Large Language Models for Financial and Investment Management: Models, Opportunities, and Challenges

Yaxuan Kong, Yuqi Nie, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren

Yaxuan Kong
is a PhD student in the Department of Engineering Science at the University of Oxford in Oxford, UK.
yaxuan.kong@eng.ox.ac.uk

Yuqi Nie
is a PhD student in the Department of Electrical and Computer Engineering at Princeton University in Princeton, NJ.
ynie@princeton.edu

Xiaowen Dong
is an associate professor in the Department of Engineering Science at the University of Oxford in Oxford, UK.
xiaowen.dong@eng.ox.ac.uk

John M. Mulvey
is a professor in the Department of Operations Research and Financial Engineering at Princeton University in Princeton, NJ.
mulvey@princeton.edu

H. Vincent Poor
is the Michael Henry Strater University Professor in the Department of Electrical and Computer Engineering at Princeton University in Princeton, NJ.
poor@princeton.edu

Qingsong Wen
is the head of AI research and chief scientist at Squirrel Ai Learning in Bellevue, WA.
qingsongedu@gmail.com

Stefan Zohren
is an associate professor in the Department of Engineering Science at the University of Oxford in Oxford, UK.
stefan.zohren@eng.ox.ac.uk

KEY FINDINGS

- Efforts are underway to evaluate the usefulness of LLM technology throughout the financial industry. These studies require massive investments in human capital, data collection, computer resources, and even electrical power.
- There are multiple opportunities to improve the efficiency and quality of investment management. Prominent examples include increasing productivity, providing a natural language interface for customers and employees, and gaining the public's confidence in the soundness of financial institutions.
- Several challenges exist for wide usage of the LLM technology, including identifying fake information, privacy concerns, addressing situational awareness, and the concentration of data in a few firms.

ABSTRACT

The intersection of artificial intelligence (AI) and financial management has gained significant attention, particularly with the rise of large language models (LLMs). These models process vast amounts of unstructured data, offering powerful tools for financial analysis and investment decision-making. This article explores the use of LLMs in finance, focusing on recent advancements, models, and technologies, while addressing the opportunities and challenges they present. It highlights the strengths and limitations of finance-specific models in handling complex tasks and identifies key challenges such as data issues, modeling complexities, and ethical concerns, which also present opportunities for innovation. The article provides a comprehensive overview of LLMs in finance, underscoring their potential to transform the field while emphasizing the need to carefully consider their limitations and risks. The integration of LLMs into financial decision making holds significant promise, offering new possibilities for research and practical applications.

The integration of large language models (LLMs) into financial and investment management represents a significant paradigm shift in how the industry processes and leverages information. As artificial intelligence (AI) continues to evolve, LLMs stand out for their remarkable ability to understand, generate, and manipulate human-like text, opening up new frontiers in financial analysis and decision-making.

The integration of LLMs into financial and investment management represents more than just a technological upgrade; it signifies a potential paradigm shift in how we interpret market dynamics, assess risks, and formulate investment strategies. Unlike traditional quantitative models, LLMs offer a unique ability to process unstructured data, discern subtle contextual nuances, and generate insights that blend numerical analysis with qualitative understanding. This capability is particularly valuable in an industry where success often hinges on interpreting complex, multifaceted information streams.

However, the application of LLMs in finance is not without its challenges. The dynamic nature of financial markets, the critical importance of accuracy in financial predictions, and the need for interpretability in decision-making processes all pose unique hurdles. Moreover, the ethical implications of deploying such powerful AI tools in a sector that directly impacts global economies and individual livelihoods cannot be overlooked.

This survey aims to provide a comprehensive overview of the current state of LLMs in financial and investment management. We explore the various models that have been developed specifically for financial applications, examining their architectures, capabilities, and limitations. We also delve into the fundamental reasons why LLMs are particularly well-suited for financial tasks, highlighting their potential to transform various aspects of the industry.

As we navigate through this landscape, we critically assess both the challenges and opportunities that arise from the intersection of LLMs and finance. From data quality issues and model interpretability to ethical considerations and regulatory compliance, we examine the multifaceted implications of this technological integration. Ultimately, this survey seeks to illuminate the path forward for researchers, practitioners, and policymakers in harnessing the power of LLMs to enhance financial decision-making while addressing the associated risks and ethical concerns. By providing a balanced perspective on the current state and future potential of LLMs in finance, we hope to contribute to the ongoing dialogue about the responsible and effective use of AI in shaping the future of financial and investment management.

MODELS

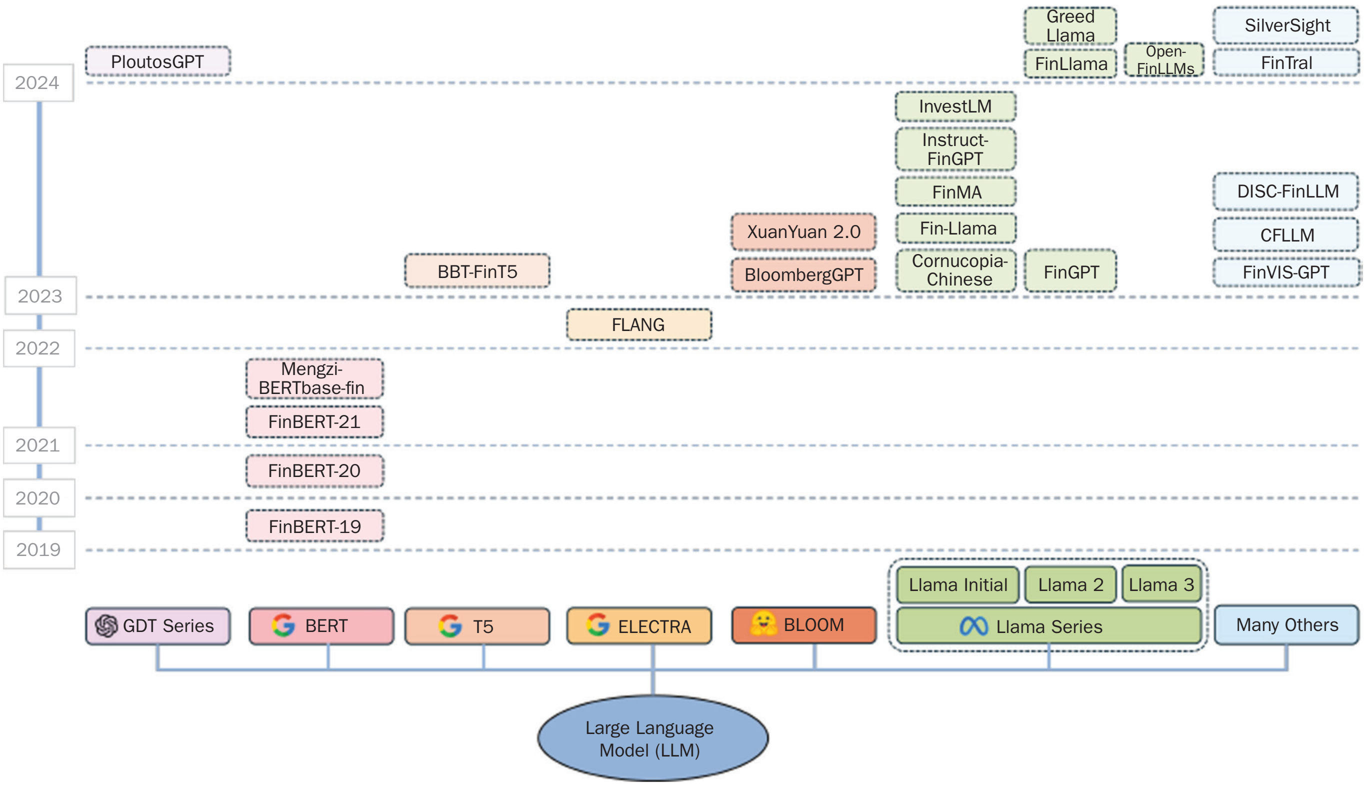
LLMs have demonstrated remarkable capabilities across a wide range of domains (Wu et al. 2023; Liu et al. 2023; Wang, Xu et al. 2024). While general-domain LLMs such as GPT-series, Llama-series, and BERT have shown impressive performance on various NLP tasks, there has also been growing interest in developing financial domain-specific LLMs. These specialized models are trained on vast amounts of financial data, allowing them to better understand and generate content related to finance, economics, and business.

We will begin by introducing several prominent financial domain-specific LLMs, exploring their strengths, limitations, and the potential in downstream financial tasks. A summary of these models is provided in Exhibit 1.

GPT-series. One of the most well-known general-domain LLMs is the GPT (generative pretrained transformers) series, developed by OpenAI (Radford, Narasimhan, Salimans, and Sutskever 2018; Radford, Wu et al. 2019; Brown et al. 2020; Achiam et al. 2023). GPT models, based on the transformer architecture, leverage self-attention mechanisms and positional embeddings to capture long-range dependencies in text. Recently, **Ploutos** (Tong et al. 2024), a novel financial LLM framework derived from GPT-4, has been proposed for interpretable stock movement prediction. Ploutos consists of two main components: PloutosGen and PloutosGPT. PloutosGen addresses the challenge of fusing textual and numerical information by integrating

EXHIBIT 1

An Overview of Financially Specialized LLMs



NOTE: The financial LLMs listed were released or developed from 2019 onward and are categorized by their foundational model types and many others.

multimodal data through a diverse expert pool, including sentiment, technical, and human analysis experts, which generate quantitative strategies from different perspectives. Conversely, PloutosGPT tackles the lack of clarity in traditional methods by using rearview-mirror prompting, which leverages historical stock data and expert analysis to guide the model and dynamic token weighting to generate accurate and interpretable rationales for stock predictions. While Ploutos demonstrates enhanced prediction accuracy and interpretability, it is constrained by potential expert selection bias, computational complexity, and limited data types. Future research could potentially focus on optimizing efficiency, expanding data variety, and mitigating biases to further improve the framework's performance.

BERT. In 2018, bidirectional encoder representations from transformers, known as BERT (Devlin et al. 2018), revolutionized the field of NLP with its deep bidirectional architecture that could learn contextual representations. This breakthrough led to the development of several domain-specific variants, particularly in the financial sector. Building upon BERT's foundation, **FinBERT-19** (Araci 2019) was developed by continually pretraining BERT on financial text to enhance its sentiment analysis capabilities. The following year, **FinBERT-20** (Yang, Uy, and Huang 2020) further evolved this approach by conducting domain-specific pretraining from scratch, focusing solely on financial communications and utilizing a large-scale financial corpus. In 2021, **FinBERT-21** (Liu et al. 2021) introduced a mixed-domain pretraining strategy, leveraging both general corpora (Wikipedia and BooksCorpus) and financial domain corpora (FinancialWeb, YahooFinance, and RedditFinanceQA). By simultaneously training on general and financial domain corpora, FinBERT-21 aims to capture a broader range of language knowledge and semantic information relevant to financial text mining. These FinBERT models have demonstrated their effectiveness in various financial downstream tasks, such as sentiment analysis, named entity recognition, question answering, and text classification within the financial domain. In addition to the Fin-BERT models already mentioned, RoBERTa (Liu et al. 2019), introduced in 2019, is another variant of BERT. **Mengzi-BERT base-fin** (Zhang et al. 2021), trained with 20GB of financial news and research reports, is a specialized version of RoBERTa designed for financial applications.

Text-to-text transfer transformer (T5). In 2019, Google introduced the text-to-text transfer transformer (Raffel et al. 2020), a unified framework that treats every text processing task as a "text-to-text" problem. The T5 model utilizes an encoder-decoder architecture and is pretrained using a self-supervised learning objective called "span corruption." This involves randomly masking contiguous spans of text in the input sequence and training the model to reconstruct the original text. Building on this, the **BBT (Big Bang Transformer)-FinT5** (Lu et al. 2023) was developed specifically for the Chinese financial sector. This model incorporates knowledge-enhanced pretraining methods and is built on the BBT-FinCorpus—a large-scale financial corpus comprising diverse sources, including corporate reports, analyst reports, social media, and financial news. BBT-FinT5 benefits from the text-to-text framework of T5, allowing it to tackle both language understanding and generation tasks within the financial domain. However, being a domain-specific model, its performance on general NLP tasks outside of finance might be limited. BBT-FinT5 can be fine-tuned for various financial applications including news classification, summarization, relation extraction, sentiment analysis, and event-based question answering.

ELECTRA. In 2020, ELECTRA (Clark et al. 2020) introduced an innovative generator-discriminator framework for pretraining language models. The model improves efficiency by training the discriminator to distinguish between real and synthetically generated tokens. Building upon this, researchers developed **FLANG** (Shah et al. 2022), a specialized variant of ELECTRA tailored for the financial domain. FLANG integrates specific adaptations such as selective token masking and span boundary objectives to effectively handle the complexities of financial language. While FLANG excels in handling financial terminology and delivers enhanced performance on tasks such as sentiment analysis and entity recognition within financial documents, its specialization may limit its effectiveness in nonfinancial contexts without further fine-tuning. Despite this limitation, FLANG has demonstrated its value in various downstream financial tasks. It enables precise analysis of market reports, accurate classification of financial headlines, and reliable identification of key financial entities.

BLOOM. In 2022, BLOOM (Le Scao et al. 2023) was released as a fundamental multilingual LLM with 176 billion parameters. It was pretrained on a vast corpus of text that included 46 natural languages and 13 programming languages. BLOOM is notable for its diversity and accessibility as an open-source model that supports a variety of languages. From BLOOM, specialized versions focused on financial applications have been created, including **BloombergGPT** (Wu et al. 2023) and **XuanYuan 2.0** (Zhang and Yang 2023). BloombergGPT, with its 50 billion parameters, was designed for the financial sector by training on Bloomberg's financial data sources. This model demonstrates enhanced performance on specific financial tasks while maintaining competitive overall competence. XuanYuan 2.0, created for the Chinese financial market, is a large open-source Chinese financial chat model. It proposes a novel hybrid-tuning strategy that combines general and finance-specific data, allowing the model to retain general language capabilities while exceling at domain-specific tasks such as financial advisory and market analysis. This strategy lowers the likelihood of catastrophically forgetting previous knowledge and enhances accuracy on finance-related tasks.

Llama-series. Llama (Touvron, Lavril et al. 2023), an LLM introduced in 2023, offers flexibility with model sizes ranging from 7B to 65B parameters. Trained on publicly available datasets for transparency, Llama outperforms larger models, including GPT-3, on most benchmarks despite its smaller size. Its financial variants, which include **FinMA** (Xie et al. 2023), **Fin-Llama** (William Todt 2023), **Cornucopia—Chinese** (Yu 2023), **Instruct-FinGPT** (Zhang, Yang, and Liu 2023) and **InvestLM** (Yang, Tang, and Tam 2023), provide specialized capabilities for various financial tasks. InvestLM, based on LLaMA-65B and a diverse investment-related dataset, offers investment recommendations comparable to cutting-edge commercial models. Llama 2 (Touvron, Martin et al. 2023), which was released later, included various enhancements over Llama, including a 40% larger pretraining corpus, a doubled context length, and the adoption of grouped-query attention for improved inference scalability. It has such financial variants as **FinGPT** (Yang, Liu, and Wang 2023), **FinLlama** (Konstantinidis et al. 2024), and **GreedLlama** (Yu, Huber, and Tang 2024). Particularly, FinGPT is an open-source model that focuses on providing accessible and transparent resources for developing financial LLMs. Despite having relatively small training data compared with BloombergGPT, FinGPT claims to offer a more accessible, flexible, and cost-effective solution for financial language modeling. In April 2024, Meta introduced Llama 3 (Meta AI 2024), featuring 8B and 70B parameter models that showcase state-of-the-art performance and improved reasoning capabilities, marking them as the most capable openly available LLM to date. The LLM community is visibly excited, and one of the notable examples is the design of the Open-FinLLMs (Xie et al. 2024). We expect more Llama 3 variants for financial LLM models to emerge soon.

In addition to the models mentioned already, there are also other financial domain-specific LLMs such as FinTral (Bhatia et al. 2024), driven from Mistral 7B (Jiang et al. 2023); SilverSight (Zhou, Li et al. 2024), based on the Qwen 1.5-7B chat model (Bai et al. 2023); DISC-FinLLM (Chen et al. 2023), used Baichuan-13B (Baichuan-inc 2023) as the backbone; CFLLM (Li, Bian et al. 2023), based on InternLM-7B (InternLM 2024); FinVIS-GPT (Wang, Li et al. 2023), a multimodal LLM for financial chart analysis based on LLaVA (Liu, Li et al. 2024). These domain-specific LLMs utilize vast financial datasets and advanced training techniques to provide more accurate and context-aware financial analysis than general-domain models. As research in this area continues to progress, we expect the development of even more sophisticated financial LLMs that could transform various sectors of the financial industry, including investment strategies, risk management, forecasting, and customer service. However, it is crucial to acknowledge the limitations and potential biases of these models and to employ them thoughtfully alongside human expertise and judgment.

Zero-Shot vs. Fine-Tuning

Zero-shot and fine-tuning are two distinct adaptation methods in the application of LLMs. Zero-shot (or few-shot) learning refers to the ability of a model to correctly predict or perform tasks it has not explicitly been trained to handle, based on its pre-existing knowledge and generalization capabilities. Fine-tuning, in contrast, involves adjusting a pretrained model on a specific dataset or for a particular task to improve its accuracy and performance on that task (Li, Wang et al. 2023).

Fine-tuning is favored when domain-specific accuracy is essential, adaptability to real-time changes is required, or customization and privacy are critical considerations. In practice, the integration of financial-related text data is a common approach in fine-tuning LLMs. Araci (2019) develops FinBERT, a tailored version of the BERT language model, achieved through extended pretraining on a comprehensive financial dataset, including news, articles, and social media posts, alongside strategic fine-tuning methods. FinBERT sets a new benchmark in finance-related text analysis, eclipsing earlier deep learning methodologies in the field.

Several technologies have been proposed to make fine-tuning more efficient. Instruction tuning (Wei et al. 2021) is a fine-tuning method for language models where the model is trained to follow specific instructions; it not only improves performance on the target tasks but also enhances the model's zero-shot and few-shot learning capabilities, making it popular among various financial applications and models. Zhang, Yang, and Liu (2023) propose an instruction-tuned FinGPT model that enhances the financial sentiment analysis capabilities of LLMs by employing instruction tuning, which transforms a small portion of supervised financial sentiment data into instruction data, thereby improving the model's numerical sensitivity and contextual understanding. Furthermore, Zhang et al. (2023) integrate instruction-tuned LLMs with a retrieval-augmentation module, which is a technique that enhances language models by supplementing their input with relevant information retrieved from external sources, to enhance the models' predictive performance by providing a richer context. Besides instruction tuning, people have also applied low-rank adaptation (LoRA; Hu et al. 2021) or quantized LLMs (Ma et al. 2024; Dettmers et al. 2024) for more efficient adaptation on financial tasks, such as FinGPT (Zhang et al. 2023), FinGPT-HPC (Liu, Zhang et al. 2024), and Llama-based models (Pavlyshenko 2023).

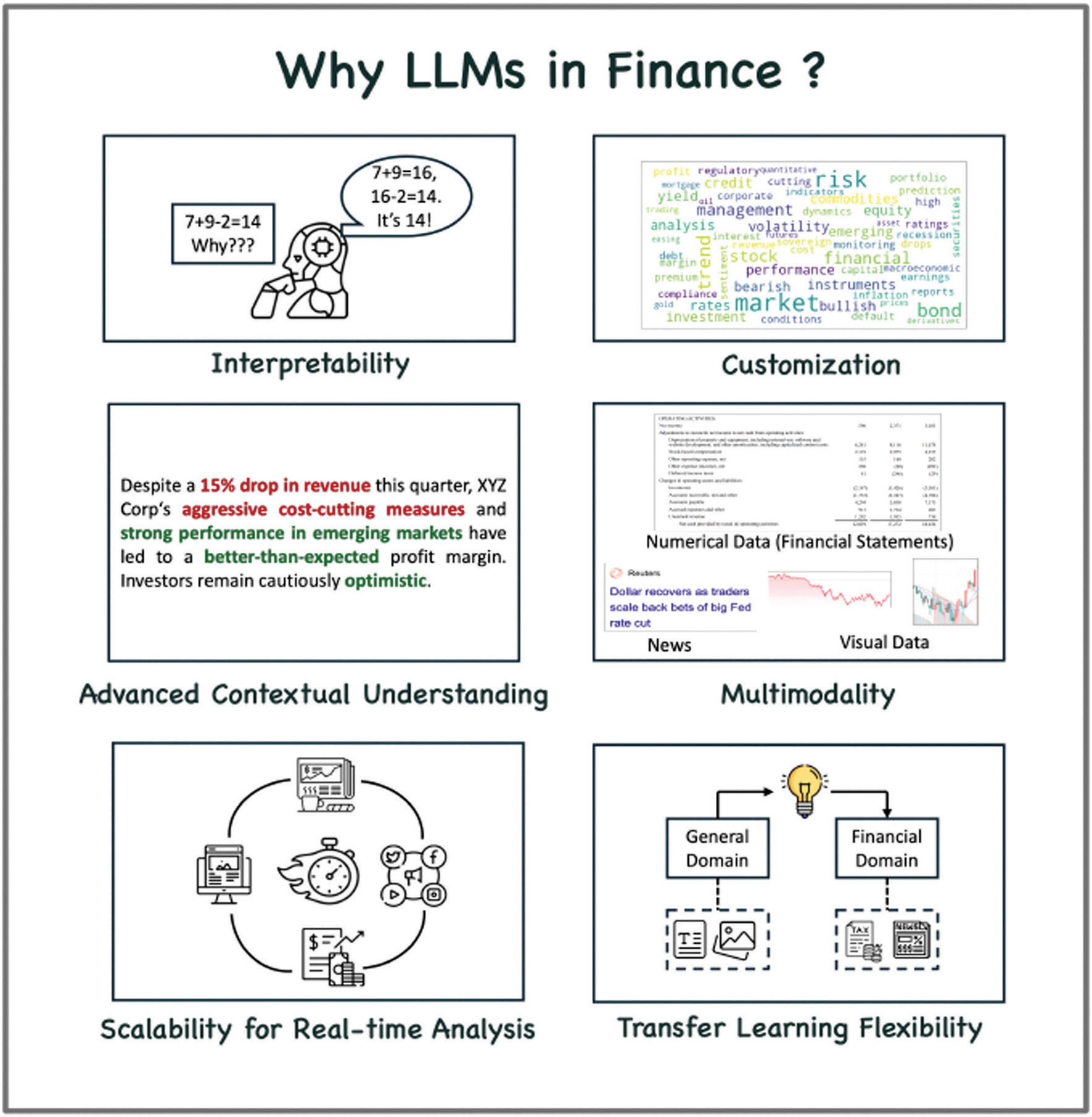
Another prevalent approach involves the consideration of smaller models, as energy efficiency and the lightweight nature of models are crucial in today's machine learning landscape (Yakar et al. 2019; Nie and Yuan 2021; Wang, Sun, and Boukerche 2022). Rodriguez Inserte et al. (2024) demonstrate that smaller LLMs can be effectively fine-tuned on financial documents and instructions to achieve comparable or superior performance to larger models. Deng et al. (2023) present a case study on utilizing an LLM for semi-supervised financial sentiment analysis on Reddit data, where the LLM generates weak sentiment labels through in-context learning and chain of thought reasoning, which are then used to train a smaller model for production use, achieving competitive performance with minimal human annotation.

While pretraining and fine-tuning allow these models to adapt to the specific linguistic characteristics and styles of various applications, zero-shot learning is preferred when labeled data are limited, rapid deployment is crucial, or modular development and interpretability are prioritized. The zero-shot and few-shot capabilities of LLMs underscore their efficiency by allowing for direct application without the need for extensive dataset-specific training. This efficiency is due to the transfer learning from the vast datasets on which LLMs are trained, as well as their emergent abilities to generate new insights or address unexpected problems during information processing (Wei et al. 2022). These features significantly broaden their usefulness across various fields without the need for further training. For instance, Steinert and Altmann (2023) explore the zero-shot capability of GPT-4 to predict same-day stock price movements of Apple and Tesla in 2017 with microblogging messages, and by comparing its performance to BERT, they highlight the importance of prompt engineering in extracting sophisticated sentiments from GPT-4 for financial applications.

Why Apply LLMs in Finance?

The integration of LLMs into financial analysis marks a fundamental transformation in how the financial sector approaches analysis and strategy. These models, underpinned by advanced machine learning techniques, offer unprecedented capabilities in processing and interpreting natural language, allowing for the extraction of insights from vast and complex data sources. This transformative potential makes LLMs compelling tools for enhancing analytical approaches in the industry. In the following discussion, we will explore several core reasons (as shown in Exhibit 2)

EXHIBIT 2
Why Apply LLMs in Finance?



for using LLMs in financial applications, setting the stage for a more detailed examination of their broader implications and specialized roles in reshaping financial analysis.

Advanced contextual understanding. LLMs are distinguished by their profound ability to understand context. This includes a comprehensive understanding of financial terminologies, jargon, and refined expressions. Such advanced contextual understanding significantly enhances the accuracy of sentiment analysis, a critical aspect when dealing with the complex and often ambiguous language found in financial documents and news articles.

Transfer learning flexibility. LLMs are initially pretrained on a vast corpus of internet text, encompassing a wide range of topics and languages. This pretraining equips LLMs with a broad understanding of language, which can then be fine-tuned for specific financial tasks. Such flexibility in transfer learning reduces the reliance on large, domain-specific datasets, allowing for efficient adaptation to new tasks with minimal domain-specific training data in finance.

Scalability for real-time analysis. The financial market's fast-paced nature demands tools that can offer timely insights. LLMs excel in processing large volumes of text rapidly, enabling real-time reasoning and sentiment analysis. This capability ensures that financial decision-makers can receive immediate insights from news articles, market information, reports, and social media, facilitating more informed and timely decisions.

Multimodality. Multimodal capabilities of LLMs extend their application beyond text to include other data forms such as images, audio, and structured data (Jiang, Kelly, and Xiu 2023; Wang et al. 2023). In finance, this is particularly useful for integrating various data sources, such as text from news articles, numerical data from financial statements, and visual data from market charts. For instance, combining textual analysis of news with visual analysis of stock price movements can provide a more comprehensive understanding of market trends and investor sentiment. This integration of different data types enhances the robustness and depth of financial analysis.

Interpretability. While deep learning models are often seen as "black boxes," LLMs' ability to generate human-like outputs opens doors to explainability. This characteristic facilitates the provision of both results and their underlying explanations, thereby enhancing the comprehensibility of the reasoning processes within LLMs and increasing trust and transparency in their financial applications.

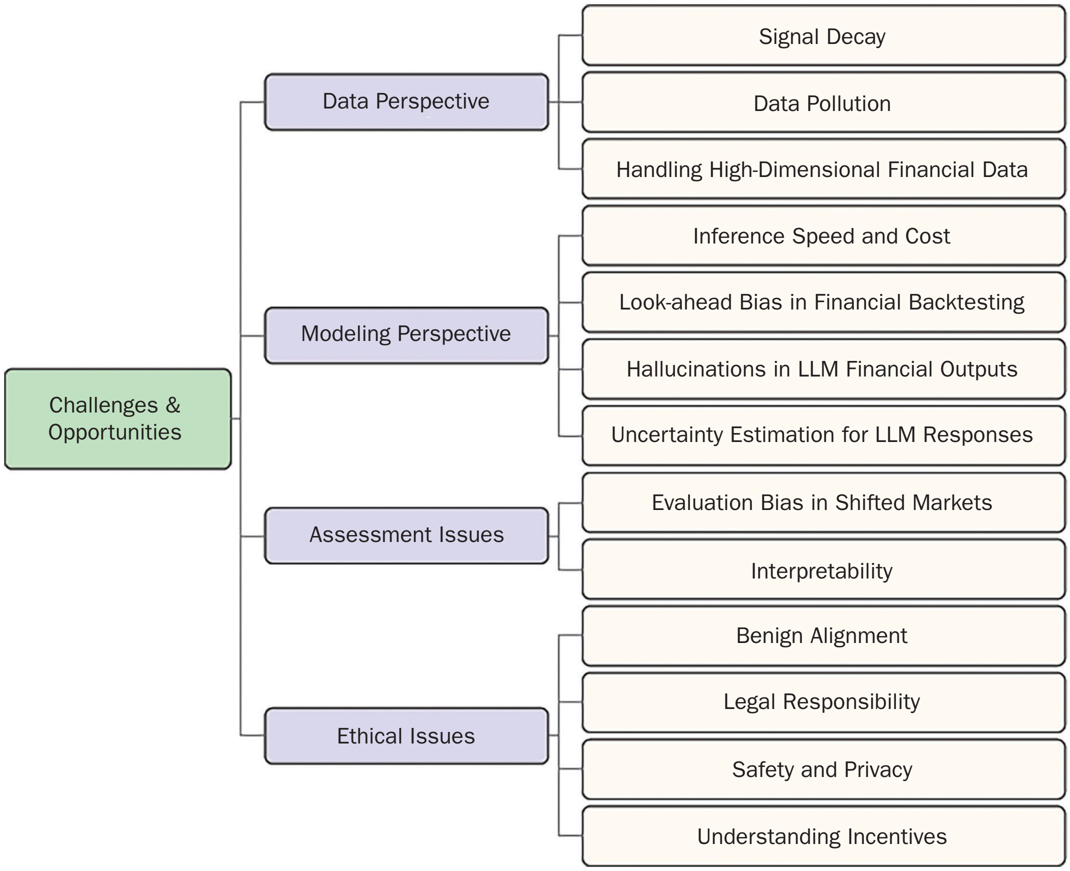
Customization. LLMs exhibit a significant degree of adaptability, enabling customization to accommodate specific financial instruments or market conditions. By integrating domain-specific data and parameters, LLMs can be trained to focus on particular aspects of financial markets, such as risk assessment for bonds or trend prediction in stock markets. This approach enhances the analytical capabilities of LLMs, allowing them to generate insights that are finely tuned to the complexities of different financial environments.

CHALLENGES AND OPPORTUNITIES

The integration of LLMs into finance offers a host of advantages, yet it is imperative to approach this innovative development with a balanced perspective. While the potential of LLMs to transform data-driven decision-making is evident, there are numerous challenges that must be acknowledged and addressed. These challenges, however, should not be viewed merely as obstacles but as avenues for further advancement and refinement. By identifying these hurdles, we also uncover significant opportunities to enhance the efficacy of LLMs in financial contexts. This section seeks to critically examine the key challenges and opportunities (as shown in Exhibit 3) inherent in the deployment of LLMs within the financial sector, highlighting how concerted efforts from both researchers and practitioners can facilitate overcoming these barriers. Such collaborations hold the promise of unlocking new frontiers in financial technology, driving more robust and informed decision-making processes across the industry.

EXHIBIT 3

An Overview of Challenges and Opportunities in LLMs for Financial and Investment Management



Data Perspective

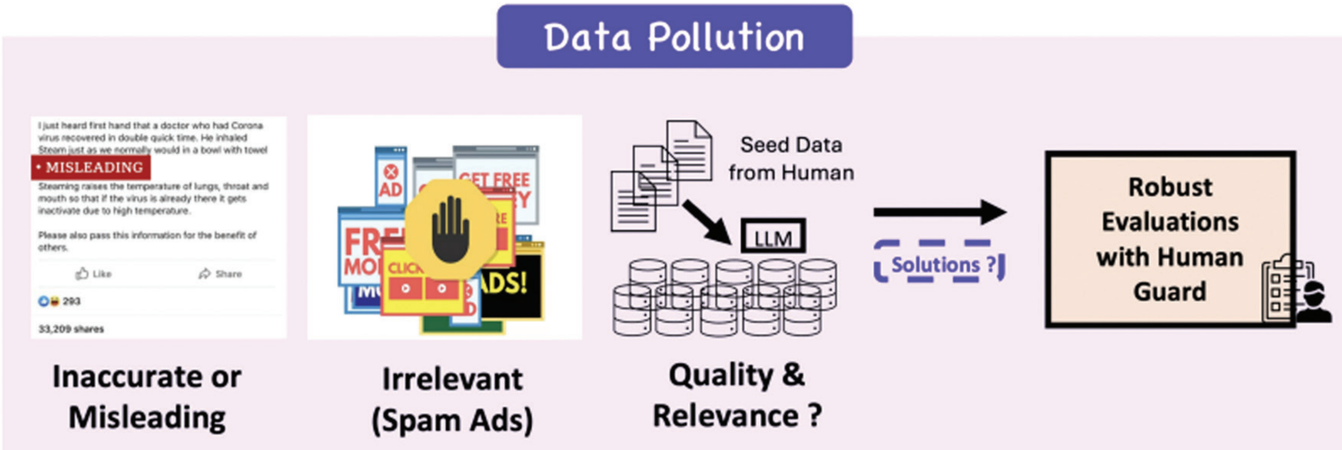
Financial data itself poses several unique challenges, including signal decay, data pollution, and the inherently complex and highly dimensional nature of financial data.

Signal decay. The widespread adoption of LLMs for generating trading strategies in the rapidly evolving financial world poses a unique challenge: signal decay. As more market participants employ LLMs, the effectiveness of these strategies may diminish over time, leading to a depletion of profitable market signals. However, this challenge also presents an opportunity to develop adaptive LLMs that continuously learn from new data and evolve alongside market conditions (Wang, Yuan et al. 2023; Wang, Yuan, Ni, and Guo 2024; Wang, Li et al. 2023; Wang, Xu et al. 2024; Yu, Li et al. 2024). By leveraging their ability to process vast amounts of financial information and identify emerging patterns, these models can maintain their effectiveness over time through continuous retraining and validation.

Data pollution. Data pollution could be a multifaceted challenge that can significantly impact the performance and reliability of these LLM models (as illustrated in Exhibit 4). The first aspect of data pollution involves the presence of inaccurate, misleading, or irrelevant data, such as spam advertisements or deliberate misinformation that has been fed into the LLM model (Zhou et al. 2024). This type of data pollution can severely degrade the performance of LLMs, leading to poor decision-making and compromised integrity of financial models, particularly when using cloud-based LLMs like ChatGPT, as the contamination can spread throughout the entire training environment.

EXHIBIT 4

Challenges and Solutions Regarding Data Pollution



A second and increasingly important aspect of data pollution lies in the growing trend of data being generated by LLMs themselves, rather than by humans (Chen and Shu 2024; Pan et al. 2023; Zellers et al. 2019). This phenomenon raises concerns about the quality and relevance of the data used to train these models. For instance, if financial reports are generated by LLMs, the models are essentially learning from their own output, which can result in increasingly rigid and inflexible learning. The models may fail to capture the true intentions and nuances of human expression, leading to a deterioration in the quality of the generated content.

To address this issue, major companies are strongly emphasizing the collection of high-quality, diverse datasets that include real human interactions. One potential solution to mitigate the impact of LLM-generated data pollution is to develop evaluation methods to assess the meaningfulness of the data created by LLMs (Zhao et al. 2021; Evans et al. 2021; Chang et al. 2024). In this case, we can enhance the performance and reliability of these models, ultimately leading to more accurate and trustworthy financial analyses and predictions.

Handling high-dimensional financial data. While LLMs have demonstrated remarkable proficiency in processing and understanding contextual information within long text sequences, their performance in handling high-dimensional financial time series data remains uncertain (Theodorou, Xiao, and Sun 2023; Li et al. 2021). The unique challenges posed by the complex and highly dimensional nature of financial data present an opportunity for further research and exploration (Fan and Li 2006; Koop and Korobilis 2019). By investigating the potential of hybrid models that combine the contextual understanding of LLMs with specialized techniques for handling high-dimensional data, domain-specific pretraining strategies, and the integration of LLMs with other machine learning techniques, researchers can develop powerful AI models tailored to analyze and understand financial time series (Cao et al. 2024). These advancements could ultimately enhance the performance and applicability of AI in the financial sector, leading to more accurate predictions, better risk management, and improved decision-making processes.

Modeling Perspective

Leveraging LLMs in finance presents unique modeling challenges, such as inference speed, look-ahead bias, hallucinations, and uncertainty estimation, all of which could potentially affect the model's performance and reliability in financial applications.

Inference speed and cost. Balancing the need for fast and cost-effective model inference with performance requirements is a significant challenge due to the high computational demands of LLMs (Bender et al. 2021). This can sometimes lead to high inference costs and slower speeds, particularly when processing large datasets. However, advances in model optimization and hardware offer exciting opportunities to reduce these costs and improve speeds (Zhou, Ning et al. 2024). This makes LLMs more accessible and practical for various financial applications, promoting more efficient resource use and wider adoption of LLM technologies in the finance industry.

For instance, a hybrid inference approach, as discussed in the work by Ding et al. (2024), proposes using a router to dynamically allocate queries to either a small or large model based on the predicted query difficulty and the required quality level. This method aims to balance the trade-off between cost and performance effectively. The router can be fine-tuned to ensure that simpler queries are handled by the smaller, less expensive models, while the more complex queries are directed to the larger, more powerful models. This approach can lead to significant cost savings—up to 40% fewer calls to the large model—without compromising the response quality. Such optimization can make LLMs more economically viable for financial applications, where precision and speed are critical, thereby enhancing their adoption and utility across various financial services and operations.

Look-ahead bias in financial backtesting. Backtesting financial models using LLMs presents a significant challenge due to the risk of future look-ahead bias (Sarkar and Vafa 2024). This bias occurs when the model inadvertently incorporates information from the future during the training process, leading to overly optimistic and misleading backtesting results. Consequently, the model's reliability and predictive accuracy come into question, as it may not perform as well on unseen, real-time data. Addressing this issue requires careful handling of data and the implementation of robust validation techniques to ensure the integrity of the backtesting process.

Despite the challenges posed by future look-ahead bias, researchers can explore innovative solutions to address this issue and design LLMs that effectively mitigate its impact. One of the straightforward methods, as addressed by Kim, Muhn, and Nikolaev (2024), is to use anonymized data that cannot be identified by the LLM. This ensures that the LLM cannot leverage its pretrained memory when dealing with specific firm questions. However, robust validation techniques should still be implemented. The authors perform formal analyses to further rule out concerns about look-ahead bias.

Similarly, recent work in Drinkall et al. (2024) specifically designed a series of point-in-time LLMs called TimeMachineGPT (TiMaGPT). These models are trained on datasets that maintain temporal integrity, ensuring that they remain uninformed about future factual information and linguistic changes. By avoiding the incorporation of future information during training, TiMaGPT models can provide more accurate and reliable insights for time-series forecasting and other dynamic contexts in financial modeling. The availability of both the models and training datasets further enhances the transparency and reproducibility of the results.

Hallucinations in LLM financial outputs. The use of LLM-generated content in various financial tasks raises significant concerns about legality and reliability. Financial reports are subject to strict legal and regulatory standards, and inaccuracies can lead to severe consequences for organizations. A primary issue with LLMs is their potential to generate fake, hallucinated, or factually incorrect statements due to their training on vast amounts of data. Ensuring LLM-generated content adheres to legal standards and is error-free is complex and requires careful consideration and monitoring, especially when the output may not undergo the same scrutiny as a full financial report. The lack of standardized frameworks and guidelines for robot-generated financial content could further complicate this process.

To address the challenge of ensuring accuracy and trustworthiness in LLM-generated financial content, leveraging advanced tools such as GenAudit (Krishna et al. 2024) presents significant opportunities. GenAudit is designed to assist in fact-checking LLM responses for document-grounded tasks. It suggests edits by revising or removing unsupported claims and presents evidence for supported facts. Comprehensive evaluations by human raters demonstrate GenAudit's effectiveness in detecting errors across various LLM outputs from diverse domains. The system is designed to increase error recall while minimizing the impact on precision, ensuring that most errors are flagged and corrected.

Uncertainty estimation for LLM responses. Estimating the uncertainty and providing confidence intervals for model predictions is critical in finance because LLM outputs are essentially sampled from a distribution, rather than being deterministic. This means that asking the LLM the same question multiple times may yield different responses, with some samples potentially having significant errors. For financial decision-making or forecasting, relying on a single sample can be misleading. Moreover, when applying these predictions practically, the error range remains unknown, making risk control challenging. Therefore, to manage risk, it is necessary to perform uncertainty estimation on LLM outputs and establish confidence intervals for their predictions (Ferdaus et al. 2024). This approach helps control errors and mitigate risks. Developing sophisticated methods for uncertainty quantification can provide more reliable confidence intervals, thereby enhancing risk management and decision-making in finance. It allows stakeholders to make more informed and confident decisions based on LLM predictions.

Assessment Issues

Evaluation and interpretability pose several challenges in the application of LLMs to finance, as traditional benchmarks often fail to capture the dynamics of markets influenced by the widespread adoption of these LLM models. Furthermore, enhanced transparency is crucial for fostering trust in AI-driven decisions, ensuring that stakeholders can clearly understand the underlying model outputs.

Evaluation bias in shifted markets. In addition to the aforementioned signal decay caused by the widespread adoption of LLM models, another significant challenge in constructing trading strategies using LLMs lies in the evaluation process. This difficulty stems from the fact that current benchmarks and back-testing periods predate the emergence and popularization of LLMs, especially since ChatGPT's surge in popularity. Consequently, these evaluation methods may not accurately reflect the performance of LLM-generated signals in contemporary market conditions. For instance, the benchmarks that were once suitable for evaluating trading signals in a pre-LLM environment may no longer be applicable, as the widespread availability of LLMs has altered the landscape. This change in the environment is not a gradual decay but rather a fundamental shift that requires a new approach to evaluation.

To address this issue, it is crucial to develop new benchmarks that are adaptable to LLMs and aligned with the current state of the market. Without such benchmarks, it becomes difficult to accurately assess the performance of LLM-generated signals, leading to uncertainty regarding their effectiveness. Therefore, in addition to the traditional signal decay problem, the evaluation difficulty posed by the changed environment should also be recognized and addressed to effectively leverage LLMs in constructing trading strategies.

Interpretability. The lack of interpretability in LLMs used within the finance industry presents a significant challenge. Stakeholders require a clear understanding of how these models arrive at their decisions to establish trust and effectively utilize their outputs.

Developing methods to enhance the transparency and interpretability of LLMs is an ongoing effort (Yu et al. 2023; Lopez-Lira and Tang 2023). By investing in research aimed at improving the interpretability of LLMs, financial institutions can build trust and transparency in their AI-driven processes, leading to better decision-making and increased adoption of LLMs in the financial sector. As described in PloutosGPT (Tong et al. 2024), two quantifiable metrics—faithfulness and informativeness—are employed to verify the quality of the interpretability of the generated rationales. Faithfulness measures whether the facts in the model's response are based on or can be inferred from the given knowledge, while informativeness measures the amount of information contained in the model's response. The development of tools that explain model decisions can help stakeholders comprehend and effectively use the insights generated by AI.

Ethical Issues

Benign alignment, legal responsibility, data privacy, and the transparency of incentives are critical ethical challenges in applying LLMs within the financial sector. Addressing these concerns is essential to ensuring that LLMs are used responsibly, fostering trust, compliance, and the ethical use of AI technologies in finance.

Benign alignment. Ensuring that LLMs output content that aligns with social values and avoids harmful recommendations is a key concern (Yao et al. 2024). This involves making sure that the outputs conform not only to ethical standards but also to legal regulations, avoiding suggestions that could lead to negative actions. This issue intersects with both attack prevention and safety measures. The challenge lies in aligning the objectives of LLMs with benign and ethical goals, as misaligned models can produce unintended and potentially harmful consequences. Therefore, it is crucial to ensure that LLMs operate within ethical boundaries and adhere to regulatory standards. The opportunity here is to proactively align LLM objectives with ethical standards to mitigate risks and ensure that these models contribute positively, particularly in the financial sector. This includes developing frameworks for ethical AI (Cao et al. 2023; He, Xia, and Henderson 2024) use in finance, which can foster trust and compliance.

Legal responsibility. As LLMs continue to play an increasingly significant role in financial decision-making, the issues of legal responsibility and accountability become more prominent (Weidinger et al. 2021). The complexity of these models and their potential for misuse present unique challenges in determining accountability when things go wrong. It is crucial to establish clear frameworks and regulations to address these concerns. The development of a well-defined legal framework for the use of LLMs in the financial sector is essential to provide certainty and foster confidence among stakeholders. By clarifying the lines of responsibility and accountability, such a framework can promote the widespread adoption of these technologies while ensuring their responsible use (Ferdaus et al. 2024; Cheong et al. 2024). This framework should allocate liability in cases where LLMs are misused or produce unintended consequences; establish standards for the development, testing, and deployment of LLMs in financial applications; and provide mechanisms for red flags and compensation in cases where LLMs cause financial harm.

Safety and privacy. The security and privacy of financial data are extremely important, given the significant threats posed by data breaches and compliance violations. Deploying LLMs in the financial sector presents unique challenges in maintaining robust data protection measures and safeguarding sensitive information (Yuan et al. 2024; Carlini et al. 2021). However, advancements in cybersecurity can bolster the security and privacy of financial data used by LLMs (Yao et al. 2024). By implementing strong security protocols, we can mitigate the risks of data leaks and ensure adherence to privacy regulations, thereby building trust and protecting sensitive information. To further prevent data leakage, especially with cloud-based GPT models, it is essential to process confidential data in a local environment. This approach minimizes the risk of breaches while still leveraging the capabilities of LLMs. With the growing availability of open-source models, organizations can now utilize LLMs within their local infrastructure, ensuring the security and privacy of their financial data while benefiting from the advanced features these models offer.

Understanding incentives. The highly competitive nature of the financial industry, coupled with the massive amounts of capital outcomes, necessitates a careful examination of the incentives driving the development and application of LLMs. As LLMs become increasingly prevalent in various domains, including finance, it is crucial to consider their potential impact on individuals and organizations, including government agencies.

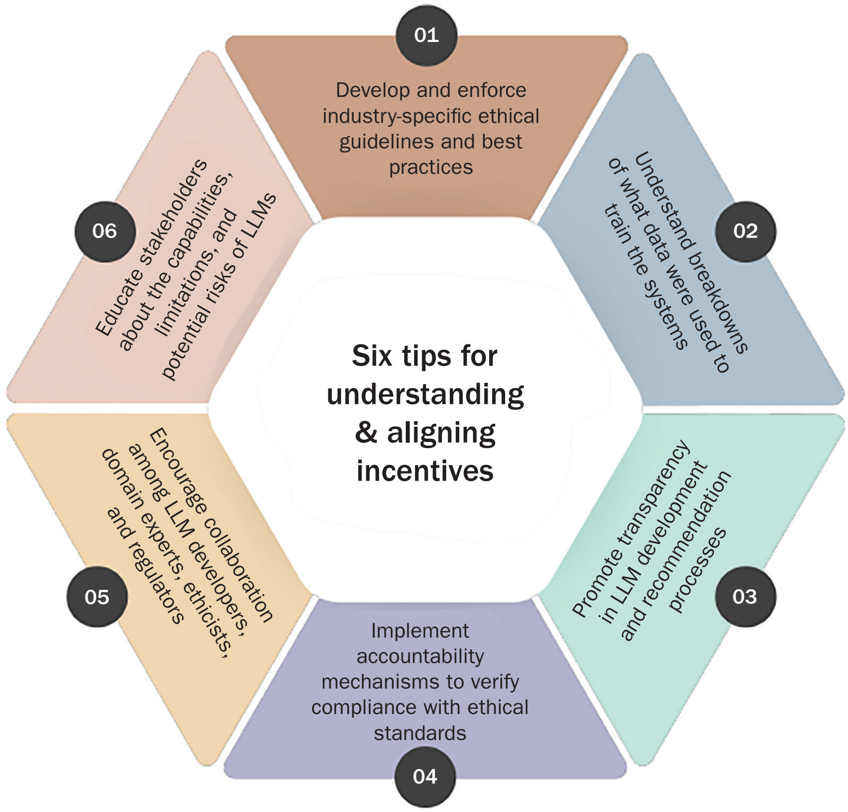
Ethical concerns surrounding AI are growing. Professional organizations such as the Association for Computing Machinery (ACM 2024) have developed codes of ethics and conduct to guide the development and use of AI technologies. However, unlike regulated professions such as medicine, law, or engineering, where practitioners are bound by professional designations and can face consequences for violating ethical standards, LLM developers are not subject to similar oversight. This lack of formal accountability mechanisms poses challenges in ensuring that LLM developers adhere to established ethical guidelines. Furthermore, LLMs themselves, when performing reasoning and decision-making, are likely to operate in an opaque manner, creating barriers to uncover and understand all their potential incentives, particularly those that may lead to negative ethical implications.

To address this, there is a pressing need for greater transparency regarding the incentives behind LLM recommendations. For instance, the fund industry has been moving toward clear reporting of management fees for fund managers. A similar approach should be adopted for LLMs to systematically evaluate their impact on stakeholders. Europe has taken proactive steps with the AI Act (Commission 2024), adopting a “risk-based approach” to regulate high-risk applications and mitigate potential harms, such as racial biases. This framework highlights the challenge of balancing effective regulation with fostering innovation.

As LLMs continue to evolve and integrate into the financial industry, understanding and aligning incentives will be critical to ensuring their responsible and beneficial application. This may involve a combination of approaches (as provided in Exhibit 5), including developing and enforcing industry-specific ethical guidelines and best practices; understanding breakdowns of what data were used to train the systems; promoting transparency in LLM development and recommendation processes; implementing accountability mechanisms to verify compliance with ethical standards; encouraging collaboration among LLM developers, domain experts, ethicists, and regulators; and educating stakeholders about the capabilities, limitations, and potential risks of LLMs.

In the end, natural language takes place in various situations: to inform, to persuade, to entertain, to educate, and so on. Thus, we would expect LLMs to be employed within these constructs. While humans have exquisite talents with situational awareness, it will be interesting to see if LLMs will be able to develop their own skills in this regard. As the financial industry increasingly adopts LLMs, a proactive and collaborative approach to addressing ethical concerns, aligning incentives, and ensuring responsible application will be essential to harnessing the benefits of this transformative technology while mitigating potential harms.

EXHIBIT 5
Tips for Understanding Incentives



CONCLUSION AND OUTLOOK

The integration of LLMs into the financial and investment management sector marks a significant leap forward in the evolution of financial technology and analysis. This article has explored the various models, challenges, and opportunities that arise from applying LLMs to financial applications.

We have examined a range of financially specialized LLMs, from GPT-series derivatives such as Ploutos to BERT variants (e.g., FinBERT), and recent specialized developments such as BloombergGPT and FinGPT. These models demonstrate the rapid progress and increasing sophistication in adapting general-purpose LLMs to the specific needs of the financial domain. The advantages of applying LLMs in finance are manifold, including advanced contextual understanding, transfer learning flexibility, scalability for real-time analysis, multimodality, improved interpretability, and customization capabilities. These features position LLMs as powerful tools for enhancing decision-making processes, risk assessment, and predictive modeling in finance.

However, the adoption of LLMs in finance is not without its challenges. From a data perspective, issues such as signal decay, data pollution, and the complexity of handling high-dimensional financial data present significant hurdles. On the modeling front, concerns about inference speed and cost, look-ahead bias in backtesting, and the potential for hallucinations in LLM outputs require careful consideration and innovative solutions. Furthermore, the ethical implications of deploying LLMs in finance cannot be overstated. Challenges related to benign alignment, legal responsibility, safety, and privacy, and understanding the underlying incentives of these models, demand ongoing attention and proactive measures from researchers, practitioners, and policymakers alike.

Despite these challenges, the opportunities for advancement are substantial. The development of adaptive LLMs, improved data quality assessment methods, novel benchmarking approaches, and enhanced interpretability techniques all represent promising avenues for future research and development. Looking ahead, the application of LLMs in finance will continue to evolve rapidly. As technology advances and our understanding of related issues deepens, LLMs are poised to become a key driver of innovation and efficiency in the financial industry. However, realizing the full potential of LLMs will require close collaboration among academia, industry, and regulatory bodies to drive technological progress while ensuring ethical and regulatory compliance. The successful integration of LLMs in finance will depend on our ability to address these challenges while capitalizing on the immense potential these models offer. This will require continued collaboration among AI researchers, financial experts, ethicists, and regulators to ensure that LLMs are deployed responsibly and effectively in investment management. By navigating these complexities thoughtfully, we can harness the power of LLMs to drive innovation, improve decision-making, and ultimately contribute to a more efficient and robust financial ecosystem.

AUTHOR NOTE

Y. Kong and Y. Nie contributed equally to this work and are listed in random order. The remaining authors are listed in alphabetical order by surname. J. M. Mulvey (corresponding author) and S. Zohren are main advisors.

ACKNOWLEDGMENTS

Y. N. and H. V. Poor acknowledge financial support from Princeton Language and Intelligence at Princeton University. X.D. acknowledges support from the Oxford-Man Institute of Quantitative Finance.

REFERENCES

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. 2023. "GPT-4 Technical Report." arXiv preprint arXiv:2303.08774.

Araci, D. 2019. "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models." arXiv preprint arXiv:1908.10063.

Association for Computing Machinery (ACM). 2024. "ACM Code of Ethics and Professional Conduct." https://www.acm.org/code-of-ethics.

Bai, J., S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. 2023. "Qwen Technical Report." arXiv preprint arXiv:2309.16609.

Baichuan-inc. 2023. "Baichuan-13B." https://github.com/baichuan-inc/Baichuan-13B.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

Bhatia, G., E. M. B. Nagoudi, H. Cavusoglu, and M. Abdul-Mageed. 2024. "FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models." arXiv preprint arXiv:2402.10986.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901.

Cao, B., Y. Cao, L. Lin, and J. Chen. 2023. "Defending against Alignment-Breaking Attacks via Robustly Aligned LLM." arXiv preprint arXiv:2309.14348.

Cao, Y., Z. Chen, Q. Pei, F. Dimino, L. Ausiello, P. Kumar, K. Subbalakshmi, and P. M. Ndiaye. 2024. "RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data." arXiv preprint arXiv:2404.07452.

Carlini, N., F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. 2021. "Extracting Training Data from Large Language Models." In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650.

Chang, Y., X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. 2024. "A Survey on Evaluation of Large Language Models." *ACM Transactions on Intelligent Systems and Technology* 15 (3): 1–45.

Chen, C., and K. Shu. 2024. "Can LLM-Generated Misinformation Be Detected?" In *The Twelfth International Conference on Learning Representations*.

Chen, W., Q. Wang, Z. Long, X. Zhang, Z. Lu, B. Li, S. Wang, J. Xu, X. Bai, X. Huang, et al. 2023. "DISC-FinLLM: A Chinese Financial Large Language Model Based on Multiple Experts Fine-Tuning." arXiv preprint arXiv:2310.15205.

Cheong, I., K. Xia, K. K. Feng, Q. Z. Chen, and A. X. Zhang. 2024. "(A) I Am Not a Lawyer, but...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice." In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2454–2469. https://dl.acm.org/doi/10.1145/3630106.3659048.

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning. 2020. "ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators." arXiv preprint arXiv:2003.10555.

Commission, E. 2024. "Regulatory Framework for AI." https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Deng, X., V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky. 2023. "What Do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis." In *Companion Proceedings of the ACM Web Conference 2023*, pp. 107–110.

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer. 2024. "QLoRA: Efficient Finetuning of Quantized LLMs." *Advances in Neural Information Processing Systems* 36.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805.

Ding, D., A. Mallick, C. Wang, R. Sim, S. Mukherjee, V. Ruhle, L. V. Lakshmanan, and A. H. Awadallah. 2024. "Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing." arXiv preprint arXiv:2404.14618.

Drinkall, F., E. Rahimikia, J. B. Pierrehumbert, and S. Zohren. 2024. "Time Machine GPT." arXiv preprint arXiv:2404.18543.

Evans, O., O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, and W. Saunders. 2021. "Truthful AI: Developing and Governing AI That Does Not Lie." arXiv preprint arXiv:2110.06674.

Fan, J., and R. Li. 2006. "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery." In *25th International Congress of Mathematicians, ICM 2006*.

Ferdaus, M. M., M. Abdelguerfi, E. Ioup, K. N. Niles, K. Pathak, and S. Sloan. 2024. "Towards Trust-Worthy AI: A Review of Ethical and Robust Large Language Models." arXiv preprint arXiv:2407.13934.

He, L., M. Xia, and P. Henderson. 2024. "What's in Your 'Safe' Data? Identifying Benign Data That Breaks Safety." arXiv preprint arXiv:2404.01099.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2021. "LoRA: Low-Rank Adaptation of Large Language Models." arXiv preprint arXiv:2106.09685.

InternLM. 2024. "InternLM." https://github.com/InternLM.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. 2023. "Mistral 7B." arXiv preprint arXiv:2310.06825.

Jiang, J., B. Kelly, and D. Xiu. 2023. "(Re-)Imag(in)ing Price Trends." *The Journal of Finance* 78 (6): 3193–3249.

Kim, A., M. Muhn, and V. V. Nikolaev. 2024. "Financial Statement Analysis with Large Language Models." Chicago Booth Research Paper No. 2024-65; Fama-Miller Working paper.

Konstantinidis, T., G. Iacovides, M. Xu, T. G. Constantinides, and D. Mandic. 2024. "Fin-Llama: Financial Sentiment Classification for Algorithmic Trading Applications." arXiv preprint arXiv:2403.12285.

Koop, G., and D. Korobilis. 2019. "Forecasting with High-Dimensional Panel VARs." *Oxford Bulletin of Economics and Statistics* 81 (5): 937–959.

Krishna, K., S. Ramprasad, P. Gupta, B. C. Wallace, Z. C. Lipton, and J. P. Bigham. 2024. "GenAudit: Fixing Factual Errors in Language Model Outputs with Evidence." arXiv preprint arXiv:2402.12566.

Le Scao, T., A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. 2023. "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model."

Li, J., Y. Bian, G. Wang, Y. Lei, D. Cheng, Z. Ding, and C. Jiang. 2023. "CFGPT: Chinese Financial Assistant with Large Language Model." arXiv preprint arXiv:2309.10654.

Li, X., F. Tramer, P. Liang, and T. Hashimoto. 2021. "Large Language Models Can Be Strong Differentially Private Learners." arXiv preprint arXiv:2110.05679.

Li, Y., S. Wang, H. Ding, and H. Chen. 2023. "Large Language Models in Finance: A Survey." In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 374–382.

Liu, F., T. Zhu, X. Wu, B. Yang, C. You, C. Wang, L. Lu, Z. Liu, Y. Zheng, X. Sun, et al. 2023. "A Medical Multimodal Large Language Model for Future Pandemics." *NPJ Digital Medicine* 6 (1): 226.

Liu, H., C. Li, Q. Wu, and Y. J. Lee. 2024. "Visual Instruction Tuning." *Advances in Neural Information Processing Systems* 36.

Liu, X.-Y., J. Zhang, G. Wang, W. Tong, and A. Walid. 2024. "FinGPT-HPC: Efficient Pretraining and Finetuning Large Language Models for Financial Applications with High-Performance Computing." arXiv preprint arXiv:2402.13533.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint arXiv:1907.11692.

Liu, Z., D. Huang, K. Huang, Z. Li, and J. Zhao. 2021. "FinBERT: A Pre-Trained Financial Language Representation Model for Financial Text Mining." In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 4513–4519.

Lopez-Lira, A., and Y. Tang. 2023. "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models." arXiv preprint arXiv:2304.07619.

Lu, D., H. Wu, J. Liang, Y. Xu, Q. He, Y. Geng, M. Han, Y. Xin, and Y. Xiao. 2023. "BBT-Fin: Comprehensive Construction of Chinese Financial Domain Pre-Trained Language Model, Corpus and Benchmark." arXiv preprint arXiv:2302.09432.

Ma, S., H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei. 2024. "The Era of 1-Bit LLMs: All Large Language Models Are in 1.58 Bits." arXiv preprint arXiv:2402.17764.

Meta AI. 2024. "Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date." https://ai.meta.com/blog/meta-llama-3/.

Nie, Y., and H. Yuan. 2021. "Neural Network Is Heterogeneous: Phase Matters More." arXiv preprint arXiv:2111.02014.

Pan, Y., L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Wang. 2023. "On the Risk of Misinformation Pollution with Large Language Models." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1389–1403.

Pavlyshenko, B. M. 2023. "Financial News Analytics Using Fine-Tuned Llama 2 GPT Model." arXiv preprint arXiv:2308.13032.

Radford, A., K. Narasimhan, T. Salimans, I. Sutskever. 2018. "Improving Language Understanding by Generative Pre-Training."

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners." *OpenAI blog* 1 (8): 9.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* 21 (140): 1–67.

Rodriguez Inserte, P., M. Nakhlé, R. Qader, G. Caillaut, and J. Liu. 2024. "Large Language Model Adaptation for Financial Sentiment Analysis." arXiv preprint arXiv:2401.14777.

Sarkar, S. K., and K. Vafa. 2024. "Lookahead Bias in Pretrained Language Models." *Available at SSRN*.

Shah, R. S., K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, and D. Yang. 2022. "When FLUE Meets FLANG: Benchmarks and Large Pre-Trained Language Model for Financial Domain." arXiv preprint arXiv:2211.00083.

Steinert, R., and S. Altmann. 2023. "Linking Microblogging Sentiments to Stock Price Movement: An Application of GPT-4." arXiv preprint arXiv:2308.16771.

Theodorou, B., C. Xiao, and J. Sun. 2023. "Synthesize High-Dimensional Longitudinal Electronic Health Records via Hierarchical Autoregressive Language Model." *Nature Communications* 14 (1): 5305.

Tong, H., J. Li, N. Wu, M. Gong, D. Zhang, and Q. Zhang. 2024. "Ploutos: Towards Interpretable Stock Movement Prediction with Financial Large Language Model." arXiv preprint arXiv:2403.00782.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. "LLAMA: Open and Efficient Foundation Language Models." arXiv preprint arXiv:2302.13971.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. 2023. "LLAMA 2: Open Foundation and Fine-Tuned Chat Models." arXiv preprint arXiv:2307.09288.

Wang, S., T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu, and Q. Wen. 2024. "Large Language Models for Education: A Survey and Outlook." arXiv preprint arXiv:2403.18105.

Wang, S., H. Yuan, L. M. Ni, and J. Guo. 2024. "QuantAgent: Seeking Holy Grail in Trading by Self-Improving Large Language Model." arXiv preprint arXiv:2402.03755.

Wang, S., H. Yuan, L. Zhou, L. M. Ni, H.-Y. Shum, and J. Guo. 2023. "Alpha-GPT: Human-AI Interactive Alpha Mining for Quantitative Investment." arXiv preprint arXiv:2308.00016.

Wang, Z., Y. Li, J. Wu, J. Soon, and X. Zhang. 2023. "FinVis-GPT: A Multimodal Large Language Model for Financial Chart Analysis." arXiv preprint arXiv:2308.01430.

Wang, Z., P. Sun, and A. Boukerche. 2022. "A Novel Time Efficient Machine Learning-Based Traffic Flow Prediction Method for Large Scale Road Network." In *ICC 2022-IEEE International Conference on Communications*, pp. 3532–3537. IEEE.

Wang, Z., Y. Sun, Z. Lei, X. Zhu, and P. Sun. 2023. "SST: A Simplified Swin Transformer-Based Model for Taxi Destination Prediction Based on Existing Trajectory." In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1404–1409. IEEE.

Wei, J., M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. 2021. "Finetuned Language Models Are Zero-Shot Learners." arXiv preprint arXiv:2109.01652.

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. 2022. "Emergent Abilities of Large Language Models." arXiv preprint arXiv:2206.07682.

Weidinger, L., J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al. 2021. "Ethical and Social Risks of Harm from Language Models." arXiv preprint arXiv:2112.04359.

William, T., and P. B. Ramtin Babaei. 2023. "Fin-LLAMA: Efficient Finetuning of Quantized LLMs for Finance." https://github.com/Bavest/fin-llama.

Wu, S., O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann. 2023. "BloombergGPT: A Large Language Model for Finance." arXiv preprint arXiv:2303.17564.

Xie, Q., W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang. 2023. "PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance." arXiv preprint arXiv:2306.05443.

Xie, Q., D. Li, M. Xiao, Z. Jiang, R. Xiang, X. Zhang, Z. Chen, Y. He, W. Han, Y. Yang, et al. 2024. "Open-FinLLMs: Open Multimodal Large Language Models for Financial Applications." arXiv preprint arXiv:2408.11878.

Yakar, O., Y. Nie, K. Wada, A. Agarwal, and İ. Ercan. 2019. "Energy Efficiency of Microring Resonator (MRR)-Based Binary Decision Diagram (BDD) Circuits." In *2019 IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1–8. IEEE.

Yang, H., X.-Y. Liu, and C. D. Wang. 2023. "FinGPT: Open-Source Financial Large Language Models." arXiv preprint arXiv:2306.06031.

Yang, Y., Y. Tang, and K. Y. Tam. 2023. "InvestLM: A Large Language Model for Investment Using Financial Domain Instruction Tuning." arXiv preprint arXiv:2309.13064.

Yang, Y., M. C. S. Uy, and A. Huang. 2020. "FinBERT: A Pretrained Language Model for Financial Communications." arXiv preprint arXiv:2006.08097.

Yao, Y., J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang. 2024. "A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly." *High-Confidence Computing*, 100211.

Yu, J., M. Huber, and K. Tang. 2024. "GreedLlama: Performance of Financial Value-Aligned Large Language Models in Moral Reasoning." arXiv preprint arXiv:2404.02934.

Yu, X., Z. Chen, Y. Ling, S. Dong, Z. Liu, and Y. Lu. 2023. "Temporal Data Meets LLM—Explainable Financial Time Series Forecasting." arXiv preprint arXiv:2306.11025.

Yu, Y. 2023. "Cornucopia-LLaMA-Fin-Chinese." https://github.com/jerry1993-tech/Cornucopia-LLaMA-Fin-Chinese.

Yu, Y., H. Li, Z. Chen, Y. Jiang, Y. Li, D. Zhang, R. Liu, J. W. Suchow, and K. Khashanah. 2024. "FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design." In *Proceedings of the AAAI Symposium Series*, Vol. 3, pp. 595–597.

Yuan, T., Z. He, L. Dong, Y. Wang, R. Zhao, T. Xia, L. Xu, B. Zhou, F. Li, Z. Zhang, et al. 2024. "R-Judge: Benchmarking Safety Risk Awareness for LLM Agents." arXiv preprint arXiv:2401.10019.

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. 2019. "Defending against Neural Fake News." *Advances in Neural Information Processing Systems* 32.

Zhang, B., H. Yang, and X.-Y. Liu. 2023. "Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models." arXiv preprint arXiv:2306.12659.

Zhang, B., H. Yang, T. Zhou, M. Ali Babar, and X.-Y. Liu. 2023. "Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models." In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 349–356.

Zhang, X., and Q. Yang. 2023. "XuanYuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters." In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4435–4439.

Zhang, Z., H. Zhang, K. Chen, Y. Guo, J. Hua, Y. Wang, and M. Zhou. 2021. "Mengzi: Towards Lightweight yet Ingenious Pre-Trained Models for Chinese." arXiv preprint arXiv:2110.06696.

Zhao, Z., E. Wallace, S. Feng, D. Klein, and S. Singh. 2021. "Calibrate before Use: Improving Few-Shot Performance of Language Models." In *International Conference on Machine Learning*, pp. 12697–12706. PMLR.

Zhou, B., D. Geißler, and P. Lukowicz. 2024. "Misinforming LLMs: Vulnerabilities, Challenges and Opportunities." arXiv preprint arXiv:2408.01168.

Zhou, Y., Z. Li, S. Tian, Y. Ni, S. Liu, G. Ye, and H. Chai. 2024. "SilverSight: A Multi-Task Chinese Financial Large Language Model Based on Adaptive Semantic Space Learning." arXiv preprint arXiv:2404.04949.

Zhou, Z., X. Ning, K. Hong, T. Fu, J. Xu, S. Li, Y. Lou, L. Wang, Z. Yuan, X. Li, et al. 2024. "A Survey on Efficient Inference for Large Language Models." arXiv preprint arXiv:2404.14294.