairness, Ethics, Accountability, and Transparency (FEAT) in developing AI solutions. This emphasizes the importance of considering model transparency. Similarly, the European Union's General Data Protection Regulation (GDPR) introduced a "right to explanation" in 2018. Under this law, individuals affected by automated decision-making solutions have the right to request an explanation of the results produced by the model.

In line with these demands for explainable AI, related research in the financial sector has been actively pursued [65]. Key focus areas include credit evaluation, financial prediction, and financial analytics.

- **Credit Evaluation:** Research in this domain has addressed explainable AI for credit assessment, credit risk management, and credit scoring.
- **Financial Prediction:** Studies have focused on asset allocation, market condition forecasting, volatility forecasting, algorithmic trading, financial growth rate predictions, economic crisis forecasts, bankruptcy prediction, fraud detection, and mortgage default prediction.
- **Financial Analytics:** Research has explored explainable AI applications in financial text classification and the analysis of spending behavior.

These efforts aim to enhance the transparency and reliability of AI systems used in finance, ensuring compliance with regulatory expectations and fostering trust in automated decision-making processes.

Some detailed method for enhancing the interpretability of LLM include *local explanation methods*. As illustrated in Fig. 15, local explanation can be categorized into four subareas that provide insights into how individual predictions are formed [66]:

1. **Attention-Based Visualization:** For example, a bipartite graph can represent the attention weights at a particular transformer layer, illustrating how the model focuses on specific parts of input sentences (see [17]).
2. **Perturbation Experiments:** Modifying an input question by removing certain words can paradoxically increase a model's confidence in nonsensical answers, demonstrating how sensitive models are to small input changes (see [67]).

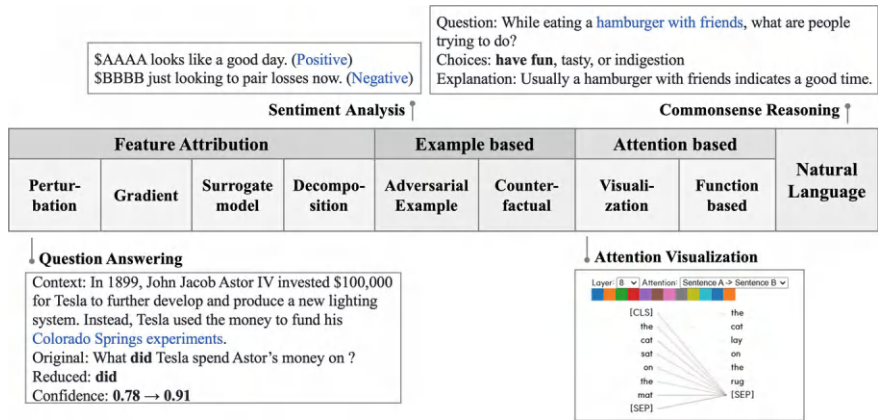


Fig. 15 Overview of local explanation framework to enhance explainability of LLM [66]

3. **Attribution Methods:** Techniques like Shapley values identify and quantify each input token’s contribution to a model’s predictions, making it clearer why certain outputs are generated (see [68]).
4. **Common-sense Reasoning and Model Robustness:** Providing explanations for which parts of the input are most crucial helps the model justify its reasoning process. Further, testing the model with negative examples and imperceptible adversarial perturbations reveals its strengths, weaknesses, and susceptibility to shifts in input structure (see [69–71]).

Let us examine a specific study related to financial prediction [5]. This research focuses on predicting weekly and monthly returns of NASDAQ-100 stock prices using time series data while also generating explanations for the results. The study leverages the ability of large language models (LLMs) to produce natural language outputs, enabling the model to articulate its reasoning process. To achieve this, the researchers used GPT-4 [72], the state-of-the-art LLM at the time, combined with structured prompts. The prompt provided instructions on the task to be performed and included the data. Additionally, the phrase “Can you reason step by step before the finalized output?” was appended to the prompt, prompting the model to output both the results and the reasoning process.

However, this study has limitations. Notably, it does not account for the potential issue of LLM hallucinations. Hallucination in LLMs refers to the phenomenon where the model generates incorrect information or presents non-existent facts as true. The evaluation in this study relies solely on automated text similarity metrics such as ROUGE [73] and BLEU [74], without addressing hallucination risks. Consequently, there is a possibility that the explanations generated for the reasoning process may not always be accurate, highlighting a critical limitation of the approach.

As highlighted in the aforementioned study, the limitation of not evaluating hallucinations in LLMs ties into a broader challenge [65]: the lack of suitable metrics

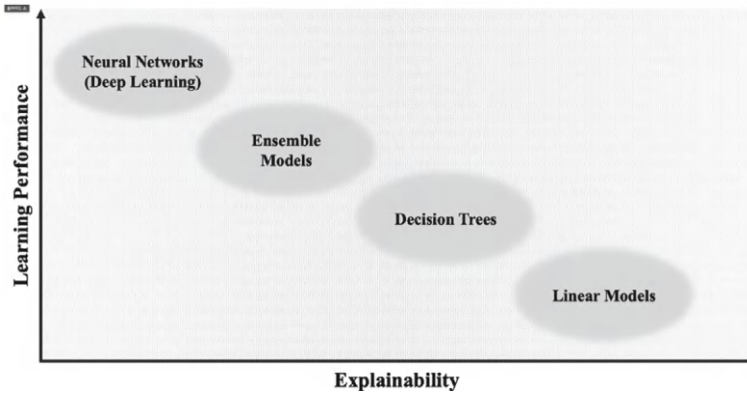


Fig. 16 *Learning performance versus explainability trade-off for several categories of learning techniques*

for assessing explanations in explainable AI (XAI). In other words, no universally accepted metric measures the quality of explanations generated by models. While some research has been conducted in this area, evaluation methods fall into two main categories: statistical evaluations and expert opinion-based evaluations.

- **Statistical Evaluations:** These involve quantitative measures like F1-score and accuracy. Although research into quantitative evaluation has been undertaken, it remains limited. Moreover, comparative studies across various XAI techniques are even rarer. This is because the models used by different techniques vary, and the structures of their explanations differ, making comparisons challenging.
- **Expert Opinion-Based Evaluations:** These rely on subjective assessments of what constitutes a “good” explanation. However, there is no consensus on appropriate metrics for expert evaluations.

This highlights the need for more extensive research into establishing suitable metrics for evaluating explanations in XAI (Fig. 16).

Another challenge [65] is the trade-off between performance and interpretability. It remains challenging to satisfy both criteria simultaneously. Current choices often involve sacrificing one for the other: for instance, opting for interpretable white-box models like decision trees at the expense of performance or using high-performance black-box models like GPTs [72, 75, 76] while compromising interpretability. However, the financial sector demands models that excel in both performance and interpretability. Therefore, it is necessary to continue advancements in XAI research.

6 Conclusion

This chapter provides comprehensive exploration of how LLMs are revolutionizing the financial industry. It begins by addressing the challenges of managing diverse and complex financial data and ongoing digital transformation within financial domains. LLMs are presented as pivotal tools that not only automate repetitive tasks but also augment human decision-making by extracting actionable insights from massive datasets. The chapter delves into the underlying mechanisms of LLMs and their adaptations for financial applications, illustrating their versatility through domain-specific models like BloombergGPT and FinGPT. These models demonstrate superior performance over general-purpose LLMs (or LMs) in specialized financial tasks like market forecasting. However, despite their impressive capabilities, LLMs face critical challenges, including high computational costs, insufficient interpretability, and difficulties in processing real-time data-factors, that are especially crucial in financial sector. To address these limitations, we thus discuss potential solutions for these challenges like Retrieval-Augmented Generation (RAG), model compression techniques (e.g., LoRA, quantization, pruning), and explainable AI frameworks to alleviate shortcomings of LLMs. In conclusion, this chapter underscores the transformative potential of LLMs in finance, emphasizing their ability to automate processes, enhance decision-making, and meet evolving industry demands.

References

1. Brown TB (2020) Language models are few-shot learners. [arXiv:2005.14165](#)
2. Khan A, Baharudin BB, Khan K (2020) Financial sentiment analysis using machine learning techniques. *Expert Syst Appl* 149:113290
3. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst* 33:9459–9474
4. Arsenault P-D, Wang S, Patenaude J-M (2024) A survey of explainable artificial intelligence (XAI) in financial time series forecasting. [arXiv:2407.15909](#)
5. Yu X, Chen Z, Lu Y (2023) Harnessing LLMs for temporal data—a study on explainable financial time series forecasting. In: *Proceedings of the 2023 conference on empirical methods in natural language processing: industry track*, pp 739–753
6. Gomber P, Kauffman RJ, Parker C, Weber BW (2018) On the fintech revolution: interpreting the forces of innovation, disruption, and transformation in financial services. *J Manag Inf Syst* 35(1):220–265
7. Chen MA, Wu Q, Yang B (2019) How valuable is FinTech innovation? The case of peer-to-peer lending. *Manag Sci* 65(12):5655–5677
8. Delen D, Kuzey C, Uyar A (2020) Exploring the analytics of big data in finance: a review of literature and future directions. *J Bus Res* 124:584–609
9. Healy PM, Palepu KG (2001) Information asymmetry, corporate disclosure, and the capital markets: a review of the empirical disclosure literature. *J Account Econ* 31(1–3):405–440
10. Gao J, Galley M, Li L (2019) Towards conversational AI: a neural approach. *ACM SIGKDD Explor Newsl* 19(2):25–35
11. Rau F, Soto I, Zabala-Blanco D (2021) Forecasting mobile network traffic based on deep learning networks. In: *2021 IEEE Latin-American conference on communications (LATINCOM)*. IEEE

12. Amjadian E et al (2021) Attended-over distributed specificity for information extraction in cybersecurity. In: 2021 IEEE aerospace conference (50100). IEEE
13. Naseem U, Razzak I, Musial K, Imran M (2020) A survey on deep neural network techniques for sentiment analysis. *IEEE Trans Comput Soc Syst* 7(3):622–634
14. Vaswani, A (2017) Attention is all you need. *Adv Neural Inf Process Syst*
15. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088):533–536
16. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
17. Vig J (2019) BertViz: a tool for visualizing multi-head self-attention in the BERT model
18. Araci D (2019) FinBERT: financial sentiment analysis with pre-trained language models. [arXiv:1908.10063](https://arxiv.org/abs/1908.10063)
19. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT 2019*, pp 4171–4186
20. Yang Y, Uy MC, Huang A (2020) FinBERT: a pretrained language model for financial communications. [arXiv:2006.08097](https://arxiv.org/abs/2006.08097)
21. Liu Z, Huang D, Huang K, Li Z, Zhao J (2020) FinBERT: a pre-trained financial language representation model for financial text mining. In: *International joint conference on artificial intelligence*
22. Clark K, Luong MT, Le QV, Manning CD (2020) ELECTRA: pre-training text encoders as discriminators rather than generators. [arXiv:2003.10555](https://arxiv.org/abs/2003.10555)
23. Shah RS, Chawla K, Eidnani D, Shah A, Du W, Chava S, Raman N, Smiley C, Chen J, Yang D (2022) When FLUE meets FLANG: benchmarks and large pretrained language model for financial domain. In: *Conference on empirical methods in natural language processing*
24. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D (2017) Deep reinforcement learning from human preferences. *Adv Neural Inf Process Syst* 30
25. Stiennon N, Ouyang L, Wu J, Ziegler DM, Lowe R, Voss C et al (2020) Learning to summarize with human feedback. *Adv Neural Inf Process Syst* 33:3008–3021
26. Wu S et al (2023) Bloomberggpt: a large language model for finance. [arXiv:2303.17564](https://arxiv.org/abs/2303.17564)
27. Xie Q et al (2023) Pixiu: a large language model, instruction data and evaluation benchmark for finance. [arXiv:2306.05443](https://arxiv.org/abs/2306.05443)
28. Yang Y, Tang Y, Tam KY (2023) Investlm: a large language model for investment using financial domain instruction tuning. [arXiv:2309.13064](https://arxiv.org/abs/2309.13064)
29. Yang H, Liu X-Y, Wang CD (2023) Fingpt: open-source financial large language models. [arXiv:2306.06031](https://arxiv.org/abs/2306.06031)
30. Liu P et al (2022) Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. [arXiv:2107.13586](https://arxiv.org/abs/2107.13586)
31. Wei J et al (2022) Chain of thought prompting elicits reasoning in large language models. [arXiv:2201.11903](https://arxiv.org/abs/2201.11903)
32. Winston PH (1980) Learning and reasoning by analogy. *Commun ACM* 23(12):689–703
33. Sun T et al (2022) BBTv2: towards a gradient-free future with large language models. [arXiv:2205.11200](https://arxiv.org/abs/2205.11200)
34. Bach N, Badaskar S (2007) A survey on relation extraction. *Lit Rev Lang Stat II* 2(1–4):1–15
35. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y et al (2023) A survey on hallucination in natural language generation. *ACM Comput Surv (CSUR)* 55(12):1–38
36. Radford A et al (2019) Language models are unsupervised multitask learners. *OpenAI blog* 1.8 9
37. OpenAI (2024) ChatGPT (December 2024 version). Accessed 14 Dec 2024. <https://gptonline.ai/ko/>
38. Lee J et al (2024) A survey of large language models in finance (FinLLMs). [arXiv:2402.02315](https://arxiv.org/abs/2402.02315)
39. Le Scao T, Fan A, Akiki C, Pavlick E, Ilić S, Hesslow D et al (2023) BLOOM: A 176B-parameter open-access multilingual language model, vol 4. *BigScience Workshop*. [arXiv:2211.05100](https://arxiv.org/abs/2211.05100). <https://doi.org/10.48550/arXiv.2211.05100>.

40. Hugging Face FinGPT: financial large language model. <https://huggingface.co/FinGPT>. Accessed 14 Dec 2024
41. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S et al (2023) Llama 2: open foundation and fine-tuned chat models. [arXiv:2307.09288](https://arxiv.org/abs/2307.09288)
42. Almazrouei E, Alobeidli H, Alshamsi A, Cappelli A, Cojocaru R, Debbah M et al (2023) The falcon series of open language models, vol 2. [arXiv:2311.16867](https://arxiv.org/abs/2311.16867). <https://doi.org/10.48550/arXiv.2311.16867>
43. Lin K, Lin C-C, Liang L, Liu Z, Wang L (2023) MPT: mesh pre-training with transformers for human pose and mesh reconstruction, vol 2. [arXiv:2211.13357](https://arxiv.org/abs/2211.13357)
44. Hu EJ et al (2021) Lora: low-rank adaptation of large language models. [arXiv:2106.09685](https://arxiv.org/abs/2106.09685)
45. Jin F, Zhang J, Zong C (2023) Parameter-efficient tuning for large language model without calculating its gradients. In: Proceedings of the 2023 conference on empirical methods in natural language processing
46. Zoph A, Pham T, Le M, Caruana RA, Le QV (2020) Finetuned language models are zero-shot learners. [arXiv:2010.11934](https://arxiv.org/abs/2010.11934)
47. Wang N, Yang H, Wang CD (2023) Fingpt: instruction tuning benchmark for open-source large language models in financial datasets. [arXiv:2310.04793](https://arxiv.org/abs/2310.04793)
48. Wang X, Zhou W, Zu C, Xia H, Chen T, Zhang Y, Zheng R, Ye J, Zhang Q, Gui T, Kang J, Yang J, Li S, Du C (2023) InstructUIE: multi-task instruction tuning for unified information extraction
49. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F et al (2023) Llama: open and efficient foundation language models. [arXiv:2302.13971](https://arxiv.org/abs/2302.13971)
50. Stanford Institute for Human-Centered AI (2024) AI index report 2024. Stanford University. <https://aiindex.stanford.edu/report/>
51. Houshy N, Giurigu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A et al (2019) Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th international conference on machine learning (ICML), Long Beach, CA, USA
52. Li XL, Liang P (2021) Prefix-tuning: optimizing continuous prompts for generation. [arXiv:2101.00190](https://arxiv.org/abs/2101.00190)
53. Ahmadian A, Dash S, Chen H, Venkitesh B, Gou S, Blunsom P, Üstün A, Hooker S (2023) Intriguing properties of quantization at scale. [arXiv:2305.19268v1](https://arxiv.org/abs/2305.19268v1) [cs.LG]
54. Chen Y, Shang J, Zhang Z, Cui S, Liu T, Wang S, Sun Y, Wu H (2024) LEMON: reviving stronger and smaller LMs from larger LMs with linear parameter fusion. In: Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers), pp 8005–8019
55. Sun M, Liu Z, Bair A, Kolter JZ (2024) A simple and effective pruning approach for large language models. In: Proceedings of the international conference learning representations (ICLR)
56. Li Z, Li H, Meng L (2023) Model compression for deep neural networks: a survey. *Computers* 12(3):60. <https://doi.org/10.3390/computers12030060>
57. Floratos P et al (2022) Online knowledge distillation for financial timeseries forecasting. In: 2022 international conference on innovations in intelligent systems and applications (INISTA). IEEE
58. Wu T, Tao C, Wang J, Yang R, Zhao Z, Wong N (2024) Rethinking kullback-leibler divergence in knowledge distillation for large language models. [arXiv:2404.02657v4](https://arxiv.org/abs/2404.02657v4) [cs.CL]
59. Zhao P et al (2024) Retrieval-augmented generation for ai-generated content: a survey. [arXiv:2402.19473](https://arxiv.org/abs/2402.19473)
60. Gao Y et al (2023) Retrieval-augmented generation for large language models: a survey. [arXiv:2312.10997](https://arxiv.org/abs/2312.10997)
61. Zhang B et al (2023) Enhancing financial sentiment analysis via retrieval augmented large language models. In: Proceedings of the fourth ACM international conference on AI in finance
62. AI4Finance Foundation FinGPT: financial large language model. <https://github.com/AI4Finance-Foundation/FinGPT>. Accessed 14 Dec 2024

63. Vijaymeena MK, Kavitha K (2016) A survey on similarity measures in text mining. *Mach Learn Appl*
64. AI4Finance Foundation (2023) FinGPT: FinGPT_RAG. GitHub repository. [Online]. https://github.com/AI4Finance-Foundation/FinGPT/tree/master/fingpt/FinGPT_RAG. Accessed 14 Dec 2024
65. Yeo WJ, Van der Heever W, Mao R, Cambria E, Satapathy R, Mengaldo G (2023) A comprehensive review on financial explainable AI. [arXiv:2309.11960](https://arxiv.org/abs/2309.11960)
66. Zhao H et al (2024) Explainability for large language models: a survey. *ACM Trans Intell Syst Technol* 15(2): 1-38
67. Feng S et al (2018) Pathologies of neural models make interpretations difficult. [arXiv:1804.07781](https://arxiv.org/abs/1804.07781)
68. Chen H et al (2023) Algorithms to estimate Shapley value feature attributions. *Nat Mach Intell* 5(6):590–601
69. Rajani NF et al (2019) Explain yourself! leveraging language models for commonsense reasoning. [arXiv:1906.02361](https://arxiv.org/abs/1906.02361)
70. Wu T et al (2021) Polyjuice: generating counterfactuals for explaining, evaluating, and improving models. [arXiv:2101.00288](https://arxiv.org/abs/2101.00288)
71. Jin D, Jin Z, Zhou JT, Szolovits P (2020) Is bert really robust? natural language attack on text classification and entailment. In: *AAAI conference on artificial intelligence (AAAI)*
72. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S et al (2023) Gpt-4 technical report. OpenAI
73. Lin C-Y (2004) ROUGE: a package for automatic evaluation of summaries. In: *Text summarization branches out, Barcelona, Spain*, pp 74–81
74. Papineni K et al (2002) Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the association for computational linguistics*
75. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. *Adv Neural Inf Process Syst* 33:1877–1901
76. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. *OpenAI Blog* 1(8):9

Voluntary Sustainability Disclosure and Third-Party Assurance: A Large Language Model Perspective



SoHyeon Kang  and Sewon Kwon 

Abstract This chapter explores the influence of third-party assurance on the sentiment and subjectivity of corporate sustainability reports. As sustainability reporting grows in importance for corporate transparency and stakeholder engagement, the role of assurance—whether limited or reasonable and provided by audit or non-audit firms—has become increasingly critical in shaping report narratives. Utilizing advanced sentiment analysis methodologies, including BERT-based models, we analyze the tonal qualities and factual content of sustainability reports. Findings indicate that reports without assurance exhibit more positive sentiment, while those with limited or reasonable assurance reflect varying degrees of narrative objectivity and sentiment neutrality. Reports assured by audit firms tend to convey more neutral and comprehensive narratives, emphasizing factuality over subjective tones, compared to those assured by non-audit entities. This analysis contributes to the understanding of how assurance practices impact the perceived credibility and faithful representation of narrative sustainability disclosures. By combining natural language processing insights with empirical data, our study underscores the transformative role of assurance in enhancing nonfinancial disclosures quality and fostering accountability. The results provide policy and practical implications for the discussion on mandating third-party assurance of nonfinancial disclosures.

Keywords Sustainability reporting · Third-Party assurance · Sentiment analysis · ESG disclosure · BERT model · Narrative objectivity · Sentiment analysis

S. Kang · S. Kwon (✉)

College of Business Administration, Ewha Womans University, Seoul, Korea

e-mail: k4js1@ewha.ac.kr

S. Kang

e-mail: laila.kang@ewha.ac.kr

1 Introduction

In recent years, sustainability or nonfinancial disclosures have become a critical element of corporate accountability. As businesses navigate the complexities of environmental, social, and governance (ESG) reporting, the role of third-party assurance has gained prominence for its potential to enhance the credibility and reliability of disclosed information. This study examines the sentiment and subjectivity of sustainability report narratives, comparing reports with different levels of assurance and those assured by different types of assurance providers. Leveraging advanced sentiment analysis techniques, including BERT-based LLM models, the analysis seeks to uncover how assurance status and provider influence the tonal and factual composition of corporate disclosures. By bridging linguistic and financial insights, this research highlights the nuanced impact of assurance on narrative sentiment and objectivity.

Sentiment analysis, in this context, provides a window into the perceived intentions and faithful representation embedded within corporate reports. The increasing reliance on natural language processing tools underscores the importance of accurately capturing public perception and linguistic subtleties. This study's approach sheds light on how different assurance frameworks and providers impact sustainability narratives, offering policy and practical insights into the discussion on mandating third-party assurance for sustainability reports, while exploring the qualitative nuances of corporate ESG communication beyond mere compliance.

2 Literature Survey

2.1 *Trends in Mandatory ESG Disclosure and Third-Party Assurance*

2.1.1 **Mandatory ESG Disclosure Trends in Global Contexts**

The rise of mandatory ESG disclosures is reshaping corporate social responsibility on a global scale. The IFRS Foundation established the ISSB, which released the S1 and S2 standards and is expected to continue issuing sustainability disclosure standards under its 'building blocks approach'. Another significant factor has been the European Union's leadership with the Non-Financial Reporting Directive (NFRD) and the subsequent Corporate Sustainability Reporting Directive (CSRD), which require extensive ESG reporting by companies. In 2024, EFRAG published the initial ESRS package regarding mandatory ESG related disclosures, which will become effective in 2026. Furthermore, on 26 February 2025, the European Commission put forward its EU Omnibus Simplification Package. Research by Cuomo et al. [1] underscores the impact of the NFRD, indicating that it has led European firms to enhance their social responsibility disclosures and comply with elevated reporting

standards. Moreover, as reporting standards develop, studies observe that mandatory disclosures increase the likelihood of ESG information affecting firm valuation, particularly in firms where ESG metrics were previously weak (INSPIRE [2]).

In the United States, mandatory ESG reporting remains less prescriptive but is increasingly reinforced by the Securities and Exchange Commission (SEC) through frameworks focused on climate-related disclosures and social impact transparency (Deloitte [3]). The SEC's focus on climate risk disclosure aligns with global standards, but the United States continues to favor a "comply-or-explain" approach, resulting in a mixed impact across sectors and often limited comparability with European disclosures (Azeus Convene [4]). Research by Krueger et al. [5] further supports that, globally, mandatory ESG disclosure is associated with enhanced corporate transparency, reduced ESG-related risks, and improved informational environments for investors.

Other countries, such as Malaysia and Hong Kong, have adopted mandatory or "comply-or-explain" ESG frameworks to address regional sustainability concerns. Malaysia, for example, requires publicly listed companies to report on diversity, human rights, and environmental policies, moving closer to international frameworks such as the Task Force on Climate-Related Financial Disclosures (TCFD) standards. Hong Kong, meanwhile, uses a "comply-or-explain" approach, while applying specific mandates for companies listed under the Main Board Listing Rules (Azeus Convene [4]). These diverse regulatory landscapes underscore the challenges in achieving global consistency, but also reflect the expanding influence of ESG disclosure mandates across financial markets and industries worldwide.

2.1.2 The Role of Third-Party Assurance in Enhancing ESG Disclosure Quality

Companies within the scope of the CSRD are required to obtain third-party assurance on ESG reports starting in 2025, for those producing their first reports based on the financial year beginning on or after January 1, 2024. Similarly, the SEC has announced that firms disclosing Scope 1 and/or Scope 2 emissions will be required to obtain an assurance report at the limited assurance level. Comparable regulations have been enacted in Australia, New Zealand, Singapore, and India. Based on current trends in ESG assurance regulation, empirical evidence suggests that third-party assurance significantly enhances the credibility of ESG disclosures by increasing transparency and minimizing risks of greenwashing. Du and Wu [6] find that external assurance is particularly effective in reducing CSR-related misconduct by providing a level of validation that stakeholders value. This finding is reinforced by Maroun [7], who examines integrated reporting in South Africa and concludes that assurance can improve report quality by aligning disclosed information with stakeholder expectations. The study indicates that externally assured reports are associated with higher quality and greater corporate accountability.

Recent studies emphasize the strategic importance of external assurance in managing ESG reputation risk, with Asante-Appiah and Lambert [8] showing that

external auditors play a key role in ensuring the reliability of ESG information, thereby protecting corporate reputations from the potential fallout of misstated disclosures.

2.2 Sentiment Analysis Before and After Large Language Models (LLMs)

Sentiment analysis has become a crucial tool for understanding subjective aspects of textual data that quantitative financial or operational metrics alone cannot capture. As organizations increasingly recognize the value of public perception, customer feedback, and social media sentiment, sentiment analysis provides a method to analyze opinions, emotions, and attitudes toward products, services, and events. These qualitative insights, extracted through linguistic and computational techniques, allow organizations to gauge public sentiment, track brand reputation, and respond proactively to potential crises. In particular, sentiment analysis is widely applied in fields such as finance, marketing, and political science, where understanding the subtleties of public opinion or consumer attitudes can drive strategic decisions (Liu [9]). For instance, in financial contexts, sentiment analysis of news articles or social media posts can reveal market sentiment trends predictive of stock price movements and investor behavior, capturing a level of human perception and market psychology not evident in numerical data alone (Pang and Lee [10]). This ability to systematically decode subjective information has propelled sentiment analysis to the forefront of computational linguistics and natural language processing research, as organizations seek to understand and quantify qualitative data through linguistic methods.

2.2.1 Sentiment Analysis Before Large Language Models (LLMs)

Sentiment analysis, also known as opinion mining, aims to identify, classify, and quantify sentiment information in text, focusing on emotions, opinions, and attitudes. Before the advent of Large Language Models (LLMs), sentiment analysis primarily relied on rule-based methods, traditional machine learning techniques, and relatively shallow neural networks.

- **Rule-Based and Lexicon-Based Approaches**

Early sentiment analysis methods were dominated by rule-based and lexicon-based approaches, which depended heavily on predefined dictionaries or lexicons that mapped specific words to sentiment labels, such as positive, negative, or neutral. For instance, the **SentiWordNet** lexicon and **VADER (Valence Aware Dictionary and sEntiment Reasoner)** were widely used tools, especially for short texts like

tweets and product reviews. These lexicon-based methods offered simplicity and interpretability but struggled with complex sentence structures and subtle sentiment nuances [9]. Lexicon-based methods provided a foundation for sentiment analysis by mapping words and phrases to predefined sentiments, but they often lacked the sophistication to detect nuanced sentiment and contextual dependencies [9].

- **Traditional Machine Learning Methods**

In the late 2000s and early 2010s, traditional machine learning algorithms, including **Naïve Bayes**, **Support Vector Machines (SVMs)**, and **Logistic Regression**, became standard in sentiment analysis [10]. These models were trained on large, labeled datasets such as the IMDB movie review dataset or Twitter sentiment datasets, using features like word frequencies, n-grams, and part-of-speech (POS) tags to improve predictive accuracy. While more adaptable than rule-based methods, these models had limitations in capturing the contextual and sequential nature of language, often failing to recognize sarcasm, complex emotions, or ambiguous language constructs. The field began to evolve as researchers recognized the need for models that could understand context better and capture sentiment more accurately across varying domain [10].

- **Shallow Neural Networks and Word Embeddings**

The introduction of word embeddings, such as **Word2Vec** and **GloVe**, marked a shift toward using vector representations to capture semantic similarities between words. Word embeddings allowed early models to detect semantic relationships in words based on their co-occurrence patterns, paving the way for more nuanced sentiment analysis [11]. Combined with shallow neural network architectures like **Recurrent Neural Networks (RNNs)**, these embeddings enabled early sentiment models to understand words in their context, which improved performance. However, shallow networks and static embeddings were limited in handling nuanced sentiment, as they could not dynamically adjust meanings based on context. This led to the development of more advanced language models that could better address the limitations of prior methods [11].

2.2.2 Sentiment Analysis with Large Language Models (LLMs)

With the advent of Large Language Models (LLMs) such as **BERT (Bidirectional Encoder Representations from Transformers)**, **GPT (Generative Pre-trained Transformer)**, and **RoBERTa**, sentiment analysis entered a transformative phase. These models, based on the Transformer architecture, are pre-trained on massive datasets and fine-tuned on specific tasks, allowing for more nuanced and context-aware sentiment analysis [12].

- **Contextualized Word Embeddings and Deep Transformer Models**

LLMs like BERT introduced contextualized embeddings, where a word's meaning varies based on surrounding context. For instance, the word “great” in “a great

deal of trouble” would convey a different sentiment than in “a great day.” BERT’s bidirectional training enables it to understand both preceding and following words, which improves its accuracy in capturing sentiment nuances and even implicit sentiment. Studies comparing BERT to traditional sentiment analysis approaches show that BERT-based models achieve superior performance on complex sentiment tasks, particularly where context shifts the sentiment meaning [12].

- **Fine-Tuning and Transfer Learning**

LLMs are typically pre-trained on large, general datasets before fine-tuning on domain-specific sentiment analysis datasets, enabling effective transfer learning. This approach has led to strong performance improvements, particularly in domains like finance, healthcare, and social media. For example, **FinBERT** has been fine-tuned on financial sentiment data to capture sector-specific language and sentiment, such as bullish versus bearish sentiments, which traditional models could not accurately distinguish (Yang et al., 2020 [13]). This level of domain adaptability, enabled by transfer learning, represents a significant advancement from previous models, which often struggled with domain-specific language [9, 10].

- **Few-Shot and Zero-Shot Learning in Sentiment Analysis**

Advanced LLMs, such as **GPT-3** and **T5**, enable few-shot and zero-shot learning, allowing sentiment analysis without extensive labeled datasets. GPT-3 can perform sentiment analysis by generating human-like text in response to specific prompts. In zero-shot settings, GPT-3, for instance, has shown a capacity to correctly infer sentiment from a brief prompt, without prior task-specific training, enabling adaptable sentiment analysis across diverse industries and languages (Brown et al. [14]). This represents a paradigm shift, as previous models typically required large amounts of labeled data for each specific application.

- **Applications and Challenges with LLM-Based Sentiment Analysis**

LLMs have facilitated sentiment analysis applications across domains, such as **social media monitoring**, **financial forecasting**, and **customer feedback**. However, challenges remain, including computational resource requirements and potential biases embedded in large datasets. Furthermore, LLMs may still struggle with sarcasm and irony, although ongoing advancements in prompt engineering and model fine-tuning continue to address these limitations [15]. Despite these challenges, the adaptability of LLMs to multiple domains has made them a powerful tool for sentiment analysis, pushing the boundaries of what qualitative sentiment analysis can capture in real-world applications [12].

2.3 *Subjectivity Analysis in Textual Data*

Subjectivity analysis, distinct from sentiment analysis, focuses on identifying whether statements in textual data are objective (fact-based) or subjective (opinion-based). While sentiment analysis aims to detect positive, negative, or neutral emotions, subjectivity analysis evaluates the **factuality versus opinionated nature** of the content. This type of analysis is critical in fields like media, social media, and content moderation, where the ability to separate objective information from opinion-based commentary enhances accuracy in reporting, public sentiment tracking, and organizational response strategies.

2.3.1 **The Role of Subjectivity Analysis in Text Mining**

Subjectivity analysis has gained prominence as organizations increasingly need to differentiate between factual information and opinion in digital content. This distinction is particularly valuable in journalism and policy analysis, where understanding the tone and nature of content is essential for accurate information dissemination and data-driven decision-making. Subjective texts often reflect personal opinions, beliefs, or emotional expressions, while objective texts present verifiable information or fact-based descriptions. As the digital media landscape grows, the line between fact and opinion becomes blurred, making subjectivity analysis even more crucial to prevent misinformation and manage public perception effectively [16].

2.3.2 **Methods and Tools for Subjectivity Analysis**

- **Lexicon-Based Approaches**

Lexicon-based methods are among the most commonly used techniques in subjectivity analysis, relying on predefined word dictionaries that tag words according to their likelihood of being subjective or objective. For example, the **MPQA Subjectivity Lexicon** assigns values to words based on their association with opinionated or factual content. These lexicons contain words that frequently appear in subjective statements, such as “believe,” “wonder,” or “think,” and words that denote factual content, like “report,” “measure,” or “confirm.”

Lexicon-based approaches are straightforward and highly interpretable, making them ideal for real-time applications in monitoring user-generated content or news articles. However, they have limitations in handling nuanced language, idiomatic expressions, or culturally specific terms that may not be covered in the lexicon. As such, they may misinterpret context-dependent expressions or fail to capture subjectivity in complex or domain-specific texts [16].

- **Machine Learning-Based Approaches**

While lexicon-based tools remain widely used, machine learning has advanced subjectivity analysis by allowing models to learn from labeled data. Traditional models, such as **Naïve Bayes** and **Support Vector Machines (SVMs)**, are trained on labeled datasets to classify texts as objective or subjective based on features like part-of-speech tags, word n-grams, and syntactic patterns. These features help capture patterns that distinguish objective from subjective statements more effectively than simple lexicons.

Machine learning models tend to improve accuracy and adaptability to linguistic diversity, but they require large labeled datasets and can lack interpretability compared to lexicon-based methods. However, their flexibility makes them well-suited for analyzing varied content types, including opinionated articles, social media posts, and reviews [17].

- **Deep Learning and Transformer-Based Models**

The development of deep learning and transformer-based models, such as **BERT (Bidirectional Encoder Representations from Transformers)**, has brought a new level of accuracy to subjectivity analysis. Transformer models are context-aware, meaning they account for the surrounding words in a sentence when determining if the language is factual or opinion-based. BERT and other transformer-based models can be fine-tuned on subjectivity-specific tasks to detect nuanced expressions of objectivity or subjectivity in complex sentences [12].

These models are particularly effective at identifying subtle shifts in language, such as sarcasm or implied subjectivity. Unlike lexicon-based or traditional machine learning models, transformers can adapt dynamically to context, making them more reliable in distinguishing subjective content within complex language structures. However, transformer models require substantial computational power and may be resource-intensive, which can limit their practical application in some contexts.

2.3.3 Applications of Subjectivity Analysis

Subjectivity analysis has broad applications across industries that require careful content categorization and moderation:

- **News and Media:** In journalism, subjectivity analysis helps separate opinion pieces from fact-based reporting. This is essential for maintaining unbiased information environments, particularly for news aggregators and fact-checking platforms that aim to distinguish factual news from commentary.
- **Social Media Monitoring:** Social media platforms host a mix of opinionated and factual posts. Subjectivity analysis enables organizations to track both objective mentions and subjective perceptions, providing a comprehensive view of public discourse that combines factual events with personal reactions.

- **Customer Feedback Analysis:** For businesses, subjectivity analysis can help differentiate objective product issues from subjective opinions in customer feedback. This allows targeted responses by focusing on factual product concerns while understanding general customer sentiment and preferences.
- **Political Discourse Analysis:** Subjectivity analysis is valuable in political science, where researchers evaluate objectivity in political statements, speeches, or policy documents. By distinguishing between factual claims and subjective statements, researchers can identify biases or partisanship in political messaging and assess public figures' reliability.

Subjectivity analysis distinguishes objective from subjective content, serving as a foundational tool in fields where the tone of communication impacts decision-making and public perception. From lexicon-based approaches to advanced transformer models, subjectivity analysis tools offer flexibility and adaptability for varied applications, helping organizations and researchers to assess the tone and factuality of textual content in real time.

3 Research Design

We collected a total of 3,718 U.S.-based firms' sustainability reports from companies with identifiable report publications in the Refinitiv database over the period from 2013 to 2023. Subsequently, the data was merged with the Compustat database, excluding 371 firm-year observations lacking financial information. As a result, a total of 3,347 firm-year observations were used for sentiment and subjectivity analysis of the sustainability reports.

We employed a combination of BERT-based sentiment analysis and TextBlob subjectivity analysis to evaluate sentiment and subjectivity in corporate responsibility reports. The primary goal was to measure sentiment intensity across texts and differentiate subjective statements from objective ones, allowing for insights into the tonal quality and factuality within the reports by assurance types and entity who provide assurance.

3.1 *BERT-Based Sentiment Analysis*

BERT (Bidirectional Encoder Representations from Transformers) was utilized for sentiment analysis, leveraging its ability to interpret text in both forward and backward contexts. The model classifies sentiment on a five-point scale from "1 star" (very negative) to "5 stars" (very positive), which allowed for a nuanced understanding of sentiment intensity within each sentence.

BERT (Bidirectional Encoder Representations from Transformers) was utilized for sentiment analysis, leveraging its ability to interpret text in both forward and backward contexts. The model classifies sentiment on a five-point scale from “1 star” (very negative) to “5 stars” (very positive), which allowed for a nuanced understanding of sentiment intensity within each sentence.

1. Sentiment Scoring

The function for BERT sentiment analysis categorized each sentence’s sentiment based on the following mapping:

- 5 stars to 1.0 (very positive),
- 4 stars to 0.5 (positive),
- 3 stars to 0 (neutral),
- 2 stars to -0.5 (negative),
- 1 star to -1.0 (very negative).

For each document, sentiment scores were aggregated to calculate the mean and median sentiment scores, offering an overview of the document’s general sentiment tone.

2. Key BERT Mechanisms in Sentiment Analysis

For each document, sentiment scores were aggregated to calculate the mean and median sentiment scores, offering an overview of the document’s general sentiment tone.

BERT’s multi-layer Transformer architecture relies on a **self-attention mechanism**, allowing it to weigh the importance of words relative to each other within a sentence. The self-attention mechanism is computed as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where

- $Q, K,$ and V represent the query, key, and value matrices from token embeddings,
- d_k is the dimensionality of the keys, acting as a scaling factor.

This mechanism enables BERT to capture contextual dependencies effectively, which is essential for interpreting nuanced sentiment. The [CLS] token, representing the sentiment of the entire sentence, undergoes this attention processing to provide a sentiment score.

3. Sentiment Classification via Softmax

After the self-attention layers, the [CLS] token’s final hidden state $h_{[CLS]}$ acts as the sentence’s sentiment representation. The softmax layer converts this representation into probabilities for each sentiment class, identifying the class with the highest probability as the predicted sentiment:

$$P(y = c|h_{[CLS]}) = \frac{\exp(w_c^T h_{[CLS]} + b_c)}{\sum_j \exp(w_j^T h_{[CLS]} + b_j)}$$

where

- c denotes each sentiment class (e.g., 1 to 5 stars),
- w_c and b_c are weights and bias for class c ,
- $P(y = c|h_{[CLS]})$ gives the probability of class c .

3.2 Subjectivity Analysis Using TextBlob

The subjectivity analysis metric, which assesses the degree of factual versus opinion-based content, relies on lexicon-based and rule-based methods. This approach utilizes well-established subjectivity lexicons, which include words and phrases commonly associated with either objective or subjective language. Here's how subjectivity scoring determines factual (objective) versus opinion-based (subjective) content.

1. Lexicon-Based Approach

The subjectivity scoring process primarily draws from **subjectivity lexicons**, such as the **MPQA Subjectivity Lexicon**. These lexicons are databases of words annotated for their likelihood of indicating subjective or objective content. Words that frequently appear in subjective expressions, such as “believe,” “think,” or “wonder,” are labeled as subjective, while words associated with factual reporting, such as “report,” “indicate,” or “confirm,” are considered objective.

- **Subjective Words:** These words often reflect personal opinions, emotions, or evaluative statements. For instance, words like “wonderful,” “terrible,” and “prefer” are classified as subjective because they tend to express opinions or preferences.
- **Objective Words:** Objective words are typically associated with neutral, factual language. Examples include words like “reported,” “calculated,” or “observed,” which are generally used in contexts that describe factual information.

When analyzing a sentence, the scoring method checks the presence and frequency of these lexicon words to generate an overall subjectivity score.

2. Rule-Based Heuristics

In addition to lexicons, subjectivity scoring applies **rule-based heuristics** to refine the score. These rules consider factors like:

- **Modifier words:** Certain modifiers, such as “very” or “quite,” can intensify subjective statements. The scoring adjusts the subjectivity level based on the presence of such intensifiers, recognizing that “very good” is more subjective than simply “good.”

- **Contextual Phrases:** Phrases structured as expressions of opinion, such as “I think” or “we believe,” increase the subjectivity score of the sentence. Conversely, phrases like “according to data” or “the study shows” reduce the subjectivity score, indicating a more factual tone.

3. Sentence-Level Aggregation and Scoring

For each sentence, the subjectivity score ranges from 0 to 1:

- **Scores closer to 0** indicate a higher likelihood of objective content, where factual information and neutral descriptions are present.
- **Scores closer to 1** suggest a more opinion-based or evaluative tone, where subjective expressions and personal viewpoints are dominant.

For a sentence like “The results clearly show significant improvement,” the subjectivity scoring would recognize “significant” and “clearly” as modifiers that introduce a subjective tone, potentially resulting in a higher subjectivity score. However, for a sentence such as “The data indicate a 5% increase in performance,” a lower subjectivity score would be assigned because “indicate” and “data” signal factual reporting.

This approach to subjectivity combines **lexicon-based** analysis and **rule-based heuristics** to distinguish between factual and opinion-based language. By leveraging pre-labeled subjectivity lexicons and adjusting scores based on linguistic context, this method produces a nuanced subjectivity score that aligns closely with how objective or subjective each sentence is likely to be. This methodology serves as a reliable tool for analyzing the tone of textual content, providing insights into the balance of factual and evaluative language in documents [18].

4 Results

4.1 Descriptives

Table 1 presents descriptive statistics for the key variables in this study, based on 3,347 observations. The variable “*Auditfirm*” denotes a dummy variable whether a report received third-party assurance from an audit firm, with a mean of 0.1264, indicating that most reports were not assured by audit firms. The mean sentiment score (*Sentiment_bert*), derived from BERT-based sentiment analysis, is 0.6613, with a median of 0.5, reflecting generally positive sentiment. The subjectivity score (*Subjectivity*), with a mean of 0.1710, suggests that most reports leaned toward factual content, though some variability is evident.

Table 2 illustrates the distribution of sustainability reports over the years, with a general increase over the past decade, indicating growing engagement with issuing sustainability reports as time goes by. Figure 1 displays the proportional distribution of assurance types, showing that limited assurance is the most prevalent, followed by reports with no assurance, and finally, reasonable assurance. This distribution

Table 1 Descriptive statistics

Variables	N	Mean	Median	S.D.	Min	Max
<i>Auditfirm</i>	3,347	0.1264	0.0000	0.3323	0.0000	1.0000
<i>Sentiment_bert</i>	3,347	0.6613	0.5000	0.2607	0.0000	1.0000
<i>Subjectivity</i>	3,347	0.1710	0.1933	0.0967	0.0000	0.3412

Table 2 Sample distribution by year

Year	Freq.	Percent
2013	95	2.84
2014	107	3.20
2015	134	4.00
2016	168	5.02
2017	216	6.45
2018	281	8.40
2019	348	10.40
2020	439	13.12
2021	476	14.22
2022	617	18.43
2023	466 ¹	13.92
Total	3,347	100

provides context for understanding the varying levels of assurance and their potential impact on narrative sentiment and subjectivity.

In Table 3, the industry with the highest frequency of sustainability reports within the sample is “Business Services,” representing 8.81% of the total observations. This is followed by “Chemicals and Allied Products” and “Holding and Other Investment Offices,” each contributing notably to the dataset’s representation.

4.2 Sentiment Analysis

Table 4 presents the results of ANOVA and *t*-tests on the sentiment of sustainability report narratives, calculated using a BERT-based sentiment analysis. This analysis tests the hypothesis that reports without third-party assurance would exhibit more positive sentiment compared to those with assurance. This expectation is grounded in the observation that sustainability reports are typically voluntary disclosures, often leading companies to emphasize positive aspects of their performance while omitting

¹ Sustainability reports are published in the year following the fiscal year, and the month of publication varies by company. As we collected data as of August 2024, sustainability reports for 2023 fiscal year published after this date may not be included.

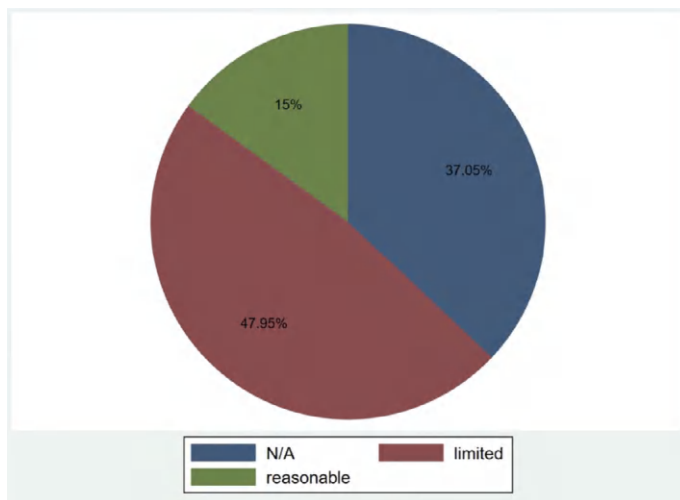


Fig. 1 Sample distribution by level of assurance

negative information. Such selective reporting practices can, if unchecked, lead to greenwashing, where companies present an overly favorable image of their environmental practices, potentially misleading stakeholders (Lyon and Montgomery, 2015 [19]). Additionally, the analysis hypothesized that reports assured by audit firms would exhibit more neutral and complete narratives compared to those assured by non-audit firms, resulting in less positive sentiment.

The results support these hypotheses, as all tests showed statistically significant differences between the groups. Panel a uses ANOVA to compare sentiment across three groups: reports with no assurance, reports with limited assurance, and reports with reasonable assurance. The results indicate significant differences in sentiment among the groups, with reports lacking assurance showing a higher positive sentiment. Panel B compares sentiment between limited assurance and reasonable assurance groups, excluding reports without assurance. This t-test also shows a statistically significant difference, highlighting variability in sentiment based on assurance level. Panel C compares reports assured by audit firms versus non-audit firms, demonstrating that reports assured by audit firms exhibit less positive sentiment. This supports the idea that audit-firm-assured reports maintain more neutrality and completeness compared to those assured by non-audit firms, as predicted.

4.3 Subjectivity Analysis

Table 5 summarizes the results of tests comparing the subjectivity and factual nature of sustainability report narratives based on whether they received third-party assurance and the type of institution providing the assurance. The analysis is based on

Table 3 Sample distribution by industry

Industry name	2-digit SIC	Freq.	Percent
Agriculture, forestry, and fishing	01	3	0.09
Metal mining	10	45	1.34
Coal mining	12	17	0.51
Oil and gas extraction	13	176	5.26
Mining and quarrying of nonmetallic minerals	14	13	0.39
General building contractors	15	12	0.36
Heavy construction, except building	16	15	0.45
Special trade contractors	17	2	0.06
Food and kindred products	20	153	4.57
Tobacco products	21	16	0.48
Textile mill products	22	10	0.3
Apparel and other textile products	23	13	0.39
Lumber and wood products, except furniture	24	6	0.18
Furniture and fixtures	25	26	0.78
Paper and allied products	26	71	2.12
Printing, publishing, and allied industries	27	2	0.06
Chemicals and allied products	28	295	8.81
Petroleum refining and related industries	29	31	0.93
Rubber and miscellaneous plastics products	30	51	1.52
Leather and leather products	31	7	0.21
Stone, clay, glass, and concrete products	32	13	0.39
Primary metal industries	33	29	0.87
Fabricated metal products	34	32	0.96
Industrial and commercial machinery and computer equipment	35	202	6.04
Electronic and other electrical equipment	36	186	5.56
Transportation equipment	37	104	3.11
Measuring, analyzing, and controlling instruments	38	123	3.67
Miscellaneous manufacturing industries	39	12	0.36
Railroad transportation	40	25	0.75
Local and suburban transit and interurban highway passenger	41	4	0.12
Motor freight transportation and warehousing	42	17	0.51

(continued)

Table 3 (continued)

Industry name	2-digit SIC	Freq.	Percent
Water transportation	44	45	1.34
Transportation by air	45	48	1.43
Pipelines, except natural gas	46	3	0.09
Transportation services	47	13	0.39
Communications	48	41	1.22
Electric, gas, and sanitary services	49	244	7.29
Wholesale trade and durable goods	50	36	1.08
Wholesale trade and nondurable goods	51	33	0.99
Building materials, hardware, garden supply, and mobile home dealers	52	5	0.15
General merchandise stores	53	16	0.48
Food stores	54	10	0.3
Automotive dealers and gasoline service stations	55	8	0.24
Apparel and accessory stores	56	31	0.93
Eating and drinking places	58	33	0.99
Miscellaneous retail	59	39	1.17
Depository institutions (Banks)	60	77	2.3
Nondepository institutions	61	28	0.84
Security and commodity brokers	62	70	2.09
Insurance carriers	63	91	2.72
Insurance agents, brokers, and service	64	14	0.42
Real estate	65	35	1.05
Holding and other investment offices	67	264	7.89
Hotels, rooming houses, camps, and other lodging places	70	35	1.05
Personal services	72	3	0.09
Business services	73	295	8.81
Automotive repair, services, and parking	75	15	0.45
Motion pictures	78	4	0.12
Amusement and recreation services	79	32	0.96
Health services	80	36	1.08
Educational services	82	3	0.09
Engineering, accounting, research, management, and related services	87	24	0.72
Nonclassifiable establishments	99	5	0.15
Total		3,347	100

Table 4 The results of ANOVA and T-Tests of sentiment analysis by groups

Panel A: Assurance type (ANOVA)				
Group/Level	Mean	Std. dev	Freq.	Statistical comparison
N/A	0.6913	0.2691	2,184	F = 43.52, p < 0.001
limited assurance	0.6083	0.2395	942	
Reasonable assurance	0.5894	0.2083	221	
Panel B: Assurance type (t-test for Limited versus Reasonable)				
Group/Level	Mean	Std. dev	Freq.	Statistical comparison
Limited assurance	0.6083	0.2395	942	Mean Diff = 0.0189 $t = 1.082$ $p = 0.2795$
Reasonable assurance	0.5894	0.2083	221	
Panel C: Assurance by audit firm (Subsample Comparison)				
Group/Level	Mean	Std. dev	Freq.	Statistical comparison
Assurance by non-audit firm (0)	0.6154	0.2404	832	Mean Diff = 0.0376 $t = 2.479$ $p = 0.0133$
Assurance by audit firm (1)	0.5778	0.2154	331	

the hypothesis that reports with third-party assurance would be less subjective and more factual compared to those without assurance. This expectation aligns with findings that “third-party assurance enhances the credibility of sustainability reports, leading to more balanced and objective disclosures” [20]. Additionally, the type of assurance provider can influence report quality; for instance, “assurance provided by accounting firms is associated with higher quality disclosures compared to non-accounting providers” [21]. These insights support the hypothesis that third-party assurance, particularly from reputable institutions which possess the specialty to provide expert assurance service, contributes to more factual and less subjective sustainability reporting.

Panel A presents an ANOVA comparing the subjectivity of narratives across three groups: reports with no assurance, reports with limited assurance, and reports with reasonable assurance. The results show that there is a statistically significant difference in subjectivity among these three groups. Panel B focuses on a t-test that compares limited assurance and reasonable assurance levels, excluding the sample with no assurance from Panel A. In this comparison, the difference between the two groups was not statistically significant. Panel C summarizes the t-test comparing reports with third-party assurance provided by audit firms versus non-audit firms. The results indicate a statistically significant difference, suggesting that reports assured by audit firms tend to be more factual and less subjective compared to those assured

Table 5 The results of ANOVA and T-Tests of subjectivity analysis by groups

Panel A: Assurance type (ANOVA)				
Group/Level	Mean	Std. dev	Freq.	Statistical comparison
N/A	0.1746	0.0994	2,184	$F = 4.40$ $p = 0.0124$
limited assurance	0.1647	0.0931	942	
Reasonable assurance	0.1626	0.0878	221	
Panel B: Assurance type (t-test for Limited versus Reasonable)				
Group/Level	Mean	Std. dev	Freq.	Statistical comparison
Limited assurance	0.1647	0.0931	942	Mean Diff = 0.0021 $t = 0.315$ $p = 0.7526$
Reasonable assurance	0.1626	0.0878	221	
Panel C: Assurance by audit firm (Subsample Comparison)				
Group/Level	Mean	Std. dev	Freq.	Statistical comparison
Assurance by non-audit firm (0)	0.1688	0.0900	832	Mean Diff = 0.0156 $t = 2.687$ $p = 0.0073$
Assurance by audit firm (1)	0.1531	0.0886	331	

by non-audit firms. These findings suggest that the existence and provider of third-party assurance influence the subjectivity and factual nature of sustainability report narratives.

5 Conclusion

The findings of this study reveal meaningful patterns in the sentiment and subjectivity of voluntary sustainability reports, driven by the presence and type of third-party assurance. Reports without assurance often exhibit a more positive sentiment, potentially reflecting a focus on unmitigated corporate achievements. In contrast, reports with limited or reasonable assurance show differentiated tones, suggesting that the assurance process influences how companies communicate their commitments and challenges. Furthermore, reports assured by audit firms demonstrate a more neutral and comprehensive narrative compared to those assured by non-audit firms, indicating an emphasis on factuality, faithful representation, and completeness that aligns with regulatory expectations.

This research contributes to the broader discourse on ESG reporting by demonstrating that assurance is more than a formal requirement; it shapes the narrative

integrity and perceived credibility of corporate sustainability disclosures. As regulatory landscapes evolve and companies increasingly seek to validate their ESG commitments, understanding the implications of assurance practices on sentiment and subjectivity can provide strategic insights for practitioners and policymakers. Future research could further explore how different assurance practices impact specific industries and geographic regions, enriching the dialogue on transparent and accountable ESG reporting.

Acknowledgements Ethan Jaesuh Kim provided excellent research assistance.

References

1. Cuomo F, Mallin C, Zattoni A (2022) The effects of the EU non-financial reporting directive on corporate social responsibility. *Eur J Financ* 30(7):726–752
2. Inspire (2024) The effects of mandatory ESG disclosure around the world. <https://inspiregreenfinance.org/publications/the-effects-of-mandatory-esg-disclosure-around-the-world/>
3. Deloitte (2021) Tectonic shifts: How ESG is changing business, moving markets, and driving regulation. Deloitte Insights. <https://www2.deloitte.com/us/en/insights/topics/strategy/esg-disclosure-regulation.html>
4. Azeus Convene (2022) The global state of mandatory ESG disclosures. <https://www.azeusconvene.com/articles/the-global-state-of-mandatory-esg-disclosures/>
5. Krueger P, Sautner Z, Starks LT (2024) The effects of mandatory ESG disclosure around the world. *J Account Res* (Forthcoming)
6. Du S, Wu H (2019) Does external assurance enhance the credibility of CSR reports? Evidence from CSR-related misconduct events in Taiwan. *Audit: J Pract Theory* 38(4):101–130
7. Maroun W (2019) Does external assurance contribute to higher quality integrated reports? *J Account Public Policy* 38(4):106670
8. Asante-Appiah E, Lambert C (2023) The role of the external auditor in managing environmental, social, and governance (ESG) reputation risk. *Rev Acc Stud* 28:2589–2641
9. Liu B (2012) Sentiment analysis and opinion mining. San Rafael, CA: Morgan and Claypool
10. Pang B, Lee L (2008) Opinion mining and sentiment analysis. *Found Trends® Inf Retr* 2(1–2):1–135
11. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781)
12. Devlin J, Chang MW, Lee K, Toutanova K (2019) 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. Association for Computational Linguistics, pp 4171–4186
13. Yang X, Liu Q, Zhang H, Tang Z (2020) FinBERT: a pre-trained financial language representation model for financial sentiment analysis. In Proceedings of the 29th ACM international conference on information and knowledge management. Association for computing machinery, pp 2621–2628
14. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Amodei D (2020) Language models are few-shot learners. *Adv Neural Inf Process Syst* 33:1877–1901
15. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 21(140):1–67
16. Wiebe J, Wilson T, Cardie C (2004) Annotating expressions of opinions and emotions in language. *Lang Resour Eval* 39(2):165–210

17. Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04). Association for computational linguistics, pp 271–278
18. Loria S (2018) TextBlob: simplified text processing. Retrieved from <https://textblob.readthedocs.io/>
19. Lyon TP, Montgomery AW (2015) The means and end of greenwash. *Organ Environ* 28(2):223–249. <https://doi.org/10.1177/1086026615575332>
20. Hodge K, Subramaniam N, Stewart J (2009) Assurance of sustainability reports: impact on report users' confidence and perceptions of information credibility. *Aust Account Rev* 19(3):178–194. <https://doi.org/10.1111/j.1835-2561.2009.00056.x>
21. Simnett R, Vanstraelen A, Chua WF (2009) Assurance on sustainability reports: an international comparison. *Account Rev* 84(3):937–967

Verbal Femininity and CEOs Compensation



Sang-Joon Kim and Juil Lee

Abstract This study investigates how gender biases among board directors influence CEO compensation, focusing on the role of male CEOs' verbal femininity. Given that verbal expression is another essential way to reveal feminine characters of CEOs, the femininity saliently cued in CEOs' verbalizing habits can bring more distorted perception, which leads to the boards' evaluation on their contributions. This is because verbal femininity, characterized by nurturing, empathetic, benevolent, and collaborative communication styles, is argued to conflict with traditional masculine stereotypes of leadership. Drawing on role congruity theory, we contend that CEOs who exhibit verbal femininity are undervalued in compensation decisions due to biases against traits perceived as incongruent with the prototype of effective leadership. In this study, we specify the feminine characteristics which induce decision errors by using a Large Language Model (LLM). Acknowledging that LLMs are subject to pre-existing social prejudices and gender biases, we utilize this attribute of the models to develop a way to capture gender biases (especially femininity). In this study, we employ an LLM-driven algorithm to measure the extent to which a certain CEO's words in official settings show femininity. Using this measure, we examine how the CEO's verbal femininity can affect the assessment of CEO quality in the boardroom, determining the level of CEO compensation. Our findings contribute to the broader discourse on implicit gender biases in determining CEO compensation.

Keywords Verbal femininity · CEO compensation · Assessment of CEO quality · Board decision · Large language models

S.-J. Kim

College of Business Administration, Ewha Womans University, Seoul, Korea

e-mail: s.kim@ewha.ac.kr

J. Lee (✉)

School of Business, Chungnam National University, Daejeon, Korea

e-mail: juil@cnu.ac.kr

1 Introduction

CEO compensation is a critical aspect of corporate governance, reflecting not only the perceived value of an executive's contributions but also the implicit biases that may influence decision-making within boardrooms [1–4]. While compensation decisions are often linked to objective performance metrics, subjective factors, such as a CEO's communication style, are increasingly recognized as shaping these evaluations [2, 5]. Among these, verbal femininity—a communication style characterized by nurturing, empathetic, benevolent, and collaborative traits—offers a unique lens through which to explore how traditional gender biases impact compensation outcomes.

Verbal femininity challenges the established norms of leadership evaluation, which are deeply rooted in traditional masculine stereotypes emphasizing assertiveness, decisiveness, and control [6–9]. Role congruity theory suggests that these stereotypes create implicit expectations for effective leadership [8, 10, 11], which may conflict with the qualities associated with verbal femininity. As a result, CEOs who exhibit verbal femininity are often undervalued in board evaluations, particularly when gender biases are prevalent among directors. This undervaluation has direct implications for compensation decisions, raising critical questions about fairness and inclusivity in corporate governance.

Existing research on CEO compensation has extensively examined factors such as financial performance, firm characteristics, and CEO demographics [12]. However, the role of communication styles—especially those that reveal feminine characteristics—remains underexplored. Moreover, while studies have acknowledged the presence of gender biases in board evaluations [2, 5, 8, 13–18], they have not directly examined how such biases interact with specific traits of CEOs, like verbal femininity, to influence compensation outcomes. Addressing these gaps is essential for understanding the subtle yet pervasive ways in which stereotypes shape decision-making processes in determining CEO compensation.

This study aims to fill these gaps by examining the relationship between CEO verbal femininity and CEO compensation, with a specific focus on gender biases among board directors. To measure verbal femininity, we employ a novel methodology using a Large Language Model (LLM)-driven algorithm that quantifies the feminine characteristics of CEOs' verbal expressions in official settings. Recognizing that LLMs themselves are subject to pre-existing social prejudices and gender biases (e.g., [19]), this study leverages these attributes of the models to capture gender biases in the capital market and identify how femininity cues in CEOs' verbal habits influence board evaluations, ultimately determining the level of CEO compensation. Our findings reveal how verbal femininity, as a signal, can trigger distorted perceptions among biased board directors, leading to undervaluation in compensation decisions.

By uncovering the implicit penalties associated with verbal femininity, this study contributes to the broader discourse on gender biases in determining CEO compensation. The results highlight the need for boards to address these biases to ensure equitable compensation practices, while also advancing theoretical insights into the intersection of communication styles, gender roles, and leadership evaluations.

2 Theory and Hypothesis

2.1 *Cognitive Biases Through the Lens of Gender Role Incongruity*

Gender stereotypes play a critical role in shaping leadership evaluations and compensation decisions, particularly in male-dominated organizational contexts [8]. According to role congruity theory, women are often associated with communal traits such as empathy, collaboration, and nurturance, while men are linked to agentic qualities like assertiveness, independence, and decisiveness. These stereotypes create a perceived incongruity between the communal traits ascribed to women and the agentic qualities typically associated with effective leadership roles. This incongruity leads to biased evaluations of female leaders, as they are perceived as less aligned with the expectations of leadership positions [8, 18].

One critical channel through which such biases manifest is the verbal communication style of CEOs. Verbal femininity, which encompasses nurturing, empathetic, benevolent, and inclusive communication, may signal a leadership style that emphasizes collaboration and relationship-building. While these qualities can be advantageous in certain organizational contexts, they may also be interpreted by board directors as indicative of a lack of authority or decisiveness. This perception arises from entrenched stereotypes that equate effective leadership with traditionally masculine communication styles, such as directness, assertiveness, and goal orientation.

Boards, as the primary evaluators of CEO performance, are not immune to these biases. When board directors interpret verbal femininity through the lens of role incongruity, they may undervalue the CEO's leadership effectiveness. For instance, feminine verbal styles may be viewed as misaligned with the agentic qualities required for navigating high-stakes corporate environments, thereby diminishing the perceived competence and strategic acumen of the CEO. This bias can result in lower evaluations of CEO performance.

According to Shin and You [20], a CEO's language serves as a signal to the board regarding their leadership qualities and strategic priorities. In male-dominated corporate cultures, signals of verbal femininity may be interpreted as weakness or an inability to assert control, reinforcing existing stereotypes. These biased perceptions not only affect the immediate evaluation of the CEO's contributions but also shape long-term decisions regarding leadership succession and strategic alignment.

The persistence of such biases underscores the challenges faced by CEOs who deviate from traditional leadership prototypes. Verbal femininity, while potentially a strength in fostering collaboration and stakeholder engagement, is often devalued in contexts where agentic qualities are prioritized. This dynamic illustrates how role incongruity perpetuates gender bias in leadership evaluations, particularly in settings where stereotypical views of effective leadership remain dominant.

By framing board-level biases through the lens of gender role incongruity, this study aims to elucidate how a CEO's feminine communication styles affect CEO evaluations and compensation decisions.

2.2 Verbal Femininity and CEO Compensation

Compensation decisions for CEOs are influenced by multiple factors, including organizational performance, leadership style, and board perceptions. This is because the difficulty in measuring intangible aspects of leadership effectiveness and the influence of external factors on firm outcomes make it challenging to objectively evaluate CEO performance [21, 22]. As a result, board directors often rely on heuristics and cognitive shortcuts, such as symbolic cues and social signals, to form their judgments [20, 23–25].

In this respect, prior studies have examined that heuristic cues such as vocal and facial signals can significantly influence board evaluations, particularly in contexts where there is limited information about a CEO's actual performance [20]. Indeed, Nair et al. [2] demonstrate that verbal masculinity, characterized by assertive and dominant language, is positively associated with CEO compensation. This is because physical strength was a key indicator of leadership effectiveness, psychological mechanisms evolved over time to link deeper, more masculine vocal tones with perceptions of leadership capability [26, 27]. Similarly, Gupta and Wowak [28] show that deeper, more masculine vocal tones are associated with authority and confidence, attributes that are valued in leadership roles. These imply that biased perceptions significantly affect how corporate leaders are perceived and the decisions they make [29]. That is, gendered biases rooted in societal stereotypes significantly shape how boards evaluate CEO effectiveness, particularly for traits or behaviors that deviate from traditional leadership norms.

According to role congruity theory [8], verbal femininity may signal traits that are incongruent with the agentic expectations of effective leadership. Building on the lens of gender role incongruity, we argue that directors' biased perceptions of verbal femininity undermine CEO evaluations and compensation. This is because gender stereotypes create a perceived incongruity between communal traits, commonly associated with women, and the agentic characteristics expected of effective leaders. Traditional leadership roles emphasize decisiveness, authority, and assertiveness—qualities historically aligned with masculine stereotypes. This incongruity can lead to negative evaluations of leaders who exhibit feminine traits, as they are perceived to lack the assertiveness required for high-stakes decision-making and strategic leadership. These biases can result in less favorable evaluations of CEO effectiveness, ultimately influencing compensation decisions. In sum, we contend that CEOs exhibiting verbal femininity are likely to face biases that devalue their leadership qualities, resulting in lower compensation. Thus, the suggested hypothesis is as follows:

Hypothesis. Male CEOs' verbal femininity is negatively related to the level of CEO compensation.

3 Method

3.1 Settings

To evaluate our argument that CEO compensation is influenced by verbal femininity, we base our research on three key assumptions. First, the sample firms are publicly traded. Public companies are required to have a board of directors responsible for decision-making, including setting CEO compensation. Additionally, since public firms separate management from ownership, decisions regarding CEO pay can be made independently and at the board's discretion. Second, CEOs are gendered in this study. Specifically, all CEOs in the sample firms are male. Female CEOs are excluded because the feminine traits of male CEOs are more likely to be perceived as incongruent with traditional gender roles, making them more pronounced in the analysis. Third, the verbal representations of CEOs are socially constructed. Verbal femininity is not considered an innate characteristic of a CEO but rather a socially constructed trait that evolves over time (e.g., [30, 31]). For empirical analysis, this means the sample firms have recorded CEOs' verbal expressions across multiple time periods, all within consistent institutional contexts.

Based on these assumptions, we selected firms listed on the U.S. capital market as our sample. We focused on this market because firm behaviors are institutionalized there, with shared market-related practices, such as earnings conference calls, aimed at enhancing firm value (e.g., [32]). From the publicly traded U.S. firms, we sampled those where a male CEO had served across multiple time periods.

To construct our sample, we utilized several databases, including Compustat, Execucomp, ISS (formerly RiskMetrics), and Seeking Alpha. First, we identified firms where a male CEO had served across multiple time periods (i.e., more than 3 years) [33–35]. Next, we combined the Compustat, Execucomp, and ISS datasets using firm and year identifiers to compile accounting and financial data, CEO information, and board data. From this merged dataset, we generated a list of sample firms. Using the identified list of firms, we retrieved earnings conference call transcripts for each firm from Seeking Alpha. Firms that did not hold any earnings conference calls between 2006 and 2020 were excluded from the sample. The time frame was truncated at 2020 to mitigate the impact of the economic disruption caused by COVID-19. This process resulted in a final dataset comprising 117 firms and 516 firm-years.

3.2 Measures

Dependent variables. The dependent variable in this study is CEO compensation, measured annually as the total sum of base salary, bonuses, restricted stock, grants, LTIP payouts, option grants, and other annual rewards received by the CEO [36–38]. Specifically, we utilized the TDC1 variable from Execucomp to represent

this measure. To address its right-skewed distribution, we applied a logarithmic transformation to the original variable.

Independent variable. The independent variables for our hypothesis are verbal femininity and verbal masculinity, which together represent culturally ingrained traits associated with gender roles and expressed through language (e.g., [39–41]). Verbal femininity refers to linguistic expressions that align with traits culturally deemed desirable for women, such as warmth, empathy, collaboration, and sensitivity. These traits are deeply rooted in societal gender norms and are reflected in communication patterns that prioritize relational and affiliative language [42–44]. For example, verbal femininity may manifest in speech through the frequent use of supportive phrases, inclusive language, and expressions of concern for others [39–41]. On the other hand, verbal masculinity refers to linguistic expressions that align with traits culturally associated with men, such as assertiveness, independence, dominance, and competitiveness [39]. These traits are reflected in communication patterns that emphasize control, authority, and task-oriented language [39, 42, 45]. Verbal masculinity may be evident in direct, commanding speech, the use of declarative statements, and language that projects confidence and certainty [39–41].

Given that our sample consists of male CEOs, we conceptualize CEOs' verbal expressions of femininity and masculinity as complementary and intertwined dimensions. To capture verbal femininity, as well as verbal masculinity, we employed a large language model. Large language models (LLMs) are advanced machine learning systems trained on vast amounts of text data to understand and generate human-like language [46]. These models leverage transformer-based architectures to capture semantic, syntactic, and contextual information in text [47–49]. By pre-training on diverse datasets and fine-tuning for specific tasks, LLMs demonstrate remarkable versatility, enabling applications such as text generation, translation, and classification [50]. Building on the capabilities of LLMs, zero-shot classification is a method that uses similarities with previously learned data to categorize new, unseen classes without requiring training examples for those classes [51]. It leverages pre-trained language models to identify semantic features in text and apply them to classification tasks. The key strength of zero-shot classification lies in its ability to predict novel classes not explicitly represented in the training data, making it particularly valuable for datasets that are difficult to label [51].

In this study, we employed zero-shot classification to quantify verbal femininity and verbal masculinity in earnings call transcripts. To represent these constructs, we defined distinct sets of labels corresponding to feminine and masculine verbal attributes [39–42]. For verbal femininity, the labels were “Collaborative,” “Supportive,” “Understanding,” “Empathetic,” and “Soft.” For verbal masculinity, the labels included “Aggressive,” “Assertive,” “Strong,” “Dominant,” and “Decisive.” The transcripts of earnings conference calls were gathered annually to assess verbal masculinity and femininity using the specified labels. For years with multiple earnings calls, all transcripts were consolidated into a single set. Since the transcripts include dialogue from various participants in the conference calls (e.g., analysts, other executives, and moderators), we specifically identified the portions of the transcripts attributed to the CEO and extracted their spoken contents for analysis.

Based on the extracted texts of each sample firm for the given year, we employed a pre-trained transformer model for zero-shot classification to compute similarity scores between the texts and each label. These scores, ranging from 0 to 1, indicate the degree to which the CEO's verbal expressions semantically align with the label [51]. To measure verbal femininity and verbal masculinity, we calculated the aggregated score for each construct as the product of the similarity scores for the five corresponding labels. This method emphasizes the simultaneous presence of all related verbal attributes, as a low score on any label will significantly reduce the overall aggregated score. This approach provides a more stringent measure of the extent to which CEOs' verbal expressions exhibit strongly feminine or masculine characteristics. Using multiplication ensures that the aggregated score reflects consistency across all labels, giving more weight to instances where multiple attributes align strongly with either femininity or masculinity.

Control variables. To account for factors influencing CEO compensation, we included control variables at the industry, firm, and board levels. For industry-level controls, industry size, environmental munificence, and environmental dynamism were measured. Industry size was calculated the aggregate total assets of all firms within a given industry, identified using the 3-digit SIC code. This metric captures the overall scale of the industry. Environmental munificence was defined as the abundance or scarcity of critical resources available to firms within an environment [52]. This was operationalized as the growth rate of the industry in which the focal firm operates. Environmental munificence is widely recognized as an essential environmental condition that facilitates firm survival and growth by providing ample resources [53–56]. On the other hand, environmental dynamism refers to “to the volatility and unpredictability of a firm's external environment” [57]. We measured this by calculating the variation in industry revenues over a five-year period. Environmental dynamism reflects the degree of uncertainty and change within the industry, potentially impacting strategic decision-making and performance. These two industry-level variables were incorporated into the regression model to control for external environmental factors that could influence CEO compensation. The specific measures for environmental munificence and dynamism are detailed in the regression equation below. This approach ensures a comprehensive understanding of how both resource availability and environmental volatility shape CEO pay dynamics.

$$R_{it} = \alpha + \beta R_{it-1} + \varepsilon_{it}, \quad (1)$$

where R_{it} represents the revenues of industry j (identified using the 3-digit SIC code) over the past five years leading up to time t . The parameter β denotes the growth rate of industry revenue and is used as a measure of environmental munificence. Meanwhile, the standard error of β serves as an indicator of environmental dynamism [54–56].

Firm-specific characteristics were defined using several key variables: market share, firm size, ROA, sales growth, current ratio, debt-equity ratio, R&D intensity, and marketing intensity. Market share was calculated as the proportion of a firm's sales relative to the total sales in the industry. Firm size was proxied by the dollar amount of total assets (in thousands), with a logarithmic transformation applied

to address skewness in the data. ROA was calculated as the ratio of net income to total assets, providing an indicator of operational efficiency. Sales growth was calculated as the change in sales divided by the previous year's sales, representing the firm's growth rate. Current ratio was calculated as the ratio of current assets to current liabilities, a measure of a firm's short-term liquidity [58, 59]. Debt-equity ratio is computed as the total value of short-term and long-term debt divided by the equity value, reflecting the firm's leverage [59]. R&D intensity was computed as the ratio of research and development expenditures to total assets, indicating the firm's investment in innovation. Marketing intensity is defined as the ratio of administrative expenditures to total assets, capturing the firm's focus on marketing and administrative activities.

Finally, the board-level control variables include the proportion of outsider directors and the number of Caucasian male directors on the board. The outsider-directors variable is measured as the proportion of directors on the board who are not part of the firm. The Caucasian-male-directors variable represents the total number of white male directors serving on the board. Table 1 presents the descriptive statistics of the variables that are used in the empirical analysis.

3.3 *Estimation Model*

To investigate the relationship between verbal femininity and CEO compensation, we employed a fixed effects model, chosen based on the results of the Hausman test [60]. The Hausman test evaluates whether the fixed effects model provides more consistent estimates than the random effects model. When the test statistic yields a p-value below the 0.05 significance threshold, it indicates the fixed effects model is the superior choice. In this study, the Hausman statistic was 25.30 ($p < 0.05$), confirming the appropriateness of the fixed effects approach for our analysis.

In addition, to address potential selection bias in the sampled firms, we accounted for sample selection bias by incorporating the Inverse Mills Ratio. This correction began with a Probit model, which estimated the likelihood of a firm being included in the sample from the entire Compustat dataset. The Probit model considered firm size (proxied by the number of employees) and sales normalized by total assets as predictors. The resulting Inverse Mills Ratio was then included in the fixed effects model to ensure robust and unbiased estimation.

Table 1 Descriptive statistics

Variables (N = 516)	Mean	SD	Min	Max	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. In CEO compensation	9.19	0.87	2.38	12.54															
2. Verbal masculinity	0.1	0.06	0	0.45	0.06														
3. Verbal femininity	0.09	0.05	0	0.3	0.02	0.74													
4. Industry size	7.67	13.08	0.54	88.77	0.13	-0.11	-0.05												
5. Environmental munificence	1.01	0.09	0.49	1.29	-0.02	0.01	0.11	-0.12											
6. Environmental dynamism	0.01	0.01	0	0.09	-0.16	-0.01	0.04	0.24	-0.04										
7. Market share	0.14	0.18	0	1	0.14	0.06	0.13	0	0.19	0.39									
8. In total assets	9.73	1.43	0	12.84	0.43	0	0.03	0.34	-0.05	0.03	0.24								
9. ROA	0.07	0.09	-1.1	0.5	-0.03	-0.02	0.09	-0.02	0.15	0.08	0.12	-0.08							
10. Sales growth	11.23	253.26	-0.44	5753	0.02	0.03	0.07	-0.01	-0.04	-0.01	-0.03	-0.3	-0.03						
11. Current ratio	1.82	1.36	0	9.59	-0.1	-0.08	-0.05	-0.19	-0.01	-0.25	-0.23	-0.28	0.2	0.01					
12. Debt-equity ratio	-0.32	35.38	-776.59	143.99	0.02	0.03	0.02	-0.03	-0.02	-0.02	0.01	-0.02	0.01	0	0.02				
13. R&D intensity	0.04	0.07	0	0.4	-0.07	0.04	-0.02	-0.25	0.07	-0.33	-0.32	-0.37	0.02	-0.03	0.45	0.02			
14. Marketing intensity	0.23	0.19	0	1.03	-0.03	0.05	-0.03	-0.33	0.01	-0.32	-0.22	-0.3	-0.05	-0.05	0.23	0.03	0.53		
15. % Outsider directors	0.04	0.18	0	0.92	0.01	0.04	-0.01	-0.07	0.03	-0.16	-0.15	0.04	0.01	-0.01	0.08	0.01	0.07	0.1	
16. % White-male directors	0.75	0.44	0	1	-0.04	-0.12	-0.03	0.05	-0.02	0.1	0.12	0.17	0.05	-0.08	0.03	0.04	-0.15	-0.2	-0.4

CEO compensation is measured as the total compensation the give CEO is given at time t. The total compensation includes salary, bonus, restricted stock, grants, LTIP payouts, option grants, and all other annual rewards. Since the distribution of CEO compensation in our sample is right-skewed, we take a logarithm to the variable. Verbal masculinity is captured through zero-shot classification by using the labels of “aggressive”, “assertive”, “strong”, “dominant”, and “decisive.” For verbal femininity, we employ the same LLM model with the labels of “collaborative”, “supportive”, “understanding”, “empathetic”, and “soft”

4 Results

4.1 Verbal Femininity and CEO Compensation

Table 2 presents the fixed-effects estimations of CEO compensation in relation to verbal femininity and verbal masculinity. Model 1 includes only the control variables, while Models 2 through 4 incorporate the focal predictors: verbal femininity and verbal masculinity. In Model 4, the results indicate that verbal femininity significantly reduces CEO compensation ($\beta = -2.6967$; $p = 0.003$), whereas verbal masculinity significantly increases it ($\beta = 1.6939$; $p = 0.024$). In the contexts of male CEOs, verbal masculinity aligns with traditional gender roles, and this congruence leads to more favorable evaluations by the board. In contrast, when a male CEO exhibits feminine traits in his narratives, this is perceived as inconsistent with gender-based expectations, resulting in less favorable evaluations from the board.

While our arguments are grounded in the theory of gender role incongruity [8, 61], we recognize that verbal femininity and verbal masculinity are not mutually exclusive. Instead, they coexist within individuals in unique androgynous configurations [42, 62–65]. To account for these androgynous aspects of CEO narratives, we identified two distinct communication styles: masculinity-focused narratives and femininity-focused narratives. Masculinity-focused narratives indicate a stronger tendency toward masculine traits compared to feminine ones, while femininity-focused narratives reflect a greater emphasis on feminine traits relative to masculinity. To examine how these communication styles influence board evaluations, as reflected in CEO compensation, we developed two variables: verbal masculinity focus and verbal femininity focus. These variables measure the extent to which verbal masculinity or verbal femininity predominates in a CEO's communication. Operationally, we calculated these variables using the measures of verbal masculinity and verbal femininity, as detailed below

$$\begin{aligned} VF_{it}^M &= VF_{it} \text{ if } VF_{it} > 0; 0 \text{ else} \\ VF_{it}^F &= |VF_{it}| \text{ if } VF_{it} < 0; 0 \text{ else} \end{aligned} \quad \text{s.t. } VF_{it} = \frac{M_{it}}{(M_{it} + F_{it})} - \frac{F_{it}}{(M_{it} + F_{it})} \quad (2)$$

where VF_{it}^M and VF_{it}^F denote verbal masculinity focus and verbal femininity focus respectively; VF_{it} indicates which a given CEO's narratives focus more on between masculinity and femininity; M_{it} and F_{it} show verbal masculinity and verbal femininity, computed through the LLM with the zero-shot classification. As seen in Model 5 in Table 2, we found the results consistent. That is CEO's communication style that is dominated by masculine traits (i.e., verbal masculinity focus) tends to receive more favorable evaluations from the board ($\beta = 0.4018$; $p = 0.074$). In contrast, communication emphasizing feminine traits (i.e., verbal femininity focus) is associated with less favorable board evaluations ($\beta = -0.5821$; $p = 0.076$).

From the perspective of gender role congruity, we found that male CEOs who exhibit feminine traits in their communication may face unfavorable evaluations

Table 2 Fixed-Effects estimation of CEO compensation with respect to verbal masculinity and verbal femininity

	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	7.2748***	7.2685***	7.4474***	7.3282***	7.1190***
	−0.9742	−0.9728	−0.9752	−0.964	−0.9738
Industry size	0.0146	0.0147	0.0148	0.017	0.0182
	−0.0137	−0.0137	−0.0137	−0.0138	−0.014
Environmental munificence	1.2603*	1.2607*	1.2260*	1.2067*	1.2451*
	−0.5742	−0.5738	−0.5751	−0.5697	−0.574
Environmental dynamism	9.9363	9.9355	9.5257	8.9646	10.49
	−7.5574	−7.5648	−7.5357	−7.3984	−7.5144
Market share	−0.2719	−0.2756	−0.3345	−0.6127	−0.4526
	−0.449	−0.4513	−0.4561	−0.4717	−0.4392
ln Total assets	0.0286	0.0291	0.0264	0.0539	0.0623
	−0.0701	−0.0711	−0.0702	−0.0711	−0.0703
ROA	0.8211+	0.8233+	0.8204+	0.9395*	0.9352*
	−0.424	−0.4269	−0.4216	−0.4313	−0.4235
Sales growth	0.0001	0.0001	0.0001	0.0002+	0.0002
	−0.0001	−0.0001	−0.0001	−0.0001	−0.0001
Current ratio	−0.0216	−0.0216	−0.0205	−0.0174	−0.0165
	−0.0415	−0.0415	−0.0412	−0.0405	−0.0404
Debt-equity ratio	0.0005	0.0005	0.0005	0.0005	0.0005
	−0.0005	−0.0005	−0.0005	−0.0005	−0.0005
R&D intensity	−1.2016	−1.1978	−1.1913	−0.973	−0.9042
	−1.5929	−1.6009	−1.6316	−1.6412	−1.6232
Marketing intensity	1.3832+	1.3841+	1.3627+	1.3906+	1.3377+
	−0.717	−0.716	−0.7202	−0.7084	−0.7031
% Outsider directors	−0.2543	−0.2555	−0.2222	−0.2456	−0.2884
	−0.1915	−0.1931	−0.1871	−0.1847	−0.1923
% White-male directors	−0.0859	−0.0852	−0.1014	−0.0849	−0.0684
	−0.1036	−0.1043	−0.1029	−0.1023	−0.1033
Inverse mills ratio	−0.3068	−0.3076	−0.3075	−0.3537	−0.3623
	−0.2436	−0.2443	−0.245	−0.2442	−0.2412
Verbal masculinity		0.0314		1.6939*	
		−0.4744		−0.7454	
Verbal femininity			−1.1957*	−2.6967**	
			−0.5859	−0.916	
Masculinity focus					0.4018+
					−0.2245

(continued)

Table 2 (continued)

	Model 1	Model 2	Model 3	Model 4	Model 5
Femininity focus					−0.5821+
					−0.3274
Firm and year fixed effects	Yes	Yes	Yes	Yes	Yes
Adj. R ²	0.629	0.628	0.735	0.737	0.738
#Firms	117	117	117	117	117
#Observations	516	516	516	516	516

This table presents the fixed-effects estimation of CEO compensation with respect to verbal masculinity and verbal femininity. CEO compensation is measured as the total compensation the give CEO is given at time *t*. The total compensation includes salary, bonus, restricted stock, grants, LTIP payouts, option grants, and all other annual rewards. Since the distribution of CEO compensation in our sample is right-skewed, we take a logarithm to the variable. Verbal masculinity is captured through zero-shot classification by using the labels of “aggressive”, “assertive”, “strong”, “dominant”, and “decisive.” For verbal femininity, we employ the same LLM model with the labels of “collaborative”, “supportive”, “understanding”, “empathetic”, and “soft.” Robust standard errors in parentheses. + $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

from the board. Verbal femininity from male CEOs reveals that their behavior deviates from traditional gender norms associated with leadership roles. In many cultural and organizational contexts, leadership is stereotypically aligned with masculine traits [66]. Given this, when male CEOs display feminine characteristics—such as empathy, collaboration, or softness—it may create a perception of gender role incongruity, where their behavior is seen as inconsistent with the expectations of their role as male leaders [67].

4.2 Additional Analyses

To gain deeper insights into the relationship between verbal femininity in male CEOs and their compensation, we examined various contingencies related to social perceptions of gender roles. One key factor we considered was growth. Specifically, we hypothesized that in growing firms, the board would place less emphasis on the CEO’s communication style, as the focus would likely be on performance metrics. In contrast, in stagnating firms, the board might pay greater attention to gendered communication traits, potentially amplifying the influence of gender roles on evaluations. To test this hypothesis, we divided the sample into two groups based on the median value of sales growth: a low-growth group and a high-growth group. We then conducted separate regression analyses for each subsample. The results indicate that a negative relationship between verbal femininity and CEO compensation exists only in stagnating firms, suggesting that the board’s sensitivity to gendered communication traits is heightened in less dynamic performance contexts.

Second, we explored firm performance as a potential contingency in the relationship between verbal femininity in male CEOs and their compensation. We conjectured that in underperforming firms, boards are likely to focus on strategies for improving performance and may pay less attention to the CEO's communication style, including verbal femininity. Meanwhile, in high-performing firms, where immediate concerns about operational stability are reduced, boards may have more capacity to evaluate and be influenced by gendered communication traits, potentially leading to greater bias. To test this, we split the sample based on the median value of ROA, dividing firms into lower-ROA and higher-ROA groups. We then conducted separate regression analyses for each group. The results indicated no significant relationship between verbal femininity and CEO compensation in the lower-ROA group, suggesting that performance concerns overshadow biases related to communication style in underperforming firms. However, in the higher-ROA group, verbal femininity was found to have a significant negative effect on CEO compensation. This finding implies that in firms with strong financial performance, boards may be more susceptible to gendered expectations and biases when evaluating a CEO's communication style, emphasizing the situational nature of gender role incongruity in leadership assessments.

Lastly, we examined the demographic composition of the board as a potential factor influencing the relationship between verbal femininity and CEO compensation. Given that gender role congruity theory is rooted in gendered social structures, cultural assumptions about leadership and communication are likely to shape how boards perceive and evaluate verbal femininity in male CEOs. To explore this, we divided the sample into two groups: firms with white-male-dominated boards and those with more diverse boards. Separate analyses were conducted for each subsample to identify how verbal femininity and masculinity affected CEO compensation. The results revealed that white-male-dominated boards were more likely to factor verbal femininity into their decisions, negatively associating it with CEO compensation. In contrast, in boards that were not dominated by white males, verbal masculinity played a more significant role in influencing CEO compensation. These findings highlight how board demographics mediate the impact of gendered communication traits in leadership evaluations.

5 Discussion and Conclusion

In this study, we explore how cognitive biases in the boardroom, especially rooted in gender roles, influence the evaluation of CEO contributions. Specifically, this study examines the femininity-derived communication style of CEOs, i.e., verbal femininity, to determine whether gender biases affect boardroom evaluations of CEOs. To operationalize verbal femininity, we utilized a Large Language Model (LLM) to identify femininity-based cues in the communication of male CEOs. In the U.S. context, our findings reveal that verbal femininity has a negative impact on male CEOs' compensation, while verbal masculinity is positively associated with higher

compensation. These contrasting results align with the principles of gender role incongruity theory, highlighting the influence of gendered expectations on leadership evaluations. This incongruity of gender roles (i.e., femininity-derived communication styles of male CEOs), can lead to biased assessments by board members, who may interpret such communication styles as a lack of leadership strength or authority, thereby undervaluing the CEO's contributions. As such, the board's evaluation of the CEO, which often influences decisions about compensation and career progression, may be negatively affected. This suggests that gendered expectations in leadership play a significant role in shaping board perceptions and evaluations.

Our findings offer theoretical implications and contributions to the literature. First, this study deepens the understanding of gender role theory by illustrating how gendered expectations shape leadership evaluations, even when the leader's biological gender conforms to traditional leadership norms. Specifically, it demonstrates that male CEOs who display verbal femininity may face penalties in compensation because their behavior conflicts with stereotypical notions of masculine leadership. This finding underscores the persistence of gender role incongruity in leadership assessments, offering new insights into how implicit biases operate in evaluating executive performance.

Second, this study contributes to the leadership communication literature by demonstrating the nuanced role of communication style in shaping perceptions of leadership effectiveness. While prior studies have emphasized the importance of assertive and dominant communication in leadership evaluations, our findings reveal a double standard: traits culturally associated with femininity, though beneficial in fostering collaboration and relational dynamics, may still be undervalued or penalized in high-stakes settings such as board evaluations. This highlights the need to reevaluate the traditional emphasis on masculine communication styles in leadership frameworks and expand them to accommodate diverse and inclusive definitions of effective leadership.

Third, this study extends compensation and governance research by linking communication style to CEO compensation decisions. While compensation is traditionally viewed as a function of firm performance and governance mechanisms, our findings suggest that subjective factors, such as perceptions of a CEO's alignment with gendered expectations, also play a critical role. This contribution underscores the importance of addressing biases in board decision-making processes to ensure that leadership evaluations and compensation outcomes are based on objective criteria rather than cultural stereotypes.

Last, this study highlights the broader need to challenge traditional gender norms within the leadership domain. As organizations increasingly prioritize diversity, equity, and inclusion, it is imperative to recognize and mitigate the biases that undervalue leadership styles associated with femininity. By integrating these findings into the leadership literature, we encourage future research to explore strategies for reducing the influence of gender role incongruity in leadership evaluations and fostering greater acceptance of diverse leadership styles.

While this study presents significant contribution to the literature, it has more opportunities to further deepen our knowledge on verbal femininity and CEO

compensation. Future research could examine how the negative impact of verbal femininity on CEO compensation varies across different organizational and cultural contexts. For example, do industries traditionally associated with collaboration and relational leadership, such as healthcare or education, exhibit similar biases against verbal femininity in male leaders? Conversely, are these biases more pronounced in highly masculine, performance-driven sectors such as finance or technology? Comparative studies across industries and cultural settings could provide deeper insights into the situational factors that amplify or mitigate these biases.

Second, this study reveals implicit biases in board evaluations that penalize verbal femininity. Future research could delve into the decision-making processes within boards to uncover the mechanisms driving these biases. Questions to explore include: How do board member demographics, such as gender, age, or professional background, influence perceptions of gendered communication styles? Do boards with greater diversity or training in implicit bias exhibit less penalization of verbal femininity? Such studies could inform governance practices aimed at reducing bias in leadership evaluations.

Third, while this study focuses on verbal femininity, future research could investigate how verbal traits interact with non-verbal behaviors or leadership actions. For example, does the combination of verbal femininity with traditionally masculine behaviors, such as decisive decision-making, neutralize the negative evaluation? Exploring the interplay between communication styles and behavioral traits could offer a more holistic view of how gendered expectations shape leadership outcomes.

Last, a longitudinal approach could explore how the use of verbal femininity or masculinity evolves throughout a CEO's tenure and how these shifts affect perceptions of leadership effectiveness and compensation over time. Are leaders who consistently exhibit verbal femininity eventually recognized for its relational benefits, or do these traits continue to be undervalued? This type of research could provide insights into the temporal dynamics of gendered communication styles in leadership.

By pursuing these research directions, scholars can advance a more nuanced understanding of the interplay between gendered communication styles and leadership evaluations while also contributing to more equitable and inclusive leadership practices. Moreover, further exploration into verbal femininity, verbal masculinity, or verbal androgyny is needed to clarify the gender biases in the boardroom.

Acknowledgements Jeemin Jo provided excellent research assistance.

References

1. Graffin SD, Boivie S, Carpenter MA (2013) Examining CEO succession and the role of heuristics in early-stage CEO evaluation. *Strateg Manag J* 34:383–403
2. Nair K, Haque W, Sauerwald S (2022) It's not what you say, but how you sound: CEO vocal masculinity and the board's early-stage CEO compensation decisions. *J Manage Stud* 59(5):1227–1252

3. Westphal JD, Zajac EJ (1995) Who shall govern? CEO/board power, demographic similarity, and new director selection. *Adm Sci Q* 40:60–83
4. Zhu DH, Westphal JD (2014) How directors' prior experience with other demographically similar CEOs affects their appointments onto corporate boards and the consequences for CEO compensation. *Acad Manag J* 57(3)
5. Cannon B, Lynch J, Shams A (2022) Perception matters: how executive vocal masculinity influences CEO selection and compensation. Available at SSRN 4314748
6. Eagly AH, Carli LL (2003) The female leadership advantage. *Leadersh Q* 14(6):807–834
7. Eagly AH, Johnson BT (1990) Gender and leadership style. *Psychol Bull* 108(2):233–256
8. Eagly AH, Karau SJ (2002) Role congruity theory of prejudice toward female leaders. *Psychol Rev* 109(3):573
9. Eagly AH, Karau SJ, Makhijani MG (1995) Gender and the effectiveness of leaders. *Psychol Bull* 117(1):125–145
10. Adams RB, Funk P (2012) Beyond the glass ceiling: does gender matter? *Manage Sci* 58(2):219–235
11. Anglin AH, Kincaid PA, Short JC, Allen DG (2022) Role theory perspectives: past, present, and future applications of role theories in management research. *J Manag* 48(6):1469–1502
12. Finkelstein S, Hambrick DC, Cannella AA (2009) *Strategic Leadership: Theory and Research on Executives, Top Management Teams, and Boards*. Oxford University Press, New York
13. Abraham M (2020) Gender-role incongruity and audience-based gender bias: an examination of networking among entrepreneurs. *Adm Sci Q* 65(1):151–180
14. Ahn J, Kim J, Sung Y (2022) The effect of gender stereotypes on artificial intelligence recommendations. *J Bus Res* 141:50–59
15. Byrne J, Radu-Lefebvre M, Fattoum S, Balachandra L (2021) Gender gymnastics in CEO succession: masculinities, femininities and legitimacy. *Organ Stud* 42(1):129–159
16. Kamiya S, Kim YH, Park S (2019) The face of risk: CEO facial masculinity and firm risk. *Eur Financ Manag* 25(2):239–270
17. Mount MP, Sharpe WH, Lai KM, Gul FA (2024) Are boards sensitive to CEO masculinity? The effect of CEO facial and vocal masculinity on CEO dismissal. *J Manag Stud*
18. Paustian-Underdahl SC, Walker LS, Woehr DJ (2014) Gender and perceptions of leadership effectiveness: a meta-analysis of contextual moderators. *J Appl Psychol* 99(6):1129–1145
19. Kaneko M, Bollegala D, Okazaki N, Baldwin T (2024) Evaluating gender bias in large language models via chain-of-thought prompting. arXiv preprint [arXiv:2401.15585](https://arxiv.org/abs/2401.15585).
20. Shin T, You J (2017) Pay for talk: how the use of shareholder-value language affects CEO compensation. *J Manage Stud* 54(1):88–117
21. Boeker W (1992) Power and managerial dismissal: scapegoating at the top. *Adm Sci Q* 400–421
22. Fitza MA (2017) How much do CEOs really matter? Reaffirming that the CEO effect is mostly due to chance. *Strateg Manag J* 38(3):802–811
23. Busenitz LW, Barney JB (1997) Differences between entrepreneurs and managers in large organizations: biases and heuristics in strategic decision-making. *J Bus Ventur* 12(1):9–30
24. Liu J, Tsang EW, Shi W (2023) The superstitious heuristic in strategic decision making. *J Manag* 01492063231198191
25. Newell A, Simon HA (1972) *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall
26. Petersen MB (2015) Evolutionary political psychology: on the origin and structure of heuristics and biases in politics. *Polit Psychol* 36:45–78
27. Vugt MV, Ronay R (2014) The evolutionary psychology of leadership: theory, review, and roadmap. *Organ Psychol Rev* 4(1):74–95
28. Gupta A, Wowak AJ (2017) The elephant (or donkey) in the boardroom: how board political ideology affects CEO pay. *Adm Sci Q* 62(1):1–30
29. Boivie S, Bednar MK, Aguilera RV, Andrus JL (2016) Are boards designed to fail? The implausibility of effective board monitoring. *Acad Manag Ann* 10(1):319–407
30. Holmes J (2007) Social constructionism, postmodernism and feminist sociolinguistics. *Gend Lang* 1(1)

31. Nyame J, Tomekyin C (2018) Social construction of masculinity and femininity as portrayed in Nzema proverbs. *Int J Innov Res Adv Stud* 5(7):227–234
32. Bushman RM, Indjekian RJ, Smith A (1996) CEO compensation: the role of individual performance evaluation. *J Account Econ* 21(2):161–193
33. Lewellyn KB, Muller-Kahle MI (2022) A configurational exploration of how female and male CEOs influence their compensation. *J Manag* 48(7):2031–2074
34. Zheng Y (2010) The effect of CEO tenure on CEO compensation: evidence from inside CEOs versus outside CEOs. *Manag Financ* 36(10):832–859
35. Darouichi A, Kunisch S, Menz M, Cannella AA Jr (2021) CEO tenure: an integrative review and pathways for future research. *Corp GovAnce: Int Rev* 29(6):661–683
36. De Angelis D, Grinstein Y (2020) Relative performance evaluation in CEO compensation: a talent-retention explanation. *J Financ Quant Anal* 55(7):2099–2123
37. Deng X, Gao H (2013) Nonmonetary benefits, quality of life, and executive compensation. *J Financ Quant Anal* 48(1):197–218
38. Li ZF (2014) Mutual monitoring and corporate governance. *J Bank Financ* 45:255–269
39. Park G, Yaden DB, Schwartz HA, Kern ML, Eichstaedt JC, Kosinski M, Seligman ME (2016) Women are warmer but no less assertive than men: gender and language on Facebook. *PLoS ONE* 11(5):e0155885
40. Leaper C, Ayres MM (2007) A meta-analytic review of gender variations in adults' language use: talkativeness, affiliative speech, and assertive speech. *Pers Soc Psychol Rev* 11(4):328–363
41. Newman ML, Groom CJ, Handelman LD, Pennebaker JW (2008) Gender differences in language use: an analysis of 14,000 text samples. *Discourse Process* 45(3):211–236
42. Bem SL (1974) The measurement of psychological androgyny. *J Consult Clin Psychol* 42(2):155
43. McCreary DR (1994) The male role and avoiding femininity. *Sex Roles* 31:517–531
44. Auster CJ, Ohm SC (2000) Masculinity and femininity in contemporary American society: a reevaluation using the Bem Sex-role inventory. *Sex Roles* 43:499–528
45. Spence JT, Helmreich RL (1978) *Masculinity and femininity: Their psychological dimensions, correlates, and antecedents*. University of Texas Press
46. Kumar P (2024) Large language models (LLMs): survey, technical frameworks, and future challenges. *Artif Intell Rev* 57(10):260
47. Sufi F (2024) Generative pre-trained transformer (GPT) in research: a systematic re view on data augmentation. *Information* 15(2):99
48. Wang J, Huang JX, Tu X, Wang J, Huang AJ, Laskar MTR, Bhuiyan A (2024) Utilizing BERT for information retrieval: survey, applications, resources, and challenges. *ACM Comput Surv* 56(7):1–33
49. Devlin J (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)*
50. Sindhu B, Prathamesh RP, Sameera MB, Kumara Swamy S (2024) The evolution of large language model: models, applications and challenges. In: 2024 international conference on current trends in advanced computing (ICCTAC). IEEE, pp 1–8
51. Yin W, Hay J, Roth D (2019) Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint [arXiv:1909.00161](https://arxiv.org/abs/1909.00161)*
52. Castrogiovanni GJ (1991) Environmental munificence: a theoretical assessment. *Acad Manag Rev* 16:542–565
53. Stinchcombe A (1965) Social structure and organizations. In: March J (eds) *Handbook of Organization*, Rand McNally: Chicago, IL, pp 142–193
54. Batjargal B, Hitt MA, Tsui AS, Arregle JL, Webb JW, Miller TL (2013) Institutional polycentrism, entrepreneurs' social networks, and new venture growth. *Acad Manag J* 56(4):1024–1049
55. Keats BW, Hitt MA (1988) A causal model of linkages among environmental dimensions macro organizational characteristics and performance. *Acad Manag J* 31:57–98
56. Karim S, Carroll TN, Long CP (2016) Delaying change: examining how industry and managerial turbulence impact structural realignment. *Acad Manag J* 59(3):791–817

57. Schilke O (2014) On the contingent value of dynamic capabilities for competitive advantage: the nonlinear moderating effect of environmental dynamism. *Strateg Manag J* 35:179–203
58. George G (2005) Slack resources and the performance of privately held firms. *Acad Manag J* 48(4):661–676
59. Bromiley P (1991) Testing a causal model of corporate risk taking and performance. *Acad Manag J* 34(1):37–59
60. Hausman JA (1978) Specification tests in econometrics. *Econometrica* 46(6):1251–1271
61. Rojahn K, Willemsen TM (1994) The evaluation of effectiveness and likability of gender-role congruent and gender-role incongruent leaders. *Sex Roles* 30:109–119
62. Barberá E (2003) Gender schemas: configuration and activation processes. *Can J Behav Sci* 35(3):176
63. McCann H (2022) Is there anything “toxic” about femininity? The rigid femininities that keep us locked in. In *Critical Femininities*. Routledge, pp 9–22
64. Lombard EJ, Azpeitia J, Cheryan S (2021) Built on uneven ground: how masculine defaults disadvantage women in political leadership. *Psychol Inq* 32(2):107–116
65. Paechter C (2006) Masculine femininities/feminine masculinities: power, identities and gender. *Gend Educ* 18(3):253–263
66. Koenig AM, Eagly AH, Mitchell AA, Ristikari T (2011) Are leader stereo types masculine? A meta-analysis of three research paradigms. *Psychol Bull* 137(4):616
67. Garcia-Retamero R, López-Zafra E (2006) Prejudice against women in male-congenial environments: perceptions of gender role congruity in leadership. *Sex Roles* 55:51–61

Integrating LLM-Based Time Series and Regime Detection with RAG for Adaptive Trading Strategies and Portfolio Management



Chenkai Li, Chi Ho Roger Chan, Seth H. Huang, and Paul Moon Sub Choi

Abstract This paper explores the latest methodologies for fine-tuning open-source Large Language Models (LLMs) to enhance quantitative trading strategies by integrating numerical data (e.g., historical prices, technical indicators) with textual data (e.g., news, earnings reports, social media sentiment). We employ Retrieval-Augmented Generation (RAG) with a vector database to efficiently handle and contextualize textual data, enabling LLMs to derive actionable insights from both structured and unstructured data. The proposed approach focuses on fully fine-tuning smaller models, such as GPT-4o Mini, for cost-effective and scalable applications in finance. The study aims to create a hybrid trading model that combines the predictive power of LLMs with traditional quantitative methods, improving accuracy and adaptability in financial markets. Key innovations include the integration of real-time data pipelines and adaptive model tuning. Experimental results demonstrate significant improvements in predictive accuracy and risk-adjusted returns, showcasing the practical value of these advanced fine-tuning methodologies in finance.

Keywords Machine learning · Large language model · Time series analysis · Textual analysis

C. Li

Industrial Engineering and Operations Research, Columbia University, New York, NY, USA
e-mail: cl4271@columbia.edu

C. H. R. Chan · S. H. Huang (✉)

Business School, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
e-mail: sethhuang@ust.hk

C. H. R. Chan

e-mail: chrchanab@connect.ust.hk

P. M. S. Choi

Cornell SC Johnson College of Business, Cornell University, Ithaca, NY, USA
e-mail: mc369@cornell.edu

1 Introduction

The proliferation of diverse data modalities [10] in financial markets, including numerical time series, textual news, and macroeconomic indicators, presents both opportunities and challenges for quantitative trading and portfolio management [11]. Traditional approaches often struggle to integrate such heterogeneous data sources effectively, limiting their ability to capture complex market dynamics. Recent advancements in LLMs [1, 3, 17] offer a promising solution, as these models are capable of processing and contextualizing multimodal data to generate actionable insights.

LLMs, pretrained [2] on vast corpora of text and fine-tuned for specific applications, have demonstrated remarkable capabilities in understanding and generating human-like language. Beyond text, their adaptability allows for the incorporation of structured numerical data, enabling a unified framework for analyzing multimodal inputs. By leveraging RAG [6], LLMs can efficiently retrieve relevant information from large-scale databases and contextualize it with other data sources, enhancing their applicability in financial markets.

Despite their potential, the application of LLMs in quantitative finance remains underexplored. Most existing studies focus on single-modality data, such as numerical time series [13] or isolated textual sentiment analysis [9]. Integrating both modalities in a cohesive framework can provide a more comprehensive understanding of market behavior, improving predictive accuracy and decision-making.

This study investigates the integration of fine-tuned LLMs [4] with RAG to enhance quantitative trading strategies. The proposed approach combines numerical financial data with textual insights from news and macroeconomic reports, aiming to create adaptive trading models capable of responding to evolving market conditions. Experimental results demonstrate the feasibility of this approach, showing improvements in predictive performance and risk-adjusted returns.

2 Data

2.1 Time Series

The time series data utilized in this study encompasses a wide range of financial metrics across multiple asset classes, including rates, credit, foreign exchange (FX), commodities, and equities. Each metric serves as a proxy for market behavior and provides insights into different facets of financial markets. Below, we detail each metric and its significance:

- **Rates:**

- **Bloomberg US Treasury Index:** Represents the total return of U.S. Treasury bonds, capturing overall performance in the government bond market.

- **U.S. Treasury 2-year/10-year yield spread:** Tracks the yield spread between 2-year and 10-year U.S. Treasury bonds, a widely used indicator of economic sentiment and potential recessionary trends.
- **Merrill Lynch Option Volatility Estimate (MOVE) Index:** Measures implied volatility in the U.S. Treasury market, often referred to as the “VIX of bonds,” indicating uncertainty in interest rate expectations.
- **Credit:**
 - **Bloomberg U.S. Investment Grade Corporate Bond Index:** Reflects the performance of investment-grade corporate bonds in the U.S. market.
 - **Bloomberg U.S. High Yield Corporate Bond Index:** Tracks the high-yield (junk) bond market, providing insights into riskier credit instruments.
- **Foreign Exchange (FX):**
 - **U.S. Dollar Index:** The U.S. Dollar Index, measuring the dollar’s value relative to a basket of major currencies.
 - **EURUSD Implied Volatility:** Represents the 3-month implied volatility of the EUR/USD currency pair, an indicator of expected future fluctuations.
 - **USDJPY, EURUSD, GBPUSD:** Exchange rates for major G10 currency pairs, providing insights into international trade and capital flows.
- **Commodities:**
 - **Brent Crude Oil Futures:** Tracks the price of crude oil, a key indicator of global economic activity and energy markets.
 - **XAUUSD:** Reflects the price of gold in U.S. dollars, often seen as a safe-haven asset during market turmoil.
- **Equities:**
 - **VIX Index:** Measures implied volatility in the SP500, often referred to as the “fear gauge.”
 - **SPX Index:** Tracks the performance of the SP500, a benchmark for U.S. equity markets.
 - **MSCI World Index:** Represents global equity market performance across developed markets.
 - **MSCI Emerging Markets Index:** Tracks equity performance in emerging markets.
 - **NVDA, AAPL, AMZN, GOOG, META, MSFT, TSLA:** Stock prices of leading technology and innovation-driven companies, representing a significant portion of market capitalization and economic influence.
- **Macro Indicators:**
 - **U.S. Consumer Price Index:** Measures year-over-year changes in the Consumer Price Index, a key inflation gauge.

- **U.S. Unemployment Rate Index:** Represents total unemployment in the U.S., a critical labor market indicator.
 - **U.S. Manufacturing Purchasing Managers' Index:** Tracks the Purchasing Managers' Index (PMI) for manufacturing, indicating economic health in the industrial sector.
 - **Consumer Confidence Index in the Eurozone:** Measures consumer sentiment in the Eurozone, reflecting economic confidence.
- **Currency Pairs:**
 - **USDJPY, EURUSD, GBPUSD:** Exchange rates for major G10 currency pairs, providing insights into international trade and capital flows.

Each metric is preprocessed to derive meaningful insights. For most metrics, we calculate percentage changes to capture relative movements over time. In some cases, such as bond yields or inflation rates, absolute changes reflect the scale of movement. Additionally, rolling 125-day standard deviations are computed to assess volatility trends. Metrics with standard deviations exceeding a threshold (e.g., 3) are flagged as potential indicators of regime shifts within their respective asset classes. These regime shifts are then aggregated into a counter, ranging from 0 to 5, representing the number of asset classes experiencing elevated volatility simultaneously.

2.2 News

The news dataset used in this study comprises macroeconomic news articles sourced from Nasdaq.com, covering global financial developments from January 2023 to October 2024. These articles include announcements related to monetary policy, economic indicators, geopolitical events, and corporate earnings reports. By integrating this textual data, the study aims to capture qualitative insights that complement quantitative time series data.

The news dataset is filtered for macroeconomic relevance by scanning articles for keywords such as “inflation,” “interest rates,” “GDP,” “unemployment,” and “central bank policy.” This ensures that only articles with potential implications for broad market dynamics are included, excluding company-specific updates that are less relevant to the study’s objectives.

To link news events with financial metrics, a timestamp alignment is performed, matching the publication date of each article with the corresponding time series data. This enables the model to contextualize numerical trends within the broader macroeconomic landscape. For example, if a spike in bond market volatility coincides with a Federal Reserve announcement, the timestamp alignment allows for this relationship to be captured in the analysis. This structured approach ensures that the textual data remains directly tied to macroeconomic events, providing actionable insights without introducing additional computational complexity.

2.3 Regime Shift

Regime shifts in financial markets refer to transitions between distinct phases of market behavior, often characterized by changes in volatility, correlations, or trends across asset classes. Detecting these shifts is critical for adaptive trading strategies and risk management, as they can signal changes in market conditions that require adjustments in portfolio allocations. For this study, we classify regime shifts based on the standard deviation of returns, identifying extreme events when the standard deviation exceeds a threshold of 3 over a 125-day rolling window.

The rolling standard deviation ($\sigma_{t:t+124}$) for a given window starting at time t and ending at $t + 124$ is calculated as

$$\sigma_{t:t+124} = \sqrt{\frac{1}{125} \sum_{i=t}^{t+124} (r_i - \bar{r}_{t:t+124})^2}, \quad (1)$$

where:

- r_i represents the i -th return within the window,
- $\bar{r}_{t:t+124}$ is the mean return over the 125-day window:

$$\bar{r}_{t:t+124} = \frac{1}{125} \sum_{i=t}^{t+124} r_i. \quad (2)$$

A regime shift is identified for a specific asset class k on day $t + 124$ if:

$$\sigma_{t:t+124,k} > 3, \quad (3)$$

indicating an extreme deviation from typical market behavior within the 125-day rolling window.

For each day $t + 124$, the aggregate regime shift score S_t across all M asset classes is calculated as

$$S_t = \sum_{k=1}^M I(\sigma_{t:t+124,k} > 3), \quad (4)$$

where:

- M is the total number of asset classes,
- $I(\cdot)$ is an indicator function that equals 1 if $\sigma_{t:t+124,k} > 3$, and 0 otherwise,
- $\sigma_{t:t+124,k}$ is the rolling standard deviation for asset class k over the window t to $t + 124$.

The daily score S_t ranges from 0 (no regime shifts across any asset classes) to M (all asset classes experience regime shifts). This provides a day-specific measure of extreme market activity across asset classes.

Interpretation: The score S_t reflects the intensity of regime shifts across asset classes on a given day. A heatmap of S_t over time can be used to visualize periods of heightened market instability. Additionally, time-series plots of $\sigma_{t:t+124,k}$ for individual asset classes can help identify specific contributors to regime activity.

The regime shift counter aggregates the number of asset classes experiencing these elevated volatility periods at any given time. This counter ranges from 0 (indicating no regime shifts) to 5 (indicating simultaneous shifts across all asset classes). This aggregated measure provides a holistic view of market turbulence and aids in identifying systemic risk events.

The identified regime shifts are further analyzed in conjunction with the news dataset. By correlating regime shift events with significant news articles, the study seeks to uncover causal relationships or patterns. For instance, a regime shift in the FX market may coincide with major central bank announcements, while shifts in the equities market could align with unexpected corporate earnings reports or geopolitical crises. This integrated approach enhances the model's ability to interpret market dynamics and adapt trading strategies accordingly.

3 Methodology

3.1 Time Series Preprocessing

Time series data in financial markets are often high-dimensional, making them computationally expensive to process and challenging to integrate into models designed primarily for text processing [8], such as LLMs. To address these challenges, this study employs Symbolic Aggregate approXimation (SAX) [5, 7], a method that transforms numerical time series into symbolic representations, facilitating dimensionality reduction and making the data more LLM-compatible.

The SAX method involves three key steps: normalization, Piecewise Aggregate Approximation (PAA), and symbolization. First, a time series $\mathbf{T} = [t_1, t_2, \dots, t_n]$ is normalized to have zero mean and unit variance:

$$\mathbf{T}' = \frac{\mathbf{T} - \mu}{\sigma},$$

where μ is the mean of \mathbf{T} , and σ is its standard deviation. Normalization ensures that the method is robust to differences in scale among time series.

Next, the normalized series \mathbf{T}' is divided into w equal-sized segments, and the mean of each segment is calculated. This process, known as Piecewise Aggregate Approximation (PAA), reduces the dimensionality of the series:

$$\mathbf{T}_{\text{PAA}} = [\bar{t}_1, \bar{t}_2, \dots, \bar{t}_w],$$

where $\bar{t}_i = \frac{1}{\Delta} \sum_{j=1}^{\Delta} t_j$, and $\Delta = \frac{n}{w}$ is the segment length.

Finally, the PAA-transformed series \mathbf{T}_{PAA} is mapped to a discrete alphabet \mathcal{A} of size a based on breakpoints derived from a Gaussian distribution. Each segment mean \bar{t}_i is assigned a symbol $s_i \in \mathcal{A}$ according to:

$$s_i = f(\bar{t}_i),$$

where $f(\cdot)$ maps \bar{t}_i to a symbol based on its range relative to the breakpoints. The result is a symbolic representation of the time series:

$$\mathbf{T}_{\text{SAX}} = [s_1, s_2, \dots, s_w].$$

The SAX representation offers several advantages. By transforming time series into symbolic sequences, it reduces computational costs and storage requirements, making it particularly suitable for scenarios involving large-scale data. Furthermore, the symbolic format aligns well with the input structure of LLMs, which are optimized for processing textual data. This enables the integration of time series data into LLM-based frameworks without significant modifications to the model architecture.

In this study, SAX is applied to financial time series such as stock prices. The symbolic representations are concatenated with textual data, allowing the model to analyze numerical and textual inputs in a unified manner. This preprocessing step not only reduces computational complexity but also enhances the interpretability of the time series data within the context of LLM-driven analysis.

3.2 *Regime Question Bank*

The Regime Question Bank is a critical component designed to enhance the model's understanding of the macroeconomic environment, particularly during regime shifts. By retrieving and analyzing relevant news and contextual data, the question bank ensures that the model remains aware of prevailing market conditions and their potential implications [14]. This integration supports a more informed and adaptive trading strategy.

Market-Wide Questions: The market-wide questions aim to capture broad trends and macroeconomic dynamics across global markets. These questions focus on understanding equity performance, interest rate trends, macroeconomic indicators, and geopolitical events. Examples include:

- What is the current state of the global equity market?
- How has the S&P 500 performed over the past quarter?
- What are the key drivers of recent market volatility?
- Are there any geopolitical events impacting global markets today?
- What are the trends in interest rates, and how are they affecting bond markets?
- What sectors are currently outperforming or underperforming in the market?
- How is the Federal Reserve's policy impacting investor sentiment?

- Are there any major earnings reports or economic data releases today?
- What is the market sentiment based on recent news and social media analysis?
- How have macroeconomic factors like inflation or GDP growth influenced the markets this year?

The embeddings for these questions are computed using a pretrained language model to encode their semantic meaning into a high-dimensional vector space. For a given question embedding \mathbf{q}_i , the top- k relevant news articles are retrieved by ranking their similarity to \mathbf{q}_i using cosine similarity. This process ensures that the retrieved information directly corresponds to the key macroeconomic drivers.

Asset-Specific Questions: To provide a detailed perspective on individual assets, asset-specific questions are dynamically generated for selected assets. These questions target asset-level metrics such as volatility, valuation, and recent news, helping the model contextualize asset performance within the broader macroeconomic regime. Examples of asset-specific questions include

- What is the historical volatility of $[asset]$?
- How does $[asset]$'s price correlate with its sector or benchmark index?
- Are there any recent news or earnings reports related to $[asset]$?
- What is the current valuation of $[asset]$ compared to its historical average?
- How have dividends or other distributions affected $[asset]$'s returns?
- What is the liquidity level of $[asset]$ in the current market?
- Are there any insider trading activities or large institutional movements related to $[asset]$?
- What is the short interest ratio or sentiment surrounding $[asset]$?
- How has $[asset]$ performed in similar market conditions in the past?
- What are the key risks and opportunities associated with $[asset]$?

For each asset \mathbf{a}_j , a set of embeddings $\{\mathbf{q}_j^{(1)}, \mathbf{q}_j^{(2)}, \dots, \mathbf{q}_j^{(m)}\}$ is computed. These embeddings are matched with news articles to retrieve the top- k items most relevant to the asset. This ensures that the model is equipped with timely and asset-specific contextual information.

Advantages of the Question Bank: The primary advantage of the Regime Question Bank is its ability to help the model understand the macroeconomic environment by maintaining awareness of regime shifts. By structuring queries to address both broad market conditions and asset-level specifics, the question bank ensures that the model integrates relevant information into its decision-making process. This approach enhances the model's interpretability and adaptability, particularly during periods of heightened market volatility or systemic changes.

Furthermore, the symbolic structure of the queries aligns well with the capabilities of LLMs, enabling efficient integration of textual and numerical data. This design allows the model to contextualize regime shifts in real-time and adapt trading strategies accordingly, improving performance and risk management during turbulent market conditions.

3.3 Retrieval Augmented Generation

RAG is a critical component of this study, enabling the integration of external knowledge into the model's decision-making process. By retrieving relevant information from a vector database and combining it with the model's inherent capabilities, RAG enhances the interpretability and effectiveness of the trading strategies.

Embedding News Articles: To facilitate retrieval, each news article is converted into a high-dimensional embedding vector using a pretrained embeddings model. This embedding captures the semantic meaning of the text, making it suitable for similarity-based retrieval. For longer news articles, the study employs an LLM to condense the content into a more concise representation before applying the embedding process. This ensures that the embeddings remain focused on the most critical information, improving retrieval efficiency and relevance.

The embeddings are stored in a vector database, such as Pinecone [12], which is used in this project. Pinecone allows for scalable and efficient storage and querying of high-dimensional vectors, enabling the retrieval of relevant news articles in real-time.

Embedding Questions: In addition to embedding news articles, questions from the Regime Question Bank are similarly converted into embedding vectors. This ensures consistency in the representation of both the queries and the news articles, facilitating accurate similarity computations.

Retrieval Process: During the retrieval phase, cosine similarity is used to measure the distance between the question embeddings and the news embeddings. For a given question embedding \mathbf{q} and a news embedding \mathbf{n} , the cosine similarity is defined as

$$\text{Similarity}(\mathbf{q}, \mathbf{n}) = \frac{\mathbf{q} \cdot \mathbf{n}}{\|\mathbf{q}\| \|\mathbf{n}\|}.$$

This similarity score quantifies the relevance of a news article to the query. The top- k news articles with the highest similarity scores are retrieved from the vector database and provided as input to the model for further processing.

Advantages of RAG: The use of RAG provides several benefits in the context of this study. By dynamically retrieving relevant information from external sources, the model is able to incorporate up-to-date and contextually relevant insights into its analysis. The embedding and retrieval process ensures that both questions and news articles are represented in a format compatible with the model, enabling seamless integration. Additionally, the ability to condense long news articles reduces noise and computational overhead, focusing the analysis on the most pertinent details.

By incorporating RAG into the framework, this study enhances the model's ability to understand complex macroeconomic environments and adapt trading strategies in response to evolving market conditions.

3.4 Chain of Thoughts

The CoT framework employed in this study is a structured, step-by-step reasoning process designed to ensure transparency and robustness in decision-making. This process integrates time series predictions, regime detection, news retrieval, and trading decisions into a cohesive analytical workflow (Fig. 1). The CoT consists of the following four steps:

Step 1: Time Series Prediction

The process begins with the analysis of the input time series $\mathbf{T} = [t_1, t_2, \dots, t_n]$, which is preprocessed using the SAX algorithm to reduce dimensionality and represent the time series in symbolic form $\mathbf{T}_{\text{SAX}} = [s_1, s_2, \dots, s_w]$, where $s_i \in \{a, b, \dots, i\}$. Using this SAX-transformed time series, the model predicts the future return time series $\hat{\mathbf{T}} = [\hat{s}_1, \hat{s}_2, \dots, \hat{s}_m]$ for the next m time steps:

$$\hat{\mathbf{T}} = f_{\text{predict}}(\mathbf{T}_{\text{SAX}}),$$

where f_{predict} represents the prediction function based on the input time series. The predicted sequence $\hat{\mathbf{T}}$ serves as the initial output for subsequent analysis.

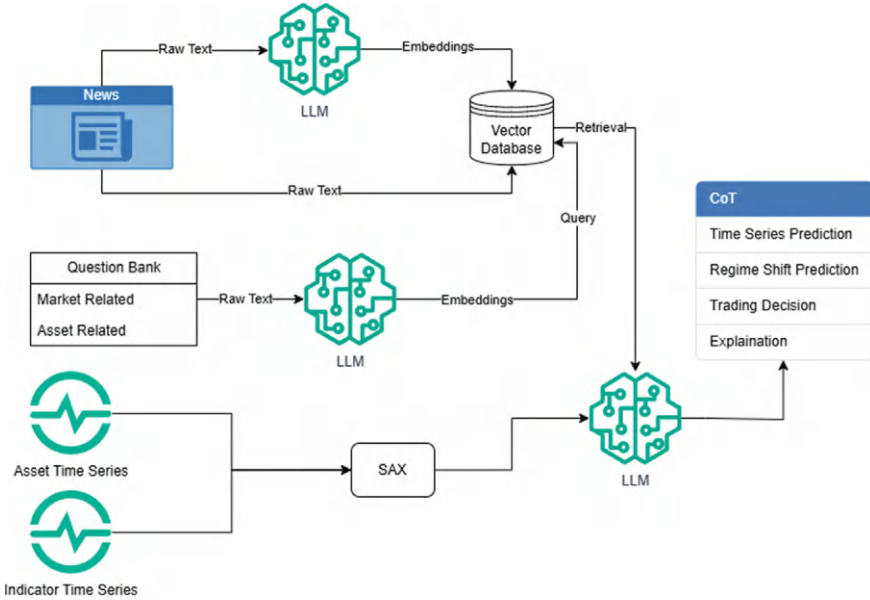


Fig. 1 Workflow of the LLM pipeline integrating SAX, RAG, and the Chain of Thoughts (CoT) [16] process. The pipeline combines asset and indicator time series data with news and question bank inputs to generate predictions, detect regime shifts, and provide trading decisions and explanations

Step 2: Regime Shift Detection and Adjustment

Following the time series prediction, the model evaluates the likelihood of a regime shift within the prediction horizon. This is based on a set of market indicators $\mathbf{I} = [i_1, i_2, \dots, i_k]$ and retrieved news embeddings $\mathbf{N} = [\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_p]$. A regime shift is detected if:

$$R = f_{\text{regime}}(\mathbf{I}, \mathbf{N}) = \begin{cases} \text{Yes,} & \text{if } P(R = \text{Yes}) > \tau, \\ \text{No,} & \text{otherwise,} \end{cases}$$

where f_{regime} represents the regime detection function, $P(R = \text{Yes})$ is the probability of a regime shift, and τ is a predefined threshold.

If $R = \text{Yes}$, the initial time series prediction $\hat{\mathbf{T}}$ is adjusted to reflect the impact of the regime shift. The adjusted prediction $\hat{\mathbf{T}}_{\text{adjusted}}$ is calculated as

$$\hat{\mathbf{T}}_{\text{adjusted}} = f_{\text{adjust}}(\hat{\mathbf{T}}, R),$$

where f_{adjust} modifies the predicted returns based on the detected market regime (Yes or No for the regime shift).

Step 3: Trading Decision

Using the adjusted time series prediction $\hat{\mathbf{T}}_{\text{adjusted}}$, the model generates a binary trading decision $D \in \{\text{Long}, \text{Short}\}$ for the next m days. The decision is based on an aggregation function that evaluates the directional trends in $\hat{\mathbf{T}}_{\text{adjusted}}$:

$$D = f_{\text{decision}}(\hat{\mathbf{T}}_{\text{adjusted}}),$$

where f_{decision} outputs `Long` if the predicted trend is upward and `Short` if the predicted trend is downward.

Step 4: Explanation

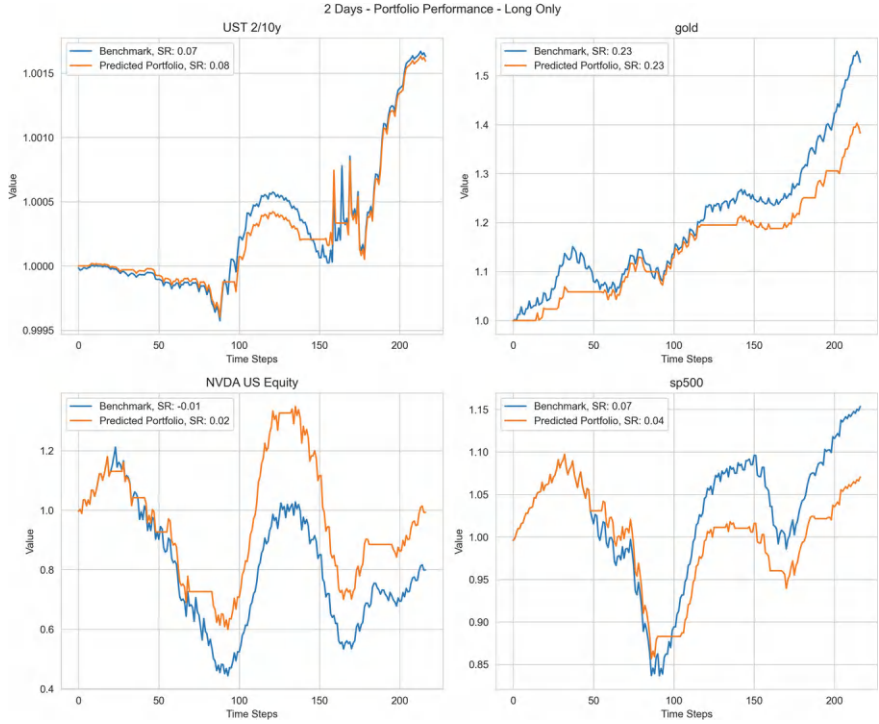
The final step involves generating a natural language explanation, providing transparency and interpretability for the model's predictions, regime detection, and trading decisions. Let \mathbf{E} represent the explanation, which is defined as

$$\mathbf{E} = f_{\text{explain}}(\mathbf{N}, R, \hat{\mathbf{T}}_{\text{adjusted}}, D),$$

where f_{explain} is a function that integrates the retrieved news embeddings \mathbf{N} , the regime shift detection result R , the adjusted time series prediction $\hat{\mathbf{T}}_{\text{adjusted}}$, and the trading decision D .

Advantages of the Chain of Thoughts Framework

The CoT framework enhances the model's ability to synthesize complex, multimodal inputs into coherent outputs. By structuring the workflow into discrete steps, the approach mitigates errors, improves interpretability, and ensures that predictions and decisions are grounded in a comprehensive understanding of market conditions [15]. This methodology is particularly effective for scenarios requiring the integration of



(a) 2 Days Long, Temp 0.7

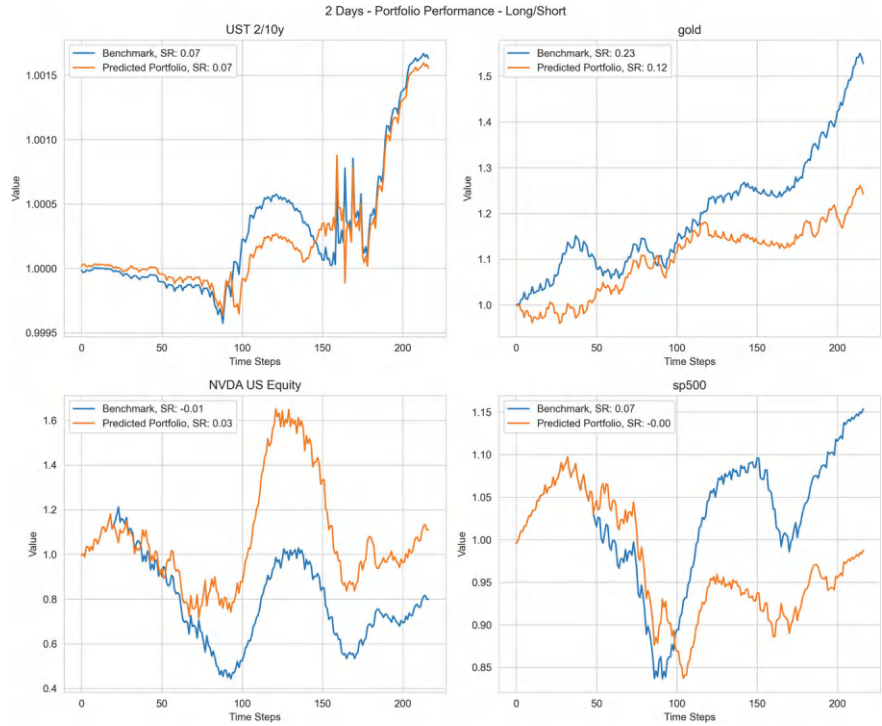
Fig. 2 Portfolio performance across different rolling windows and trading strategies on the test set

time series data, macroeconomic indicators, and real-time news retrieval, as it ensures the model remains adaptable to dynamic financial environments.

4 Experiments

The experiments evaluate the integration of fine-tuned LLMs with RAG in adaptive trading strategies. The dataset spans from January 1, 2023, to September 30, 2024, divided into training (70%), validation (15%), and test (15%) sets.

Each input sequence includes 128 days of historical time series and retrieved textual data, with a prediction horizon of 7 days. Fine-tuning is applied exclusively to the GPT-4o Mini [1] model due to computational constraints, while GPT-4o is used without fine-tuning to benchmark performance. Key hyperparameters include model temperature, which controls prediction randomness, and the rolling window, which determines portfolio adjustment frequency.



(b) 2 Days Long/Short, Temp 0.7

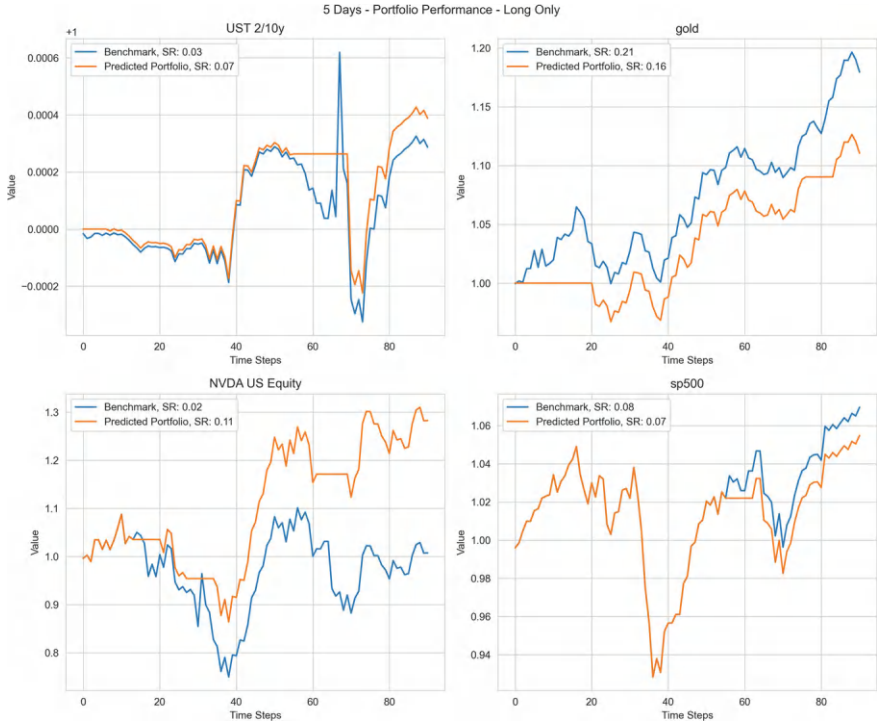
Fig. 2 (continued)

Figure 2 illustrates the portfolio performance for different rolling windows (2-day, 5-day, and 7-day) and trading strategies (long and long/short) at a model temperature of 0.7. The results are evaluated on the test set, demonstrating how fine-tuned GPT-4o Mini adapts to varying horizons and trading strategies.

Table 1 presents results for regime shift detection and trading decision accuracy, along with their respective F1 scores.

Fine-tuning the GPT-4o Mini model improves regime shift detection and trading decision accuracy, demonstrating its utility in this framework. For instance, at a temperature of 0.7, fine-tuned GPT-4o Mini achieves 53.7% regime shift accuracy and 50.8% trading accuracy, outperforming the non-fine-tuned GPT-4o. This highlights the effectiveness of fine-tuning in task-specific adaptation. At a temperature of 0, fine-tuned models exhibit deterministic behavior, often outputting “No” for regime shifts and “Long” for trading decisions. While this ensures stable predictions, it limits adaptability to dynamic market conditions.

Non-fine-tuned GPT-4o performs consistently but is less effective than fine-tuned GPT-4o Mini, particularly in capturing regime shifts. This underscores the value of



(c) 5 Days Long, Temp 0.7

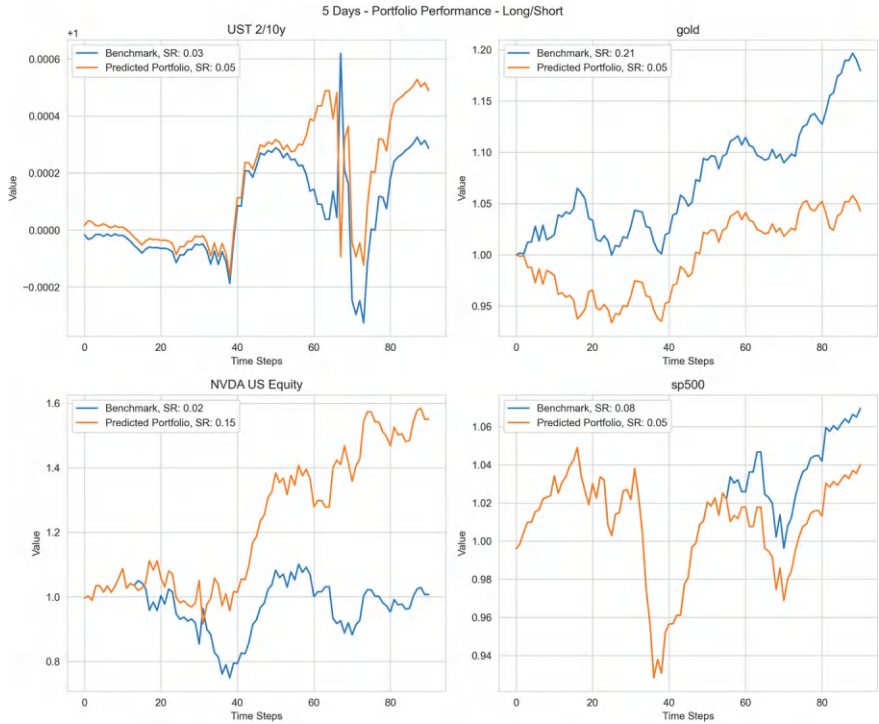
Fig. 2 (continued)

fine-tuning, which tailors smaller models to task-specific requirements while maintaining computational efficiency.

The results validate the framework’s effectiveness, with visualized portfolio performance further supporting the adaptability of fine-tuned models across varying horizons and trading strategies.

5 Conclusion

This study demonstrates the integration of fine-tuned LLMs with RAG for adaptive trading strategies and portfolio management. By combining numerical time series data and textual insights from news and macroeconomic indicators, the proposed framework addresses the challenges of multimodal data integration in financial markets.



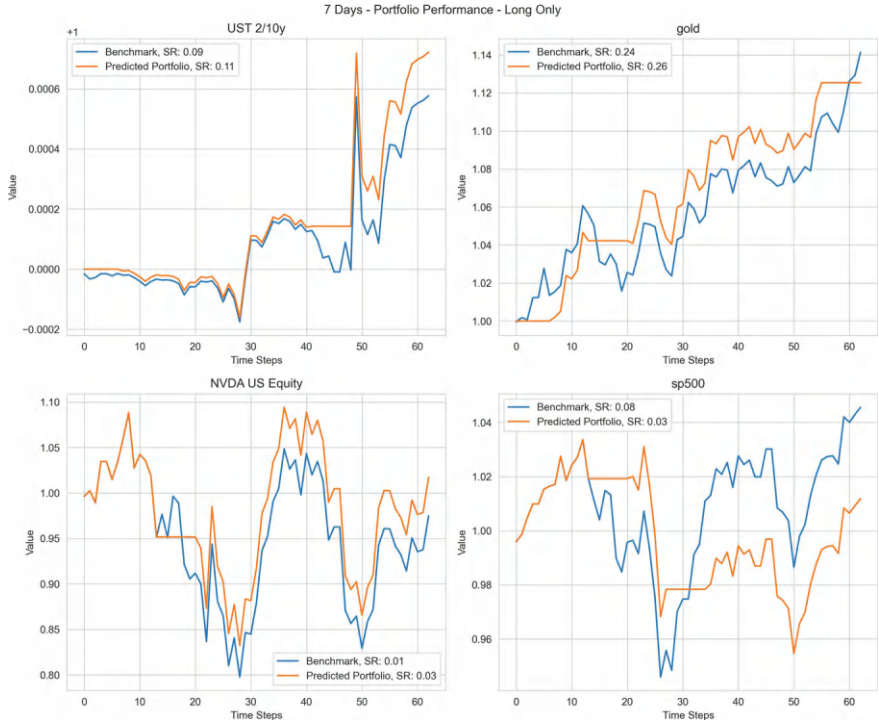
(d) 5 Days Long/Short, Temp 0.7

Fig. 2 (continued)

The experimental results highlight the value of fine-tuning smaller LLMs, such as GPT-4o Mini, which improves regime shift detection and trading decision accuracy while maintaining computational efficiency. The application of SAX enhances the compatibility of time series data with LLMs, while the CoT framework ensures transparency and robustness in decision-making. This proof-of-concept establishes a solid foundation for integrating advanced LLMs in quantitative finance.

6 Future Work

This study serves as a proof-of-concept, demonstrating the integration of fine-tuned LLMs with RAG for adaptive trading strategies and portfolio management. While the results establish the feasibility of the proposed approach, further research is needed to expand its scope and enhance its robustness.

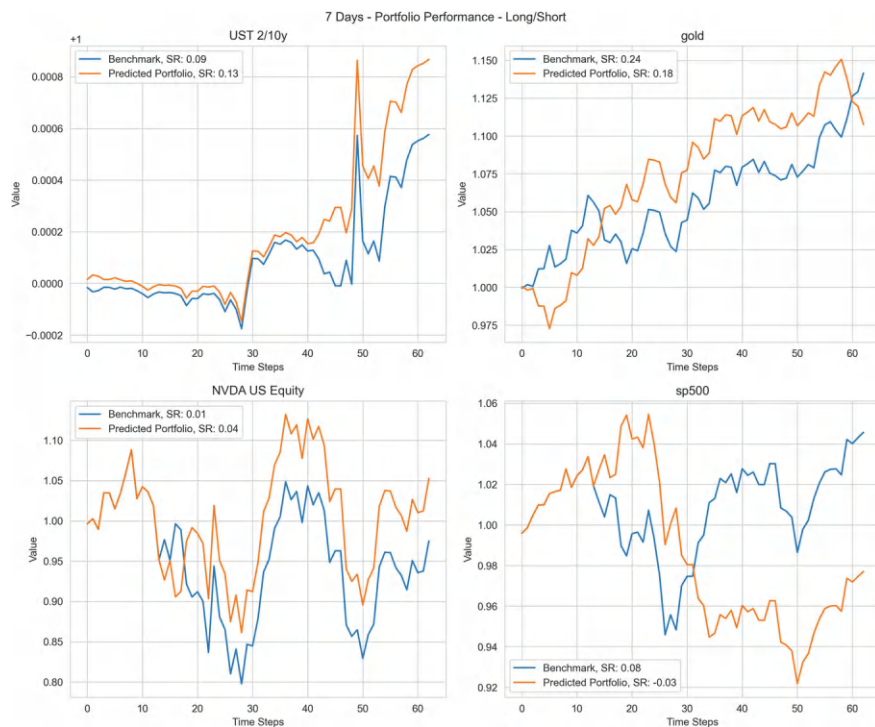


(e) 7 Days Long, Temp 0.7

Fig. 2 (continued)

Table 1 Experimental results for regime shift and trading decision

Model	Fine-tuning	Temperature	Regime/trading accuracy (%)	Regime/trading F1 score (%)
GPT-4o Mini	Yes	1.0	48.8/48.8	49.2/46.6
GPT-4o Mini	Yes	0.7	53.7/50.8	53.7/46.5
GPT-4o Mini	Yes	0.5	48.4/52.9	46.8/45.3
GPT-4o Mini	Yes	0.2	48.4/50.8	45.5/35.8
GPT-4o Mini	Yes	0.0	52.5v52.0	36.1/35.6
GPT-4o	No	1.0	48.8/48.8	44.4/47.9
GPT-4o	No	0.7	47.1/46.7	42.0/46.3
GPT-4o	No	0.5	47.5/45.5	39.9/44.6
GPT-4o	No	0.2	44.7/47.5	30.0/44.8
GPT-4o	No	0.0	45.9/33.1	31.0/43.1



(f) 7 Days Long/Short, Temp 0.7

Fig. 2 (continued)

First, extending the methodology to incorporate higher-frequency data, such as intraday or tick-level time series, could enable strategies that respond to shorter time horizons and real-time market dynamics. Second, instead of relying on simple long/short and long-only strategies, future work could focus on models that directly generate trading signals or portfolio weights. This would facilitate seamless integration with execution frameworks and support portfolio optimization processes.

Incorporating continual learning frameworks would allow the models to adapt to new data as it becomes available, ensuring their relevance and effectiveness in dynamic market environments. Additionally, exploring more complex trading strategies, such as pairs trading, options-based approaches, and multi-asset portfolio optimization, could further leverage insights generated by fine-tuned LLMs.

These future directions provide a path for advancing LLM-based trading systems, improving their performance, adaptability, and application in financial markets.

References

1. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S et al (2023) Gpt-4 technical report. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774)
2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. *Adv Neural Inf Process Syst* 33:1877–1901
3. Devlin J (2018) Bert: pre-training of deep bidirectional transformers for language understanding. [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
4. Dodge J, Ilharco G, Schwartz R, Farhadi A, Hajishirzi H, Smith N (2020) Fine-tuning pretrained language models: weight initializations, data orders, and early stopping. [arXiv:2002.06305](https://arxiv.org/abs/2002.06305)
5. Faouzi J, Janati H (2020) pyts: a python package for time series classification. *J Mach Learn Res* 21(46):1–6. <http://jmlr.org/papers/v21/19-763.html>
6. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T et al (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv Neural Inf Process Syst* 33:9459–9474
7. Lin J, Keogh E, Lonardi S, Chiu B (2003) A symbolic representation of time series, with implications for streaming algorithms. In: *Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, pp 2–11
8. Liu C, Xu Q, Miao H, Yang S, Zhang L, Long C, Li Z, Zhao R (2024) Timecma: towards llm-empowered time series forecasting via cross-modality alignment. [arXiv:2406.01638](https://arxiv.org/abs/2406.01638)
9. Mishev K, Gjorgjevikj A, Vodenska I, Chitkushev LT, Trajanov D (2020) Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE Access* 8:131662–131682
10. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp 689–696
11. Philippon T (2016) The fintech opportunity. Tech. rep, National Bureau of Economic Research
12. Pinecone Systems, Inc (2023) Pinecone: vector database for machine learning. <https://www.pinecone.io/>. Accessed 16 Dec 2024
13. Sezer OB, Gudelek MU, Ozbayoglu AM (2020) Financial time series forecasting with deep learning: a systematic literature review: 2005–2019. *Appl Soft Comput* 90:106181
14. Smales LA (2017) The importance of fear: investor sentiment and stock market returns. *Appl Econ* 49(34):3395–3421
15. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D et al (2022) Emergent abilities of large language models. [arXiv:2206.07682](https://arxiv.org/abs/2206.07682)
16. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D et al (2022) Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst* 35:24824–24837
17. Yang H, Liu XY, Wang CD (2023) Fingpt: open-source financial large language models. [arXiv:2306.06031](https://arxiv.org/abs/2306.06031)

Empirical Factor Identification for Artificial Intelligence in Finance: Indian Evidence



Rohit Kaushik 

Abstract In the fast-changing landscape of financial markets, artificial intelligence is making inroads into it. Artificial intelligence is a concept or technological advancement which aims to mirror intelligence exhibited by the human beings and executes all the jobs to the perfection. Financial or investment domain is not an exception to this. Newer technologies are making their way to the horizon like Large Language Models (LLMs) as they present themselves into innovative and interactive methods of extracting information for making informed investment decisions. Before making any investment decision one has to look at various aspects and pay attention to various reports and constantly monitor the changes in stock prices, which is a very difficult job every person can't claim to master this art, in this artificial intelligence comes to the rescue of such persons who can't perform such tasks carefully. The present study is an attempt in this regard and it consists of respondents who are living in Delhi national capital of India. Data was collected with the help of questionnaire and analyzed by employing multiple regression method to ascertain the effect of artificial intelligence on investment decisions in addition to overconfidence bias which was another factor studied with artificial intelligence.

Keywords Artificial intelligence · Robo-Advisory · LLMs · Investment decisions · Overconfidence bias

1 Introduction

Artificial intelligence has become a new development that is pervasive in all aspects. Most of the tasks in daily routine nowadays are performed to a great extent by the artificial intelligence as Alexa is doing all the household activities and managing all the appliances in the home of any individual. Financial sector is not unaffected by the advent of the artificial intelligence. This convergence of minds and machines has effected a momentous change in all the spheres of life. Human beings are prone

R. Kaushik (✉)

Jagannath International Management School, Vasant Kunj, New Delhi, Delhi, India
e-mail: rkaushik46@gmail.com

to biases and ultimately they make inferior decisions and suffer losses. Artificial intelligence on the other hand is a software which does not rely on instincts and biases and very practical and takes very tough decisions without getting emotional.

Artificial intelligence analyzes the market condition comprehensively and suggest course of action or takes action pragmatically. In stock markets, artificial intelligence is playing its role in the form of algorithm trading where with the help of intelligence algorithms investors take right decisions and execute many orders at a given point in time which gives them additional edge over other investors in the market. In many global institutions, these changes have taken place and Bank of Tokyo and Bank of America are examples in this regard Marinova et al. [1], Rosman et al. [2].

In the age of novel technology, people are always found doing their important work on smartphones and spending considerable time and fintech companies sensed the absence of such things or facilities in the banking or investment arena. Robo-advisors are a result of such concerted efforts of fintech companies to develop digital facilities for customers with smart technology and interactive too. A robo-advisor gathers information from the clients on regular basis pertaining to their financial condition and their financial goal and advise them accordingly. Evolution of robo-advisory is crucial to maintaining the transparency for ethical and round-the-clock functioning of the fintech companies and banking sector in general but its revolutionary nature has not resulted in full adoption by the masses in today's scenario robo-advisors have totally changed the way investment services are offered and in a way disrupted the market Lukkanen et al. [3].

Large language models on the other hand is a new development in the domain of the technological advancements which help in the revolutionizing the financial behavior of investors. Earlier financial analysis was used to be done by the experts or analysts but with the advent of Large Language Models, a systemic change has occurred in which Large Language Models are performing financial analysis on the companies resulting in less errors. Expertise of Large Language Models lies in its ability to process and extract meaningful information from huge volumes of unstructured data. But application of Large Language Models in the financial sector is not without challenges as it requires integration with data sets, data training and concerns in relation to data privacy and security.

The research intends to study the understand the preferences of consumers in relation to robo-advisory with respect to investment decisions of retail investors in Indian financial environment and it also provides valuable insights into extent of adoption of robo-advisory in Indian financial landscape, identify those factors which are going to result in faster adoption of robo-advisory by keeping in center the factors like perceived utility, external influence and internal influence.

2 The Literature

The advent of robo-advisory and artificial intelligence will transform the entire gamut of activities in financial sector and will have huge transformational value for all sectors cutting across all the dimensions of business sector Acemoglu et al. [4]. On the other hand Huang et al. [5] emphasized upon the eventuality which is awaiting the mankind in case the artificial intelligence is properly implemented in the system and human being are no longer needed to work, they also anticipated that artificial intelligence will develop all abilities ranging from mechanical to intuitive intelligence.

Simultaneously researches were also focusing upon the challenges in the way of introduction of automation in the service sector, special emphasis was placed upon sectors where direct communication with consumers was needed. These researches noted that consumer interaction is going to be shaped profoundly with the introduction of artificial intelligence and will also change the consumer behavior in the market, but the level of adoption of technology in service sector will depend on the presence of humans and relative capacity of artificial intelligence to socialize with consumers on daily basis Singh et al. [6], Han et al. [7], Van Doorn et al. [8], Grewal et al. [9], Ji [10].

In totality, there is a huge realization among companies to introduce artificial intelligence in their operation as part of increasing efficiency and range of products. Management strategies adopted by their companies in relation to their operations also get a boost and passively increase competitive advantage of all the companies employing artificial intelligence.

Robo-advisors have been depicted as interactive and intelligent systems that are built on the foundations of information technology and helped consumers in taking crucial investment decisions Jung et al. [11]. Initially these systems first gather information from consumers through questionnaire, on the basis of information provided by the consumers in the questionnaires, these systems start making recommendation to the consumers.

Robo-advisors present a range of benefits to the entire community of consumers, unlike human advisors they are always on service for the human beings, resulting in cost reduction on the part of companies and also helping consumer in seeking professional advice regarding diverse investment options in an efficient manner and eliminating the scope of fraud which is always looming large in case of dealing with human advisors Faubion [12], Park et al. [13]. During the period 2019–22 assets managed by robo-advisors crossed \$880,000 million.

Apart from Robo-advisors one new technological development has taken place which is known as Large Language Models (LLMs), introduction of these sophisticated models have resulted in the way of interaction of companies with their clients and also augment efforts of companies in providing 24/7 services to the consumers and giving expert investment advice. Besides this LLMs are also extremely useful in conducting sentiment analysis by processing large datasets available on the internet and providing valuable information to the company about the prevailing trends in

the market and facilitate better preparation on the part of company professionals de Kok [14], Paul et al. [15].

LLMs are also exhibiting their utility in the area of fraud detection as the traditional methods of stopping frauds are proving to be insufficient and LLMs are becoming a potent weapon to understand the modus operandi of the fraudsters and minimizing the damage to the unsuspecting customers Ali et al. [16].

Even though this concept is evoking huge interest, but some studies have pointed out the obstacles in the introduction of robo-advisors and found that corporate world is more interested in exploring the capabilities of artificial intelligence and its application in diverse areas whereas consumers are not ready to trust applications based on artificial intelligence due to security reasons. Customization is the unique advantage of robo-advisors identified by the studies in relation to investment management Faloon et al. [17], studies have also found a correlation between increased usage of artificial intelligence and informed decision-making Heinrich et al. [18].

Considering freshness associated with the concept of robo-advisory in the process of investment management, there is lack of awareness about the key factors which determine the prospects for adoption of robo-advisory by the consumers, with help of robo-advisory the financial sector is making advisory services available to all the consumers for better decision-making Sironi [19].

The consumers' thoughts and perceptions are conditioned by the thought process of people who are close to the consumer. The societal behavior acts as a pressure for the consumer to adopt a technology or disapprove of it Taylor et al. [20], Fishbein et al. [21].

People with limited information and also facing the situation or specter of a disruptive technology, as they lack relevant experience in the concerned field tend to behave like their relatives, peers, friends, basing their decisions on the experiences of others in order to enhance their assimilation in the society and appear modern in the eyes of others Belanche [22].

In the area of finance, there is considerable role played by the close social circle of an individual and the other non-human source of information like media and by this one can easily understand that consumers tend to gain some confirmation from others with respect to their decision Bhattacharjee et al. [23], Ryu [24].

Familiarity with any concept or new technology results in the adoption of that technology as that is going to enhance or decrease the willingness on the part of consumers to actually engage with the new technology in question. Same is the case with robo-advisors, in the 1980s and 90s all the banking jobs and investment management services were provided by the human resources of the concerned organization but now robo-advisors are playing that role with increased efficiency and professionalism but all the consumers are not welcoming of the new change and a substantial part of the population is not exhibiting faith in adopting robo-advisors for investment management and with different level of knowledge and familiarity one can sense the difference in the extent of adoption Mäenpää et al. [25], consumers with limited information about new technology will be more reliant on the experiences of others Venkatesh et al. [26], Young et al. [27].

3 Empirics

3.1 Hypotheses

By examining the available research and literature in the public domain, questionnaire was modeled to address the basic requirements of the present research. First part of the questionnaire was related to the socio-economic profile of the respondents and the other two parts were dedicated to the core areas of the research like perceived trust, internal and external influence and intention to use in relation to robo-advisory. The questions were designed with the help of Likert Scale.

Sample size of the present study was of 100 people and all of them were residing in Delhi as Delhi is the political and financial capital of India. People from diverse background participate in the financial activities taking place in the city resulting in diverse opinions making it a near-perfect fit for carrying out research.

Data collected in the study was analyzed with the help of SPSS software using Multiple Regression Methods to ascertain the influence of perceived trust, internal and external influence on intention to use robo-advisory. To streamline the process of research following hypothesis were formed (Table 1):

Table 1 Demographic profile

Variables	Category	Frequency	%
Gender	Male	54	54
	Female	46	46
Age	18–30	59	59
	31–45	29	29
	46–60	12	12
Marital status	Unmarried	46	46
	Married	54	54
Education	School education	11	11
	Graduation	44	44
	Post-Graduation	36	36
	Others	9	9
Occupation	Private employees	43	43
	Public employees	10	10
	Business	13	13
	Others	34	34
Income	< 5 lakhs	68	68
	5–10 lakhs	25	25
	> 10 lakhs	7	7

- *H1: Perceived trust does not share a significant relationship with an intention to use robo-advisory services.*
- *H2: Internal influence does not share a significant relationship with an intention to use robo-advisory services.*
- *H3: External influence does not share a significant relationship with an intention to use robo-advisory services.*

3.2 Data Analysis

For the validity of a research model, all the variables in the research have to fulfill some norms to avoid lopsidedness and biasedness. The data was collected from 100 respondents. Acceptable value of a Cronbach Alpha test is 0.70 and a variable or construct must measure more than 0.70 Hair et al. (2006). Cronbach Alpha value more than 0.80 points toward the soundness of the constructs in terms of validity. The values attained by the constructs in the present study are in the range of 0.75–0.82 (Tables 2, 3, 4 and 5).

As can be observed in the results of the data analysis that independent variables have significantly affected the dependent variable $(3, 96) = 75.221, p < 0.01$ which is reflected by the effectiveness of the model. The model in the present study is able

Table 2 Reliability analysis

Construct	No. of items	Alpha (α)
Internal influence	3	0.82
External influence	3	0.75
Perceived utility	3	0.82
Intention to use	3	0.78

Table 3 Model summary

R	R ²	Adjusted R ²	Standard error of the estimate
0.838	0.702	0.692	0.477

The predictors are the intercept term, EF₁, PU₁, and IF₁

Table 4 Anova

	Sum of squares	df	Mean square	F	Sig.
Regression	51.332	3	17.111	75.221	0
Residual	21.838	96	0.227		
Total	73.17	99			

The dependent variable is IU₁. The predictors are the intercept term, EF₁, PU₁, and IF₁

Table 5 Coefficients

Variable	Unstandardized coefficients		Standardized		
	Beta	Std. error	Beta	t-stat.	Sig.
Intercept	−0.184	0.295		−0.623	0.535
PU ₁	0.384	0.086	0.359	4.472	0.000
IF ₁	0.29	0.117	0.225	2.482	0.015
EF ₁	0.41	0.078	0.387	5.268	0.000

The dependent variable is IU₁

to reflect the variations in the dependent factor due to independent factors in the question, to the tune of 70 percent as reflected by an R^2 value of 0.70.

As we delve deeper into the individual role played by the independent factors in relation to the dependent factor we find that Hypotheses 1 examined the relation between perceived utility and intention to use robo-advisory and the results in the study throw up significant relation between the factors ($\beta = 0.384$, $t = 4.472$, $p = 0.00$) and clearly showcase that utility clearly shapes the usage behavior of the respondents in relation to the robo-advisory and Hypothesis 1 is rejected in view of the results. H2 evaluates the relationship of advises of close relatives and family members and the adoption of robo-advisory services and intention to use it ($\beta = 0.290$, $t = 2.48$, $p = 0.015$) and clearly points out that views of immediate family members affect the perception of any person to use the robo-advisory services in a significant services. H3 studied the impact of external influence on the intention to use robo-advisory services in the form of friends and peer circle and the results found a significant relation between the two ($\beta = 0.410$, $t = 5.268$, $p = 0.00$) and in a much profound manner than internal influence and shaped immensely the intention of the individual to use the robo-advisory services and the hypothesis is rejected.

3.3 Discussion

In the ongoing emergence of FinTech companies, robo-advisory services caught the attention of investors and individuals in general as robo-advisers rely only on artificial intelligence and their advent and usage is resulting in fast replacement of humans from this area. From a customer's perspective, it is very essential to identify those factors which result in the adoption of robo-advisory services in the area of investment management and personal financial management. As this study noted that subjective norms and perceived utility are the factors which immensely influenced the decision to adopt or use robo-advisory services.

As the findings of the present study concluded that internal and external influence share a significant relation with the intention of usage of robo-advisory service and it is in conformity with the findings of other studies like [21] who have earlier concluded in their study that subjective norms like internal and external influence, found to be

impacting intention of consumers regarding adoption of robo-advisory or artificial intelligence in the domain of investment management.

On the other hand perceived utility was another factors which was studied in its impact on intention of consumer and it was also found to be sharing a significant relation with dependent variable in question, this relationship as thrown up by the findings is also consistent with previous studies' findings Hernandez et al., 2009 as it has also concluded that perceived utility as a factor not only contributes to the decision of individuals before the actual usage of the robo-advisory but even after using the services it positively conditions the mind of individual. As the millennium generation is actively involved in the financial decision-making this results also takes into account the advance exposure of the current generation to the artificial intelligence and their willingness to adopt robo-advisory services is reflected in the results or findings of the study.

As the major contribution of the research lies in divulging key aspects which can be focused upon by managers while introducing artificial intelligence in the form robo-advisory services, they will have to present robo-advisors in a way which is very user friendly so that consumer adoption process is simplified as due to simple interface there is higher possibility of consumers switching to robo-advisory services. Financial institutions need to organize conferences to interact with their customer base to spread awareness about the concept of robo-advisors so that subjective norms can also be managed which have found to be impacting ultimate decisions of consumers.

4 Concluding Remarks

In the present study, small sample size plays a inhibiting role in terms of generalization study for that similar studies with comparatively large sample size need to be carried out in adjoining big cities in other studies. People were not willing to share their personal details and were giving biased responses to the questions in the questionnaire. The present study is going to be useful for the financial institutions to focus on those features which consumers attach a lot of importance and hence will result in enhanced customer adoption of robo-advisors in the area of investment management.

References

1. Marinova D, de Ruyter K, Huang MH, Meuter ML, Challagalla G (2017) Getting smart: learning from technology-empowered frontline interactions. *J Serv Res* 20(1):29–42
2. Rosman C (2018) Mad about erica: why a million people use Bank of America's chatbot. *American Banker* 183:114
3. Laukkanen T, Pasanen M (2008) Mobile banking innovators and early adopters: how they differ from other online users? *J Financ Serv Mark* 13:86–94

4. Acemoglu D, Restrepo P (2020) Robots and jobs: evidence from US labor markets. *J Polit Econ* 128(6):2188–2244
5. Huang MH, Rust RT (2018) Artificial intelligence in service. *J Serv Res* 21(2):155–172
6. Singh J, Brady M, Arnold T, Brown T (2017) The emergent field of organizational frontlines. *J Serv Res* 20(1):3–11
7. Han S, Yang H (2018) Understanding adoption of intelligent personal assistants: a parasocial relationship perspective. *Ind Manag Data Syst* 118(3):618–636
8. Van Doorn J, Mende M, Noble SM, Hulland J, Ostrom AL, Grewal D, Petersen JA, (2017) Domo arigato Mr. Roboto: emergence of automated social presence in organizational frontlines and customers' service experiences. *J Serv Res* 20(1):43–58
9. Grewal D, Guha A, Saturnino CB, Schweiger EB (2021) Artificial intelligence: the light and the darkness. *J Bus Res* 136:229–236
10. Ji M (2017) Are robots good fiduciaries: regulating robo-advisors under the investment advisers act of 1940. *Colum L Rev* 117:1543
11. Jung I, Sun H, Kang J, Lee CH, Lee S (2018) Bigdata analysis model for MRO business using artificial intelligence system concept. *Int J Eng Technol* 7(3):134–138
12. Faubion B (2016) Effect of automated advising platforms on the financial advising market
13. Park JY, Ryu JP, Shin HJ (2016) Robo advisors for portfolio management. *Adv Sci Technol Lett* 141(1):104–108
14. de Kok T (2024) ChatGPT for textual analysis? How to use generative LLMs in accounting research. How to use Generative LLMs in Accounting Research
15. Paul J, Ueno A, Dennis C (2023) ChatGPT and consumers: benefits, pitfalls and future research agenda. *Int J Consum Stud* 47(4):1213–1225
16. Ali A, Abd Razak S, Othman SH, Eisa TAE, Al-Dhaqm A, Nasser M, Saif A (2022) Financial fraud detection based on machine learning: a systematic literature review. *Appl Sci* 12(19):9637
17. Faloon M, Scherer B (2017) Individualization of robo-advice. *J Wealth Manag* 20(1):30
18. Heinrich P, Schwabe G (2018) Facilitating informed decision-making in financial service encounters. *Bus Inf Syst Eng* 60:317–329
19. Sironi P (2016) *FinTech innovation: from robo-advisors to goal based investing and gamification*. John Wiley & Sons
20. Taylor S, Todd PA (1995) Understanding information technology usage: a test of competing models. *Inf Syst Res* 6(2):144–176
21. Fishbein M, Ajzen I (1977) *Belief, attitude, intention, and behavior: an introduction to theory and research*
22. Belanche D, Casaló LV, Flavián C (2019) Artificial intelligence in fintech: understanding robo-advisors adoption among customers. *Ind Manag Data Syst* 119(7):1411–1430
23. Bhattacharjee A (2000) Acceptance of e-commerce services: the case of electronic brokerages. *IEEE Trans Syst, Man, Cybern-Part A: Syst Hum* 30(4):411–420
24. Ryu HS (2018) What makes users willing or hesitant to use Fintech? The moderating effect of user type. *Ind Manag Data Syst* 118(3):541–569
25. Mäenpää K, Kale SH, Kuusela H, Mesiranta N (2008) Consumer perceptions of internet banking in Finland: the moderating role of familiarity. *J Retail Consum Serv* 15(4):266–276s
26. Venkatesh V, Davis FD (2000) A theoretical extension of the technology acceptance model: four longitudinal field studies. *Manage Sci* 46(2):186–204
27. Young JE, Hawkins R, Sharlin E, Igarashi T (2009) Toward acceptable domestic robots: applying insights from social psychology. *Int J Soc Robot* 1:95–108

Federated and Decentralized Finance: Decentralized Reward Mechanisms for Advanced AI Learning



Hyoseok Jang, Sangchul Lee, Haneol Cho, and Chansoo Kim

1 Introduction

Federated Learning (FL) is a distributed machine learning paradigm in which multiple clients collaboratively train a global model while keeping their private data local [2, 6, 8]. This setup significantly reduces the privacy risks associated with conventional centralized training, yet FL still faces important challenges. First, data across different clients often follow *non-Independent and Identically Distributed (non-IID)* distributions [11], which can slow convergence and introduce biases in the global model. Second, there is limited transparency or incentive for clients to contribute high-quality updates, creating vulnerabilities to malicious or low-quality participants [1]. Third, communication and aggregation overhead grows with the number of participating clients, complicating large-scale deployments.

Concurrently, blockchain technology has garnered attention for its decentralized consensus mechanisms and transparent record-keeping [4, 9, 10]. These properties align well with FL's core principle of distributed collaboration, suggesting a natural synergy. In particular, blockchain's reward systems—originally devised to incentivize honest participation in consensus—could help address FL's incentive gap. By assigning tokens or other rewards proportionate to a node's verified contribution, nodes in a federated network can be encouraged to provide accurate updates rather than harmful or trivial ones. Furthermore, a blockchain-based ledger can record all contributions in an auditable manner, aiding in the detection and penalization of malicious actors.

H. Jang · S. Lee · H. Cho · C. Kim (✉)

AI, Information and Reasoning Laboratory, Computational Science Centre, Korea Institute of Science and Technology, Seoul, Korea

e-mail: eau@ust.ac.kr

H. Jang · C. Kim

Department of AI-Robot, University of Science and Technology, Seoul, Korea

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

P. M. S. Choi and S. H. Huang (eds.), *Finance and Large Language Models*,

Blockchain Technologies, https://doi.org/10.1007/978-981-96-5833-6_9

157

This paper introduces a **Blockchain-Based Reward System** (BBRS) for Federated Learning, outlining how tokenomics, delayed reward schemes, and transparent logging can mitigate FL’s trust and incentive problems. Specifically,

- Incorporate *non-IID data considerations* into FL’s aggregation and reward design, ensuring clients are fairly compensated regardless of heterogeneous data distributions.
- Leverage blockchain *consensus algorithms* (e.g., Proof of Stake, Practical Byzantine Fault Tolerant) to securely track node updates with lower overhead than classic Proof of Work.
- Propose a *delayed payout mechanism* that allows additional validation (e.g., via cross-validation or anomaly detection) before finalizing rewards.
- Illustrate how real-world applications-ranging from *IoT networks* and *smart cities* to *healthcare* and *autonomous vehicles*-can benefit from this integrated system.

By fusing FL with blockchain rewards, we seek to create an *open, trustworthy, and autonomous* learning environment, where stakeholders have both a technical framework and a strong economic incentive to collaborate honestly. In the following sections, we present the theoretical underpinnings of FL, describe how blockchain can address FL’s vulnerabilities, and detail our proposed reward mechanism and its potential impact.

2 Theoretical Foundations of Federated Learning

Federated Learning (FL) distributes machine learning tasks across multiple clients (or nodes), each holding a local dataset. Rather than pooling all data on a central server, FL coordinates local model training and then aggregates the results into a global model. The key objectives are to (1) protect data privacy by keeping raw data on-device, and (2) leverage distributed computational resources efficiently.

2.1 Global Objective

Consider K clients, where client k holds dataset \mathcal{D}_k . Due to variations in data-collection environments, these local datasets often follow *non-Independent and Identically Distributed (non-IID)* distributions [11]. Formally, the global model parameters \mathbf{w} are obtained by minimizing a weighted sum of local loss functions:

$$\min_{\mathbf{w}} F(\mathbf{w}) = \sum_{k=1}^K p_k F_k(\mathbf{w}),$$

where

- \mathbf{w} represents the model parameters (e.g., weights of a neural network),
- $F_k(\mathbf{w})$ is the local loss on client k 's data,
- $p_k \geq 0$ are scaling factors, often set as $p_k = \frac{|\mathcal{D}_k|}{\sum_{j=1}^K |\mathcal{D}_j|}$ to reflect the dataset size.

In non-IID contexts, each $F_k(\mathbf{w})$ may differ substantially, posing additional convergence challenges (e.g., slower or biased training).

2.2 Federated Learning Process

Initialization A central coordinator (server) initializes a global model with parameters $\mathbf{w}^{(0)}$. The parameters are broadcast to all clients.

Local Update At each training round t , client k receives the latest global model $\mathbf{w}^{(t)}$. It then performs local optimization (often mini-batch gradient descent) on \mathcal{D}_k for one or more local epochs:

$$\mathbf{w}_k^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla F_k(\mathbf{w}^{(t)}),$$

where η is the local learning rate. With non-IID data, each client's gradient may vary significantly, affecting the global model's convergence properties.

Aggregation Upon completing local training, clients transmit their updated parameters (or gradients) back to the server. The server aggregates these updates:

$$\mathbf{w}^{(t+1)} = \sum_{k=1}^K p_k \mathbf{w}_k^{(t+1)}.$$

This step ensures the global model reflects contributions from all participating clients. However, *aggregator overhead* increases with the number of clients K . In large-scale scenarios (e.g., thousands or millions of devices), communication frequency and payload sizes can become a bottleneck. Hierarchical or asynchronous FL approaches have been proposed to mitigate this overhead.

Iteration and Convergence The process (local update \rightarrow aggregation) repeats until certain termination criteria, such as a global accuracy threshold or a maximum number of rounds, are met. In non-IID settings, additional rounds or adaptive methods (e.g., personalized FL) may be required to achieve satisfactory performance.

2.3 Advantages and Limitations

Advantages:

- *Data Privacy*: Since only model parameters or gradients are exchanged, sensitive data remains local.
- *Parallelism*: Clients train concurrently, leveraging distributed computational resources.
- *Scalability*: FL can, in principle, incorporate large numbers of heterogeneous clients without necessitating a single data warehouse.

Limitations:

- *Non-IID Distributions*: Heterogeneous local data can slow global model convergence or skew results [11].
- *Reliability and Security*: Malicious or compromised clients can inject harmful updates [1], demanding robust aggregation or anomaly detection.
- *Communication Overhead*: Frequent model updates become expensive when K is large, necessitating efficient aggregator strategies or reduced synchronization frequency.
- *Incentive Structure*: Clients currently have limited motivation to invest computational resources or share meaningful updates, as FL frameworks often lack explicit reward or compensation mechanisms.

In light of these challenges, research has explored improved aggregation rules (e.g., secure aggregation [2], Byzantine-resilient updates), personalization strategies for non-IID data, and incentive frameworks to motivate honest participation. Later sections of this paper detail how *blockchain-based rewards* can help tackle both the *trust gap* and the *economic incentive gap* in FL.

3 Using Blockchain to Address Federated Learning Challenges

Although federated learning (FL) offers data privacy and distributed computation, it faces persistent challenges related to trust, incentive structures, and resilient model aggregation. Blockchain technology can address many of these gaps through its decentralized consensus mechanisms, transparent record-keeping, and token-based reward frameworks [4, 9, 10]. This section discusses how FL and blockchain align conceptually, then explores the role of blockchain reward systems and consensus algorithms in fortifying FL's reliability and scalability.

3.1 Conceptual Parallels Between FL and Blockchain

Decentralization and Data Sovereignty Both FL and blockchain operate without a single controlling entity:

- *Federated Learning*: Multiple clients retain control over their data and conduct local model training.
- *Blockchain*: Each node in the network holds a replica of the ledger, enforcing a decentralized trust model.

This shared focus on decentralized decision-making forms a natural foundation for their integration.

Consensus and Global State Updates In FL, a “global model” emerges by aggregating locally computed updates. In blockchain networks, consensus protocols (e.g., Proof of Work [9], Proof of Stake [5], Practical Byzantine Fault Tolerant (PBFT) [3]) ensure agreement on the “global ledger state.” Analogous to FL’s model averaging, blockchain’s consensus reconciles possibly conflicting updates into a canonical version of the ledger.

Security and Auditability FL preserves data privacy by localizing training to each client’s device, but verifying the integrity of these updates can be challenging [7]. Blockchain, by contrast, provides a tamper-resistant log of transactions (or any recorded data), enabling reliable auditing of node contributions. Combining these strengths allows for a verifiable record of local model updates while keeping raw data private.

3.2 Blockchain Reward Systems and Consensus for Federated Learning

Why a Blockchain-Based Reward? A key limitation in FL is the lack of incentive for nodes to contribute meaningful or high-quality updates. Clients may also behave maliciously by injecting erroneous gradients [1]. Blockchain was originally designed to reward honest actors who maintain the network’s consensus. By integrating a *token reward mechanism* into FL, we can

- *Motivate Participation*: Nodes receive tokens (or other digital assets) based on the value (accuracy gain, resource usage) of their updates.
- *Encourage Fair Play*: Malicious or low-quality contributions can be identified and penalized, due to transparent on-chain records and delayed reward distribution.
- *Create Sustainable Economies*: A well-designed tokenomics model fosters a long-term cooperative environment, offsetting computation/communication costs.

Lower-Overhead Consensus: PoS and PBFT Classic Proof of Work (PoW) blockchains (e.g., Bitcoin [9]) ensure high security but incur heavy computational

costs. In large-scale FL, such overhead is often prohibitive. Alternative algorithms like *Proof of Stake (PoS)* [5] or *Practical Byzantine Fault Tolerant (PBFT)* [3] provide consensus with significantly lower energy consumption and higher throughput. Deployed in a federated learning context, these protocols can

- Reduce the resource burden required to validate updates,
- Minimize latency for on-chain confirmations,
- Achieve robust fault tolerance and mitigate Sybil attacks through staking or Byzantine-resilient replication.

On-Chain vs. Off-Chain Storage Recording every local model update fully on-chain may overwhelm blockchain storage. Instead, *hybrid* approaches can be considered:

- **On-Chain Metadata:** Store hashes or encrypted proofs of contributions, ensuring immutability and auditability.
- **Off-Chain Bulk Storage:** Large models remain off-chain, either on clients' devices or in secure distributed storage. Smart contracts verify integrity by comparing hashes or zero-knowledge proofs.

Such architectures balance *transparency* with *scalability*, preventing block size bloat while retaining a verifiable ledger of contributions.

Malicious Node Detection and Penalization In federated learning, malicious nodes can degrade the global model or embed backdoors [1]. Blockchain-based solutions can incorporate:

- *Delayed Rewards:* Final payouts for each update are withheld until additional validation (cross-validation, anomaly detection) confirms the update's legitimacy.
- *Reputation or Slashing Mechanisms:* Nodes found to produce harmful updates lose staked tokens or forfeit future rewards, mirroring slashing in PoS blockchains.

This adds a credible deterrent against adversarial or low-quality participation.

Extensions to Layer-2 Solutions Should the system require higher throughput, off-chain or "Layer-2" protocols (e.g., payment channels, sidechains) can further reduce on-chain transaction load. A *sidechain* could, for instance, handle micro-rewards in near real-time, periodically anchoring results to the main chain for security. Such *multi-layer* designs allow FL networks to scale without sacrificing the security or integrity benefits of a fully decentralized ledger.

Summary By choosing a suitable consensus mechanism and adopting on-chain/off-chain partitioning, blockchain networks can handle verifiable rewards distribution for large-scale FL. This synergy not only addresses the incentive gap but also introduces robust audit trails and fault tolerance, thereby reinforcing both the security and sustainability of federated learning (Fig. 1).

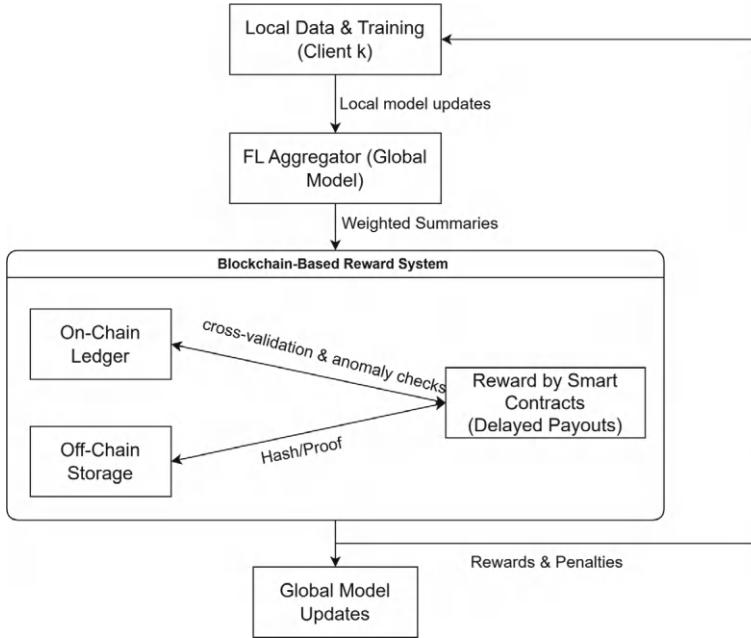


Fig. 1 Conceptual figure illustrating how Federated Learning (FL) synergizes with a Blockchain-Based Reward System (BBRS). Local data remains on each client, while weighted summaries are passed to the BBRS for verification and delayed payouts. On-chain/off-chain mechanisms, along with smart contracts, handle records, cross-validation, and anomaly checks. Malicious updates can be penalized, ensuring a transparent and trustworthy learning ecosystem

4 Mathematical Formulations for Blockchain + FL

This section presents key formulations for *contribution metrics*, *reward distribution*, and *synergy quantification* to illustrate how a Blockchain-Based Reward System (BBRS) can reinforce FL's reliability, fairness, and performance.

4.1 Defining Node Contribution

Let $\mathbf{w}^{(t)}$ denote the global model parameters at round t , and $\mathbf{w}_k^{(t+1)}$ be the locally updated parameters from client k . In a non-IID setting, each client's contribution can differ substantially. We define a *contribution quality* metric $Q_k^{(t+1)}$ that captures both *model improvement* and *resource usage*:

$$Q_k^{(t+1)} = \lambda_1 \cdot \underbrace{\frac{\Delta \text{Acc}_k^{(t+1)}}{\sum_{j=1}^K \Delta \text{Acc}_j^{(t+1)}}}_{\text{relative accuracy gain}} + \lambda_2 \cdot \underbrace{\frac{R_k^{\text{usage}}(t+1)}{\sum_{j=1}^K R_j^{\text{usage}}(t+1)}}_{\text{relative resource usage}},$$

where

- $\Delta \text{Acc}_k^{(t+1)} = \text{Acc}(\mathbf{w}_{\text{with } k}^{(t+1)}) - \text{Acc}(\mathbf{w}^{(t)})$ measures how much client k 's update improves the global model accuracy,¹
- $R_k^{\text{usage}}(t+1)$ denotes the resource usage (e.g., GPU hours, CPU cycles) client k expends,
- λ_1, λ_2 are weighting coefficients balancing accuracy gains vs. resource contributions.

This formulation encourages nodes that either (i) significantly boost model performance, or (ii) invest substantial computation. One can extend Q_k to account for other factors (e.g., data quality, timeliness) as needed.

4.2 Reward Distribution and Delayed Verification

In a blockchain-based setting, each training round $t+1$ can be associated with a reward pool $R_{\text{pool}}^{(t+1)}$ funded by minted tokens, transaction fees, or a predefined budget. A simple proportional allocation might be

$$R_k^{(t+1)} = \alpha \cdot Q_k^{(t+1)} \cdot R_{\text{pool}}^{(t+1)},$$

where α is a scaling constant controlling the system's overall payout level.

However, to mitigate malicious updates [1], we employ a *delayed reward* mechanism:

$$\text{Final_Payout}(k, t + \Delta T) = f(R_k^{(t+1)}, \text{Verification_Results}(k, t + \Delta T)),$$

where

- $\text{Verification_Results}(k, t + \Delta T)$ includes metrics from cross-validation, anomaly detection, or consensus checks to confirm k 's update is not malicious or low-quality.
- $f(\cdot)$ reduces or nullifies $R_k^{(t+1)}$ if the update is deemed harmful (e.g., a backdoor insertion), mirroring “slashing” concepts in Proof-of-Stake blockchains [5].

Thus, nodes only receive their full reward after a vetting period ΔT , incentivizing honest behavior and discouraging adversarial attacks.

¹ Or alternatively use reduction in loss, $\Delta \text{Loss}_k^{(t+1)}$.

4.3 Accounting for Aggregator Overhead

As the number of participating clients K grows, the server (or aggregator) must handle more frequent and larger model updates. Let $H(t)$ denote the *aggregator overhead function* at round t . For instance:

$$H(t) = \beta \cdot K^\gamma \quad \text{or} \quad H(t) = \beta \cdot \log(K),$$

where β and γ depend on the communication protocol and model size. Higher aggregator overhead can slow training, especially in large-scale FL. While not directly part of each node's reward, $H(t)$ can factor into a *system-level* synergy analysis, as described below.

4.4 Synergy Quantification: Blockchain + FL

To measure the net benefit of integrating blockchain-based rewards into FL, we define a synergy score S comparing two scenarios:

- *Baseline FL*: Standard federated learning without on-chain verification or rewards; final model $\hat{\mathbf{w}}^{(T)}$ after T rounds.
- *BBRS-Enhanced FL*: FL with a blockchain-based reward system; final model $\tilde{\mathbf{w}}^{(T)}$.

We propose:

$$S = [\text{Acc}(\tilde{\mathbf{w}}^{(T)}) - \text{Acc}(\hat{\mathbf{w}}^{(T)})] - \beta_1 [H_{\text{BBRS}} - H_{\text{Baseline}}] + \beta_2 [\text{PartRate}_{\text{BBRS}} - \text{PartRate}_{\text{Baseline}}] + \beta_3 [\text{ResFair}_{\text{BBRS}} - \text{ResFair}_{\text{Baseline}}].$$

Here:

- $\text{Acc}(\cdot)$ is the final model accuracy,
- H_{BBRS} vs. H_{Baseline} denotes aggregator overhead (or total communication cost) in each setup,
- $\text{PartRate}_{\text{BBRS}}$ vs. $\text{PartRate}_{\text{Baseline}}$ measures how many clients actively participate (possibly due to newly introduced incentives),
- ResFair refers to a “resource fairness” index, indicating how well resource contributions correlate with actual rewards (closely related to $Q_k^{(t+1)}$),
- $\beta_1, \beta_2, \beta_3$ weight each term's importance.

A higher S suggests the blockchain-based approach yields better accuracy, higher participation, or fairer resource compensation—even after accounting for extra overhead.

Interpretation – $\text{Acc}(\cdot)$ gain captures improvements in global model quality thanks to honest participation or better-quality updates. – The overhead penalty

$\beta_1[H_{\text{BBRS}} - H_{\text{Baseline}}]$ acknowledges that using a blockchain could increase communication or computational costs. – Additional terms (PartRate and ResFair) highlight how incentives attract more nodes and reward them proportionally to their efforts.

4.5 Discussion

These mathematical building blocks outline how to track and incentivize individual client contributions in a federated setting. By factoring in aggregator overhead, malicious node detection, and delayed rewards, the Blockchain-Based Reward System can systematically bolster FL’s *trustworthiness* and *scalability*. The synergy score S helps evaluate trade-offs in different deployments, guiding future protocol design and system refinements.

5 Potential Benefits and Remaining Challenges

The proposed Blockchain-Based Reward System (BBRS) for Federated Learning (FL) aims to reinforce trust, incentivize active participation, and handle the complexities of non-IID data. While promising, several hurdles must be addressed to achieve a secure, scalable, and sustainable ecosystem. Below, we highlight the advantages and outstanding questions.

5.1 Enhanced Reliability and Trust

- *Transparent Contribution Tracking*: By storing audit trails (e.g., hashed model updates) on a blockchain, both honest participants and oversight entities can verify the legitimacy of local training outputs.
- *Delayed Rewards and Verification*: The use of delayed payouts (cf. Sect. 4.2) discourages malicious behavior, as harmful updates are flagged before final rewards are released.
- *Byzantine Resilience*: Consensus algorithms like PBFT [3] provide fault tolerance even in adversarial conditions, complementing FL’s need for robust aggregation [1].

5.2 Fairness and Incentives

- *Meaningful Rewards*: Nodes that supply high-quality updates or invest substantial computation receive proportionate compensation, as captured by the node contribution metric Q_k (Sect. 4.1).
- *Attracting Broader Participation*: Token-based incentives can motivate resource-constrained or otherwise reluctant clients to join FL, increasing the diversity and volume of data.
- *Resource Fairness*: Incorporating computational effort (GPU/CPU hours) into the reward function helps ensure that participants who shoulder heavier training loads are recognized (Sect. 4.1).

5.3 Tokenomics and Sustainability

- *Reward Pool Dynamics*: Determining the appropriate rate of token issuance ($R_{\text{pool}}^{(t+1)}$) is crucial; too high may cause inflation, while too low may discourage participation.
- *Slashing Mechanisms*: Staking or collateral-based approaches (inspired by Proof of Stake [5]) could penalize malicious nodes by reducing their locked tokens or future reward eligibility.
- *Long-Term Incentive Alignment*: A sustainable token economy should balance rewards with *real* FL improvements, ensuring long-term cooperation rather than short-term speculation.

5.4 Scalability and Communication Overhead

- *Aggregator Bottlenecks*: As the number of participating clients K grows, the communication overhead $H(t)$ (Sect. 4.3) can impede timely model aggregation.
- *Hybrid On/Off-Chain Solutions*: Storing large model updates fully on-chain is infeasible. Instead, on-chain references (e.g., hashes) combined with off-chain bulk storage can mitigate ledger bloat [4].
- *Layer-2 Protocols*: Payment channels or sidechains may further reduce on-chain load while periodically committing aggregated results back to the main ledger for finality.

5.5 *Non-IID Data and Personalization*

- *Heterogeneous Updates*: Divergent local data distributions can slow or bias the global model's convergence [11], challenging straightforward reward schemes.
- *Personalized FL Approaches*: Techniques that fine-tune per-client models or cluster clients with similar data distributions could be integrated with the BBRs framework, potentially adding reward premiums for specialized, yet beneficial contributions.

5.6 *Security and Malicious Actors*

- *Backdoor Attacks*: Nodes injecting backdoor gradients [1] undermine the global model. Delayed rewards and reputation systems must detect and penalize such attempts (Sect. 4.2).
- *Smart Contract Vulnerabilities*: If rewards are managed via on-chain contracts, these contracts themselves can be targets of exploits, demanding rigorous security audits.
- *Sybil Attacks*: Bad actors could spawn numerous pseudo-clients to manipulate rewards. Consensus mechanisms like PBFT or PoS reduce Sybil risk by requiring real stake or computational identity.

5.7 *Regulatory and Ethical Considerations*

- *Data Privacy Compliance*: While FL localizes data, tokenized rewards for cross-border collaborations may raise legal or jurisdictional issues related to data protection (e.g., GDPR).
- *Cryptocurrency Regulations*: Laws regulating digital token issuance and trading vary by region, adding complexity to implementing a robust token economy.
- *Ethical Resource Use*: Energy consumption, hardware usage, and the potential for environmental impact must be weighed against FL's privacy-preserving benefits.

5.8 *Real-World Implementations*

- *IoT and Smart Cities*: Sensor networks across a city could collectively train anomaly detection or traffic forecasting models, receiving rewards for consistent, high-quality updates.

- *Healthcare Collaborations*: Hospitals contribute patient data insights without exposing sensitive records, using on-chain logs to validate secure model exchanges and compensation distribution.
- *Autonomous Vehicles*: Fleets share driving scenarios and model refinements, earning tokens for rare but valuable edge-case data (e.g., extreme weather conditions).

Summary Despite these potential benefits, the road to a fully deployed blockchain-integrated FL ecosystem involves substantial *technical, economic, and regulatory* challenges. Addressing aggregator overhead, malicious node detection, non-IID distributions, and tokenomic sustainability are key tasks for future research. A robust framework must provide verifiable rewards, protect against adversarial manipulation, and ensure the system scales in heterogeneous environments.

6 Conclusion, Vision and Future Directions

This work proposes a **Blockchain-Based Reward System** (BBRS) to strengthen Federated Learning (FL) in terms of trust, participation, and resilience. By uniting blockchain’s transparent, decentralized record-keeping with FL’s privacy-preserving, distributed training paradigm, we aim to create an *open, trustworthy, and autonomous* learning ecosystem.

6.1 Summary

We began by outlining the core challenges of FL, including non-IID data distributions, communication overhead, and a lack of incentives for honest node participation. The proposed BBRS addresses these concerns through:

- *Rewarding Quality and Resource Usage*: Each client’s local update is quantitatively evaluated, incorporating accuracy gains (or loss reduction) and resource consumption.
- *Delayed Payouts for Security*: Verification periods reduce the risk of backdoor or malicious updates [1], mirroring “slashing” concepts in Proof-of-Stake [5].
- *Aggregator Overhead Modeling*: Communication and computation costs in large-scale FL scenarios are explicitly accounted for, enabling more robust system-level design.
- *Synergy Quantification*: A proposed metric S encapsulates gains in model accuracy, participation, and fairness, balanced against increased overhead.

6.2 Vision for an Open, Trustworthy, and Autonomous Learning Model

Looking beyond immediate integration challenges, the overarching goal is a global network where

- *Nodes Participate Freely*: Devices or institutions (e.g., hospitals, IoT sensors, autonomous vehicles) join or leave based on transparent, on-chain incentives and verifiable contribution scores.
- *Trust Emerges from Decentralization*: No single party dictates the process; rather, consensus protocols (e.g., PBFT [3], PoS [5]) ensure all legitimate updates are incorporated.
- *Data Remains Local*: Privacy is preserved as raw data never leaves client devices, aligning with legal frameworks (GDPR, HIPAA) and ethical standards.
- *Self-Sustaining Ecosystem*: Tokenomics spur ongoing model improvement. High-quality, high-effort contributions are rewarded, while poor or malicious behavior is penalized.

In essence, the model fosters an *autonomous collective intelligence*, where participants' incentives align with building robust global models.

6.3 Future Directions

Although initial modeling and simulations show promise, real-world adoption of BBRS-based FL demands further exploration:

- (1) **Large-Scale Pilot Implementations** Applying the proposed system to industry-scale datasets (e.g., healthcare, finance, or city-scale IoT) can reveal unanticipated bottlenecks in consensus overhead, aggregator scaling, or token distribution.
- (2) **Enhanced Malicious Node Detection** Methods like robust aggregation, anomaly detection, or advanced cryptographic techniques (e.g., homomorphic encryption, zero-knowledge proofs) can further thwart adversarial behaviors. Integrating these with *delayed payout* logic remains an active research area.
- (3) **Tokenomics and Regulation** Balancing reward pools, preventing inflation, and aligning with evolving cryptocurrency regulations pose practical hurdles. Smart contract vulnerabilities (e.g., reentrancy or manipulation) also require rigorous audits.
- (4) **Personalized and Fair FL** Future protocols might allocate additional rewards for specialized data distributions or underrepresented groups. Personalized FL could tailor global models to diverse node populations while preserving fairness and avoiding resource imbalances.
- (5) **Layer-2 and Hybrid Designs** High-throughput off-chain (or Layer-2) solutions can reduce ledger overhead, making microtransactions feasible for frequent updates.

Periodic anchoring of FL states to the main blockchain may strike a balance between scalability and security.

6.4 Conclusion

By combining FL's privacy-centric data handling with blockchain's transparent consensus and incentive mechanisms, the BBRs paradigm offers a compelling new direction for collaborative AI. Realizing this vision entails careful attention to technical detail (e.g., aggregator overhead, non-IID data handling), economic design (e.g., token issuance, slashing policies), and regulatory contexts. As these pieces come together, the ultimate outcome is an *open, distributed intelligence* that responsibly harnesses the world's collective data and compute resources to solve complex problems without sacrificing trust, privacy, or fairness.

References

1. Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V (2020) How to backdoor federated learning. In: international conference on artificial intelligence and statistics (AISTATS), vol 108, pp 2938–2948
2. Bonawitz K, Ivanov V, Kreuter B, Marcedone A et al (2017) Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security (CCS), pp 1175–1191
3. Castro M, Liskov B (1999) Practical byzantine fault tolerance. In: Proceedings of the 3rd USENIX symposium on operating systems design and implementation (OSDI), New Orleans, LA, USA. USENIX Association, pp 173–186
4. Hyperledger Project. Hyperledger Fabric Documentation (2021) Hyperledger.org. <https://www.hyperledger.org/use/fabric>
5. Kiayias A, Russell A, David B, Oliynykov R (2017) Ouroboros: a provably secure proof-of-stake blockchain protocol. In: Annual international cryptology conference (CRYPTO). Lecture notes in computer science, vol 10401. Springer, pp 357–388
6. Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D (2016) Federated learning: Strategies for improving communication efficiency. [arXiv:1610.05492](https://arxiv.org/abs/1610.05492)
7. Lu Y, Huang X, Dai X, Maharjan S, Zhang Y (2020) Blockchain empowers federated learning: a survey, framework, and future directions. [arXiv:2007.03782](https://arxiv.org/abs/2007.03782)
8. McMahan HB, Moore E, Ramage D, Hampson S, y Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th international conference on artificial intelligence and statistics (AISTATS), pp 1273–1282
9. Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. Bitcoin.org. <https://bitcoin.org/bitcoin.pdf>
10. Wood G (2014) Ethereum: a secure decentralised generalised transaction ledger. Ethereum project yellow paper. <https://ethereum.github.io/yellowpaper/paper.pdf>
11. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V (2018) Federated learning with non-iid data. [arXiv:1806.00582](https://arxiv.org/abs/1806.00582)

AI-Driven Financial Chart Analysis with Benchmarks: A Domain-Specific Large Language Model Approach



Hyoseok Jang, Sangchul Lee, Haneol Cho, and Chansoo Kim

1 Introduction

Since the release of GPT-3.5, Large Language Models (LLMs) have taken a dramatic leap forward, surpassing previous state-of-the-art achievements in tasks such as reading comprehension, code generation, and creative text composition [1, 2]. This surge in capability has sparked wide-ranging interest: from customer service chatbots and legal document analysis to clinical support systems in healthcare [3, 4]. In the finance sector, in particular, emerging research suggests that LLM-driven tools can significantly bolster market predictions, investment decisions, and automated trading strategies [5, 6]. However, these gains do not always translate seamlessly into advanced financial analytics—where domain-specific expertise, strict regulatory considerations, and high-stakes outcomes demand specialized approaches.

Despite LLMs' remarkable progress, there remains a scarcity of rigorous and quantitative benchmarks that can thoroughly evaluate performance in specialized domains like finance [7, 8]. Many recognized LLM evaluation protocols—initially designed for broad language understanding—offer limited insights into tasks unique to financial markets, such as real-time market intelligence, complex derivatives pricing, and portfolio risk assessments. As an illustrative example, while GPT-3.5 achieved a 92% accuracy on a commonly cited reading comprehension benchmark, its score dropped to 65% on a domain-specific test focusing on derivatives pricing and market risk [8]. This discrepancy underscores the nuanced demands of financial analytics and the limitations of generic evaluation methods.

H. Jang · S. Lee · H. Cho · C. Kim (✉)

AI, Information and Reasoning Laboratory, Computational Science Centre, Korea Institute of Science and Technology, Seoul, Korea

e-mail: eau@ust.ac.kr

H. Jang · C. Kim

Department of AI-Robot, University of Science and Technology, Seoul, Korea

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

P. M. S. Choi and S. H. Huang (eds.), *Finance and Large Language Models*,

Blockchain Technologies, https://doi.org/10.1007/978-981-96-5833-6_10

From another perspective, given the rapid pace at which LLMs continue to evolve—driven by increasingly sophisticated training paradigms, larger parameter sizes, and domain-specific fine-tuning—today’s 65% success rate could soon climb toward near-perfect performance. Much like buying clothes a size larger for a rapidly growing child, we need to anticipate these imminent breakthroughs, developing more demanding tests that remain relevant even as models become significantly more capable. While several efforts have attempted to address the finance domain with specialized benchmarks [9, 10], they still tend to focus narrowly on isolated tasks and rarely capture the full complexity of financial technical chart analysis.

Technical chart analysis itself poses particular challenges for LLMs. Far from being a purely language-based task, it often requires interpreting time-series data, recognizing intricate price-action patterns, and integrating non-textual signals (e.g., volume or order flow) in near-real-time settings. In such high-stakes environments, even minor errors can lead to substantial financial or reputational consequences, emphasizing the need for robust, domain-specific testing and validation.

In light of these considerations, this article proposes the construction of a specialized dataset for financial technical chart analysis, focusing on four key areas: (1) pattern recognition, (2) sentiment integration, (3) anomaly detection, and (4) multi-market correlation. By proactively designing more demanding benchmarks, we aim not only to bridge the current performance gap but also to establish a scalable evaluation framework that evolves alongside future AI advancements. Ultimately, this work seeks to foster the continued growth and reliability of LLM-driven solutions in the increasingly complex world of finance.

2 The Need for Elaborate Benchmarking

As LLMs continue to show remarkable progress in general natural language tasks, many existing evaluation protocols are still rooted in earlier, broadly scoped NLP benchmarks that do not capture the unique—and often urgent—realities of financial markets [7, 8]. In this environment, even seemingly small delays or misinterpretations can lead to material losses. For instance, when earnings reports or regulatory announcements are released, market sentiment can shift dramatically in a matter of hours, sometimes minutes. Models that cannot integrate fresh data and respond appropriately risk providing outdated or misleading outputs. Consequently, timeliness, domain knowledge, and reliability are crucial in finance, yet often overlooked in standardized benchmarks.

A recent survey indicates that fewer than 20% of existing finance-oriented benchmarks incorporate near real-time data feeds or domain-specific parameters such as volatility spreads and leverage ratios [9]. This gap reflects a broader set of challenges:

- **Regulatory Constraints and Compliance:** Financial institutions operate under strict legal frameworks, and any AI-driven recommendation or analysis may be subject to compliance scrutiny. Generic NLP benchmarks do not address the

need for explainability in decision-making processes, especially in contexts where audits, legal liability, and investor protection measures come into play.

- **Rapid Sentiment Shifts and Timely Analysis:** When major events—such as earnings announcements, M&A news, or unexpected geopolitical developments—unfold, sentiment can pivot quickly. LLMs in finance must be able to parse new information on short notice, delivering context-aware insights that remain valid in a rapidly evolving environment.
- **Complex Instruments Beyond Simple Price Predictions:** Real-world finance involves a wide array of products like exotic derivatives, structured notes, and credit-based instruments. Accurately modeling and explaining these instruments demands specialized knowledge that generic NLP benchmarks neither require nor measure [8].
- **Scalability and Vast Amounts of Unstructured Data:** Financial analysts frequently sift through massive volumes of textual information—research reports, social media chatter, regulatory filings—on top of numerical data such as intra-day price movements and macroeconomic indicators. Designing benchmarks that reflect these multimodal, large-scale data requirements poses a level of complexity absent in typical NLP tasks.

Together, these factors underscore the urgent need for more elaborate, finance-specific benchmarks—ones that incorporate near real-time data handling, explainability, and the capacity to tackle advanced instruments. Moreover, the finance sector itself continuously evolves with the emergence of new asset classes (e.g., digital assets, decentralized finance), meaning benchmarks must not only address current gaps but also remain flexible for future developments.

To illustrate how such benchmarks might take shape, the following sections delve into four key pillars—pattern recognition, sentiment integration, anomaly detection, and multi-market correlation—and propose a specialized dataset and evaluation framework built around them. By addressing these domain-specific requirements, we aim to bridge the gap between generic LLM capabilities and the sophisticated demands of institutional investors, regulators, and market participants. Ultimately, a robust, up-to-date suite of tests is essential for ensuring LLMs can perform reliably in the high-stakes world of finance, laying a solid foundation for the strategies we introduce next.

3 Toward a Specialized LLM Evaluation Benchmark in Finance

Recent advances in LLMs highlight their capacity to handle an ever-growing range of tasks, yet their performance on specialized financial problems often lags behind general NLP benchmarks. To address this gap, we propose a **domain-specific evaluation framework** centered on four key pillars essential to technical chart analysis: (1) *pattern recognition*, (2) *sentiment integration*, (3) *anomaly detection*, and (4)

multi-market correlation. By going beyond generic language tests, these pillars aim to capture the real-world complexities of modern trading and risk management.

3.1 *Pattern Recognition*

Traders and analysts frequently rely on chart patterns—such as head-and-shoulders, double or triple tops and bottoms, triangles, wedges, pennants, and flags—to gauge market sentiment and anticipate price movements. These formations can signal trend reversals or breakouts, making them integral to both short-term tactical trades and longer-term strategy decisions [10]. An LLM that accurately identifies such patterns from raw or aggregated (OHLC) data demonstrates a strong grasp of time-series interpretation and market microstructure.

- **Practical Relevance:** In real-world trading, the ability to detect patterns quickly and reliably can yield better risk–reward outcomes and bolster investor confidence.
- **Scope for Scalability:** A robust benchmark might include commonly tracked geometric patterns (e.g., triangles) as well as more advanced or less frequent shapes (e.g., harmonic patterns), ensuring flexibility and extensibility.

3.2 *Sentiment Integration*

While traditional technical analysis focuses on price and volume, market sentiment—shaped by news articles, social media chatter, and public opinion—increasingly drives intraday price moves. For instance, an unanticipated regulatory development or a viral social media post can trigger momentum shifts far faster than would be predicted by technical factors alone. Consequently, LLM-based models capable of integrating textual sentiment signals into their forecasts can deliver deeper, context-aware insights [5].

- **Practical Relevance:** During earnings season or unexpected announcements, sentiment often shifts in a matter of hours or minutes, directly influencing volatility.
- **Broader Benchmark Utility:** A sentiment integration test can differentiate between models that merely recognize positive/negative wording and those that effectively correlate sentiment changes with price action.

3.3 *Anomaly Detection*

Financial markets are prone to rapid swings triggered by unexpected events—such as liquidity shortages, geopolitical shocks, or systemic sell-offs. Past incidents, including flash crashes or abrupt market reactions to global news, highlight the need for *anomaly detection* that extends beyond conventional volatility tracking [8].

- **Core Motivation:** Detecting anomalies helps institutions mitigate losses by identifying early warning signs (e.g., sudden volume spikes, out-of-distribution moves), thereby enabling timely intervention or risk rebalancing.
- **Scalable Complexity:** Benchmarks can measure whether LLMs correctly categorize anomalies (e.g., an illiquid market anomaly versus a news-driven crash) and adapt to shifting conditions across different trading sessions.

3.4 *Multi-market Correlation*

Today's global financial system is highly interconnected: equity indices can move in tandem with commodities under certain market regimes, while emerging-market currencies respond to shifts in oil prices, and digital assets may exhibit unique correlation patterns [11]. Evaluating an LLM's ability to account for these cross-asset relationships tests its capacity to synthesize data beyond a single chart or asset class.

- **Holistic Analysis:** In practice, traders watch multiple markets—bonds, equities, currencies—to gain a fuller picture of market sentiment and liquidity.
- **Benchmark Construction:** This pillar can involve correlation matrices or dynamic correlation scores as ground truth, including scenarios where correlations temporarily break down.

3.5 *Synergies and Future-Proofing*

By bringing these four pillars together, we capture a broad swath of real-world financial challenges that purely text-focused NLP benchmarks overlook. Importantly, each pillar can **scale** with the increasing sophistication of LLMs:

- *Pattern Recognition & Anomaly Detection* naturally converge when unexpected price patterns emerge (e.g., flash crashes).
- *Sentiment Integration* complements technical signals, as abnormal volume or specific chart formations may correlate with sentiment spikes.
- *Multi-Market Correlation* adds a macro-level perspective, revealing whether patterns or sentiment changes in one asset are reinforced or contradicted by another.

Table 1 Overview of the four key pillars for domain-specific LLM evaluation in financial technical chart analysis

		Asset type	
		Single-asset focus	Cross-asset focus
Bias type	Technical bias (price-based)	Pattern recognition	Multi-market correlation
	Fundamental bias (external news)	Anomaly detection	Sentiment integration

This comprehensive approach helps ensure **future-proofing** as financial markets continue to evolve. New asset classes (e.g., digital tokens, carbon credits) or emerging factors (e.g., climate risk, geopolitical realignments) can be integrated into the framework by adding new datasets under one or more of these pillars. Thus, we not only address current performance gaps but also establish a scalable blueprint that can adapt to coming trends—ultimately paving the way for more robust and reliable LLM-driven financial tools (Table 1).

4 Constructing a Technical Chart Analysis Dataset

Building on the capabilities outlined in previous sections, we now turn to the practical steps required to assemble a robust dataset for LLM evaluation in **financial technical chart analysis**. By incorporating diverse market data, carefully chosen indicators, and clear labeling protocols, we aim to capture the multi-faceted nature of real-world trading environments. This section also highlights the importance of *annotation consistency*, offers *example-driven guidance* for each pillar, and briefly addresses how new data sources and advanced correlation structures can be integrated.

4.1 Chart Data

Raw Data (Tick Data) For the most granular view of trading activity, tick-by-tick price and volume data can be collected from high-frequency feeds provided by brokers, exchanges, or reputable data vendors (e.g., Refinitiv, Bloomberg). Although sub-second latency is not always essential for medium- to long-term strategies, having tick data available allows researchers to investigate microstructure effects such as *order book depth* or *bid–ask spreads*. This level of detail can be especially relevant for studies of liquidity shocks or rapid volatility expansions.

OHLC Data A more conventional approach is to aggregate tick data into Open-High-Low-Close (OHLC) bars at various intervals (e.g., 1-minute, 15-minute, hourly, daily). Widely adopted by retail and institutional traders, OHLC data presents a

clearer, noise-filtered view of market momentum [10]. Common data sources include CRSP for equities, ICE Data Services for futures, and Binance APIs for crypto assets, enabling researchers to capture broad market coverage.

Indicator Augmentation To make the dataset richer for technical analysis, each OHLC bar can be supplemented with technical indicators—moving averages, Bollinger Bands, oscillators (e.g., RSI), volume-based measures, and so on. Using a range of parameters (e.g., 15-day versus 50-day moving averages) can help the benchmark reflect diverse trading strategies. More advanced measures such as Ichimoku Clouds or pivot points can also be included, mirroring real-world chart setups while preserving flexibility to accommodate new indicators as markets evolve.

Chart Images For multimodal evaluation—where LLMs may leverage image processing—a subset of data can be converted into **annotated candlestick charts**. These images could contain overlays such as trendlines, support/resistance levels, or volume profiles. Such visual cues are invaluable for testing an LLM’s capacity to recognize price patterns, compare them across different instruments, and integrate textual signals (e.g., news headlines) with *visual* chart insights.

4.2 Labeling for Four Benchmarking Tasks

High-quality, consistent labeling is vital to ensure objective and reproducible evaluations, particularly in finance where subtle misinterpretations can lead to tangible monetary losses. Whenever possible, we recommend using *inter-annotator agreement* metrics (e.g., Cohen’s kappa) to measure label reliability, especially for tasks with subjective boundaries or ambiguous signals [12, 13].

4.2.1 Pattern Recognition

- *Labeling Chart Segments*: Identify and label classic formations (e.g., head-and-shoulders, triangles, flags) across different time frames. For instance, a 1-hour chart showing a head-and-shoulders formation can be marked with exact start and end points.
- *Segmentation*: Consistent segmentation rules (e.g., a minimum length for pattern formation) help minimize label drift. Annotators may also assign a *confidence level* to each pattern for borderline cases.
- *Example Figure*: A labeled screenshot might highlight a head-and-shoulders pattern on a 15-minute chart of the S&P 500, with bounding boxes or colored lines indicating the left shoulder, head, and right shoulder.

4.2.2 Sentiment Integration

- *News and Media Data*: Collect headlines and articles from multiple feeds (e.g., Dow Jones Newswires, social media platforms) linked to specific time windows.
- *Event Tagging & False Positives*: Alongside labeling sentiment (positive, negative, neutral), annotators should mark events where company or asset mentions do not actually influence market sentiment (false positives).
- *Examples*: An earnings-call transcript revealing unexpectedly high costs might be flagged as negative, correlating with a short-term price drop.

4.2.3 Anomaly Detection

- *Unusual Market Events*: Label segments showing sudden volatility spikes, wide price gaps, or drastic volume surges. Real-world examples include the dramatic market plunge on March 16, 2020, when the S&P 500 opened significantly down amid COVID-19 fears.
- *Categorization*: Differentiating anomalies caused by liquidity shortfalls versus macro news versus technical glitches helps models learn to distinguish among varied types of extreme behavior.
- *Interpretability*: Marking the start/end time of an anomaly and the subsequent recovery period offers a clearer temporal footprint for downstream analysis.

4.2.4 Multi-market Correlation

- *Combined Datasets*: Merge price data from multiple markets—equities, commodities, forex, crypto—sourced from platforms like CRSP, ICE, or Binance.
- *Correlation Labels*: Precompute correlation matrices or rolling correlation scores for relevant asset pairs. In addition, note any *time-lagged relationships* (e.g., currency swings that precede or follow equity moves by hours or days) [11].
- *Scenario Coverage*: Include epochs where correlations break down entirely (e.g., a flight-to-safety event) to challenge models relying on stable asset linkages.

4.3 Transition and Outlook

Collectively, these labeling protocols form a holistic view of financial technical chart analysis across multiple data modalities. By balancing objectivity (e.g., clear start/end points, systematic sentiment scoring) with real-world considerations (e.g., ambiguous patterns, time-lagged correlations), the resulting dataset aligns closely with the

Table 2 Summary of chart data components (Sect. 4.1)

Data type	Key characteristics	Use cases/notes
Tick data	<ul style="list-style-type: none">– Sub-second or per-trad data– Highest granularity– Potentially large volume	<ul style="list-style-type: none">– Captures microstructure details (e.g., order book depth)– Useful for liquidity or rapid volatility studies
OHLC data	<ul style="list-style-type: none">– Aggregated Open-High-Low-Close at fixed intervals (1 m, 15 m, 1h, 4h, daily, weekly...)– Noise-filtered view of price momentum	<ul style="list-style-type: none">– Widely adopted by traders– Simpler than tick data, but loses intra-bar detail
Indicator augmentation	<ul style="list-style-type: none">– Adds technical indicators (MA, Bollinger Bands, RSI, etc.)– Flexible parameter choices (15-day, 50-day, etc.)	<ul style="list-style-type: none">– Mimics common real-world chart setups– Enhances pattern recognition tasks
Chart images	<ul style="list-style-type: none">– Visual candlestick or bar chart snapshots– Possible overlays: trendlines, support/resistance	<ul style="list-style-type: none">– Enables multimodal LLM or CV+NLP– Tests pattern recognition in a visual context

demands of practitioners. In the next section, we detail how this labeled data integrates into an evaluation protocol, including metrics, stress-testing procedures, and baseline comparisons for LLM-based models. Such a framework will not only validate current performance but also adapt as financial markets and AI techniques continue to evolve (Tables 2 and 3).

5 Benchmarking LLMs for Chart Analysis

With a domain-specific dataset in place, we now focus on designing an evaluation protocol that thoroughly measures LLM performance across the four pillars introduced in Sect. 3—*pattern recognition*, *sentiment integration*, *anomaly detection*, and *multi-market correlation*. The goal is a benchmark that not only reports raw accuracy metrics but also reflects the practical complexities of finance, including multi-step reasoning, interpretability demands, and exposure to dynamic market conditions.

Table 3 Key labeling considerations for the four benchmarking tasks (Sect. 4.2)

Task	Labeling focus	Example/special notes
Pattern recognition	<ul style="list-style-type: none">– Identify chart formations (head-and-shoulders, triangles, etc.)– Mark pattern boundaries and confidence levels	<ul style="list-style-type: none">– E.g., label left/right shoulder points in head-and-shoulders– Capture partial/ambiguous patterns with caution
Sentiment integration	<ul style="list-style-type: none">– Annotate news or social media items with sentiment scores– Mark event relevance and false positives	<ul style="list-style-type: none">– E.g., earnings call triggers a “negative” label– Distinguish truly influential news versus irrelevant mentions
Anomaly detection	<ul style="list-style-type: none">– Tag sudden volatility spikes, price gaps, unusual volume– Add context for cause (liquidity short-fall, major news, glitch)	<ul style="list-style-type: none">– E.g., labeling flash crash on a specific date/time– Mark start/end times + recovery period
Multi-market correlation	<ul style="list-style-type: none">– Correlation metrics among equities, forex, commodities, crypto– Track correlation regime shifts or break- downs	<ul style="list-style-type: none">– E.g., a rolling correlation window for S&P 500 versus Gold– Note time-lagged effects (leading/lagging correlation)

5.1 Key Considerations

5.1.1 Multi-dimensional Scoring Metrics

Each of the four pillars benefits from tailored metrics. For classification-driven tasks (e.g., identifying specific chart patterns or labeling sentiment), F1-scores or precision–recall curves can be more informative than simple accuracy. Anomaly detection tasks should capture both *timeliness* (how quickly anomalies are flagged) and *severity* (e.g., large price gaps versus minor blips). In multi-market correlation, measuring *the accuracy of predicted correlations* or *the error in dynamic correlation transitions* is key [11]. Additionally, real-world regulatory environments often demand **explainability** [10], making interpretability or auditability metrics highly relevant.

5.1.2 Multi-step Evaluation Pipeline

Unlike many NLP tasks that focus on a single, static input (e.g., a passage of text), financial LLM applications often require sequential or iterative processes. We recommend dividing the evaluation into incremental stages:

1. *Data Parsing*: Assess the model’s ability to handle raw or OHLC data and technical indicators.

2. *Pillar-Specific Tests*: Evaluate core competencies in pattern recognition, sentiment integration, anomaly detection, and correlation analysis independently.
3. *Integrated Tasks*: Challenge the model to generate trading signals, risk assessments, or scenario analyses that blend the four pillars.

Such a pipeline approach can reveal which part of the workflow remains most challenging for an LLM, guiding focused improvements or additional fine-tuning efforts.

5.1.3 Stress Testing

To gauge robustness, we propose subjecting models to **extreme or conflicting market conditions**. For instance:

- *Volatile Periods*: Historical segments with large price swings, such as sudden liquidity drops or shock events (e.g., major policy announcements).
- *Data Gaps*: Partial data availability or missing feeds, reflecting real-world issues like system outages.
- *Conflicting Signals*: Situations where chart patterns suggest a bullish trend, yet social media sentiment is overwhelmingly negative.

Such scenarios help distinguish LLMs that only excel under stable market conditions from those capable of adaptive, context-aware reasoning.

5.1.4 Model Comparison and Reproducibility

We encourage benchmarking multiple LLM architectures—from fine-tuned GPT-3.5 variants to specialized Transformers designed for time-series data—to measure relative strengths and weaknesses. Transparent reporting of

- *Hyperparameters and Training Protocols*: Learning rates, batch sizes, optimizer settings.
- *Hardware Environment*: GPU/TPU types, memory constraints.
- *Data Splits and Preprocessing*: Consistent train/validation/test partitions, date ranges, and normalization steps.

ensures that results can be replicated. Such reproducibility is crucial in finance, where model-related decisions carry substantial monetary and regulatory implications.

5.2 Beyond Raw Metrics

While quantitative metrics (e.g., F1, MSE, correlation error) are important, financial institutions often require *interpretability*, *compliance*, and *operational feasibility* in real-world deployments. For instance:

- **Interpretability:** Regulators may require clear audit trails or model explanations for trading decisions, especially under stress scenarios [10].
- **Regulatory Alignment:** Compliance frameworks (e.g., MiFID II in Europe, FINRA in the U.S.) could mandate disclosures about how AI-driven recommendations are generated.
- **Computational Efficiency:** In fast-moving markets, latency can matter. Benchmarking inference speed or resource usage ensures scalability across varied trading infrastructures.

Incorporating these considerations makes the benchmark more realistic and aligns the measured performance with genuine financial-sector needs.

In sum, a robust benchmark for LLMs in technical chart analysis must extend beyond standard accuracy metrics to encompass stress-tested scenarios, interpretability, regulatory factors, and multi-step analytical processes (Fig. 1).

6 Expert Collaboration and Ongoing Directions

Constructing and maintaining a specialized financial dataset demands more than just technical know-how; it requires **continuous engagement** with market practitioners, risk managers, and industry regulators. Their expertise ensures that labeling criteria remain accurate, new market realities are swiftly incorporated, and the benchmark itself remains aligned with evolving regulations and best practices.

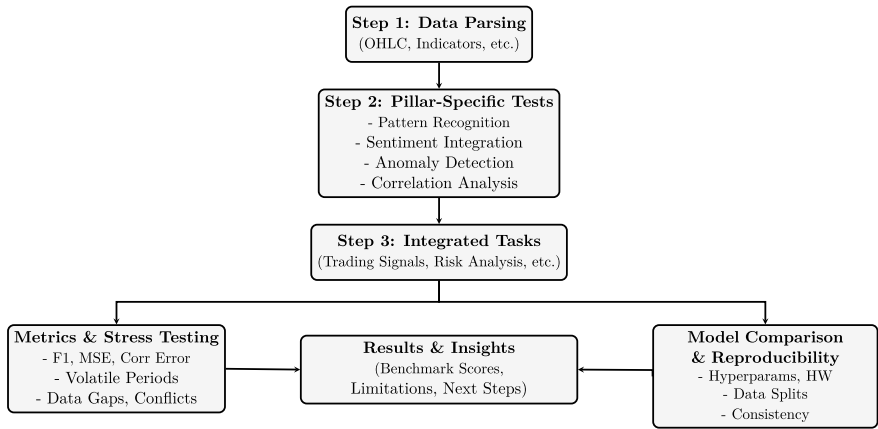


Fig. 1 A multi-step LLM evaluation pipeline with parallel processes, resized to fit within A4

Role of Financial Experts: Domain experts can play a pivotal role at every stage:

- **Annotation and Quality Control:** Beyond identifying common patterns or anomalies, experienced traders and analysts can refine label definitions for borderline cases (e.g., partial formations, conflicting sentiment) and validate outlier events.
- **Scenario Selection and Validation:** Practitioners with first-hand experience of historical market events (e.g., sudden regulatory announcements, major credit shocks) can guide which time segments best test an LLM's adaptive capacity.
- **Regulatory and Compliance Insights:** Legal specialists and compliance officers help ensure that models' outputs meet transparency and accountability requirements, reflecting the realities of financial oversight (e.g., MiFID II, FINRA guidelines).

Continuous Updating and Expansion: Financial markets constantly introduce novel instruments (e.g., tokenized securities, carbon credits) and face emerging risk factors (e.g., geopolitical shifts, climate-related impacts). To maintain relevance:

- **Regular Data Refreshes:** Instituting a rolling update mechanism allows the dataset to integrate *recent* events, ensuring that benchmarks challenge LLMs with current trends or regime shifts.
- **Feature Enhancements:** As new analytic techniques (e.g., alternative data sources, advanced volatility measures) gain traction, expert consultation ensures that these dimensions are captured in future labeling efforts.
- **Iterative Benchmark Improvement:** Even well-established patterns can evolve, and new patterns can emerge in response to changing market structure. Periodic reviews with domain experts guard against dataset stagnation.

Roadmap for Ongoing Collaboration: To streamline expert engagement, we propose

1. **Scheduled Review Cycles:** Biannual or quarterly workshops where domain experts, data scientists, and compliance officers evaluate annotation quality, discuss newly identified anomalies, and propose expansions to coverage.
2. **Open-Source Contribution Channels:** Hosting the dataset on a public repository (with anonymized data samples, if needed) can facilitate community-driven improvements, subject to appropriate data licensing.
3. **Industry-Academic Partnerships:** Joint initiatives between financial institutions and academic labs can accelerate robust labeling processes, risk analysis, and method validation under real-world constraints.

By weaving expert insights into every iteration of dataset development, we ensure that our finance-specific benchmark remains **up-to-date, comprehensive, and practical**. This strategy aligns with the rapid evolution of LLMs and the unceasing transformation of financial markets, fostering a long-term synergy between technical advancement and real-world applicability.

7 Conclusion

This paper has highlighted the vital need for domain-specific benchmarks in financial technical chart analysis, demonstrating how current LLM evaluation methods can fall short when confronted with the complexities of real-world markets. We introduced a four-pillar framework encompassing *pattern recognition*, *sentiment integration*, *anomaly detection*, and *multi-market correlation*—each reflecting a core challenge faced by practitioners and regulators. By outlining the construction of a specialized dataset, complete with chart-based annotations, sentiment-labeled events, anomaly tags, and cross-asset correlations, we provided a roadmap for more meaningful and robust AI assessments.

Beyond improving raw accuracy measures, this approach emphasizes **interpretability**, **regulatory compliance**, and **adaptability** to a rapidly evolving financial ecosystem. The proposed dataset design ensures that new asset classes, emerging risk factors, or dynamic trading patterns can be integrated over time. Crucially, our findings underscore the importance of ongoing collaboration with industry experts and compliance officers to maintain labeling consistency, update scenario coverage, and validate anomalous market events.

Although we primarily focus on technical chart analysis, the methodology is readily extensible to broader financial tasks, from quantitative risk modeling to algorithmic portfolio allocation. We envision that future iterations of our benchmark will incorporate additional data modalities (e.g., alternative data feeds, textual transcripts of earnings calls) and explore advanced evaluation metrics suited to real-time decision-making contexts.

References

1. Bubeck S, Chandrasekaran V, Eldan R et al (2023) Sparks of artificial general intelligence: early experiments with gpt-4. [arXiv:2303.12712](https://arxiv.org/abs/2303.12712)
2. OpenAI (2022) Gpt-3.5: Technical overview and performance benchmarks. <https://openai.com/research/gpt-3-5>
3. Nay J (2021) Natural language processing and law: annotated corpus creation for legal NLP applications. *Artif Intell Law* 29(1):47–73
4. Shen J, Zhang C, Jiang X (2022) Adopting LLMs in healthcare: a survey of use cases, benefits, and challenges. *Health Inform J* 28(3):205–217
5. Cheng W, Fang W (2023) Language models in finance: from sentiment analysis to automated trading systems. *Quant Financ* 23(2):153–172
6. Yang Y, Sung Y (2023) Predictive power of large language models in equity markets: an empirical study. *J Financ Data Sci* 5(1):66–75
7. Lopez R, Sung M, Jin Y (2023) Beyond accuracy: a framework for assessing LLM deployment in financial risk management. In: *Proceedings of the 40th international conference on machine learning*
8. He L, Wang G (2023) Challenges in benchmarking AI for financial applications: a critical review. *J Comput Financ* 17(3):210–228
9. Wang X, Li F, Zhou J (2023) A survey of real-time financial data benchmarks for AI applications. *J Quant Financ* 12(4):89–102

10. Lo AW, Mamaysky H, Wang J (2000) Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. *J Financ* 55(4):1705–1765
11. Engle R (2002) Dynamic conditional correlation: a simple class of multivariate GARCH models. *J Bus & Econ Stat* 20(3):339–350
12. Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20(1):37–46
13. Artstein R, Poesio M (2008) Inter-coder agreement for computational linguistics. *Comput Linguist* 34(4):555–596