Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination

Haoqiang Kang[1,*] **Xiao-Yang Liu**[1,2]
[1]Columbia University,
[2]Rensselaer Polytechnic Institute
haoqik88@gmail.com, XL2427@columbia.edu

Abstract

The hallucination issue is recognized as a fundamental deficiency of large language models (LLMs), especially when applied to fields such as finance, education, and law. Despite the growing concerns, there has been a lack of empirical investigation. In this paper, we provide an empirical examination of LLMs' hallucination behaviors in financial tasks. First, we empirically investigate LLM model's ability of explaining financial concepts and terminologies. Second, we assess LLM models' capacity of querying historical stock prices. Third, to alleviate the hallucination issue, we evaluate the efficacy of four practical methods, including few-shot learning, Decoding by Contrasting Layers (DoLa), the Retrieval Augmentation Generation (RAG) method and the prompt-based tool learning method for a function to generate a query command. Finally, our major finding is that off-the-shelf LLMs experience serious hallucination behaviors in financial tasks. Therefore, there is an urgent need to call for research efforts in mitigating LLMs' hallucination.[2]

1 Introduction

Motivation. Large language models (LLMs) [2, 16, 20] have emerged as transformative tools, demonstrating unprecedented prowess in comprehending and generating human-like text across diverse applications. LLMs are revolutionizing the interface[3] between humans and machines in fields such as finance, education, and law. Among many LLM applications, the finance domain stands out as a notable area of impact. Various specialized models, such as FinBERT [27], BloombergGPT [24], and FinGPT [14, 26], have been customized to understand and generate financial texts, showing considerable promise in aiding humans in a variety of financial tasks, from portfolio management [10] to predictive analysis of market trends [28] and sentiment analysis [28, 29]. However, a fundamental deficiency inherent in these models is *hallucination — generating plausible but unsupported or factually incorrect content to a reference text* — which may be highly risky when deploying financial large language models (FinLLMs). Considering the sensitive and intricate characteristics of finance use cases, inaccuracies and misinformation may lead to severe consequences, such as substantial monetary losses and erosion of trust.

Challenges. The hallucination issue is considered as a major challenge of deploying FinLLMs in real-world applications. Addressing the hallucination issue is nontrivial. First, the problem of properly measuring hallucinations, *determining how often and to what extent LLMs produce incorrect or "hallucinated" information*, particularly amidst intricate financial concepts. Second, the challenge of deployment in real-world financial scenarios. Finance is a multifaceted field, while tasks such as

[*]Haoqiang Kang completed this work as a research assistant at Columbia University.
[2]We release our code and data at https://github.com/mk322/fin_hallu.
[3]There may be a switch from graphical user interface (GUI) to language user interface (LUI).

I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models, at 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

querying historical stock prices demand pinpoint accuracy. *Will LLMs be able to consistently deliver on these demands*?

Contributions. In this paper, taking an empirical approach, we examine hallucination behaviors of LLMs in financial tasks. We summarize our work as follows:

- **Examining LLMs' ability of financial knowledge**: We conduct an empirical investigation to elucidate the extent of hallucinations in finance, produced by LLMs. Our analysis delves into the models' capacity of memorizing and retaining financial domain knowledge, providing critical insights into their reliability in grasping intricate financial concepts and terminologies.
- **A case study of analyzing a real-world financial task**: As a case study, we further analyze the models' performance in a fundamental financial task, namely the ability of querying historical stock prices accurately. This enables us to discern the models' practical utility and effectiveness in addressing tasks that are quintessential in the finance sector.
- **Evaluating mitigation methods for hallucinations**: We assess four practical methods, few-shot prompting, Decoding by Contrasting Layers (DoLa) [3], the *Retrieval Augmentation Generation (RAG)* method and the *prompt-based tool learning method* that generates correct function calls. These strategies are designed to enhance the factual accuracy of the generated outputs and enable models to interact with up-to-date financial data, ensuring the relevance and reliability of the information provided.

These contributions underscore the need to reduce hallucinations of LLMs and enhance their reliability for practical financial tasks in the real world.

2 Background and Related Works

The concept of "hallucination" in the context of LLM refers to instances where the generated text conflicts with either the input instructions (*instruction inconsistency*), the input context (*input context inconsistency*), the previously generated text (*generated context conflict*), or established world knowledge (*factual inconsistency*) [6, 30]. In this work, we focus particularly on *factual inconsistency*, as it represents a serious and frequent type of error. For instance, as illustrated in Figure 1, the GPT4 model incorrectly interprets the financial acronym "TIF" as "Time in Force", which is not the common full name for "TIF", instead of the correct "Tax Increment Financing."

Hallucinations in general-purpose LLMs. Previous research has investigated several factors contributing to LLM hallucination, such as imperfect learning and decoding methods, and knowledge gaps in the training data [7]. In addition, various methods have been proposed to mitigate hallucination in LLMs. One line of work involves factuality-enhanced decoding techniques [3, 19, 13, 12]. Other works leverage external tools, notably the RAG technique, which enhances factuality by incorporating external knowledge sources in the generation process [11, 8, 1], and the application of precise API calls [17, 18] to obtain up-to-dated knowledge. In this study, we examine the effectiveness of one of the most representative decoding method of DoLa [3] method, a traditional RAG approach [11], and an prompt-based tool learning method in mitigating hallucinations within the finance domain.

Hallucinations in specific domains. The deployment of LLMs in specialized domains faces a significant challenge due to the risk of hallucinations. This is particularly critical in areas like finance, medicine and law where accuracy is paramount. Research in this area is growing, with notable efforts like the Med-HALT benchmark by Umapathi et al. [22] assessing hallucinations in the medical domain. Similarly, a study of ChatGPT [23] in the Chinese medical sector demonstrates GPT4's advancements, yet also underscores the ongoing challenges. In the legal field, the ChatLaw model [4] employs a vector database and keyword retrieval to tackle hallucinations in legal data extraction. Additionally, Elaraby et al. [5] advocate for incorporating domain-specific knowledge, such as NBA data, into supervised-finetuned (SFT) datasets to reduce such inaccuracies.

3 Methodology for Empirical Examination

We introduce an empirical framework tailored for evaluating the hallucination behaviours of LLMs within the financial field, as shown in Fig. 1. The framework presents a method to assess LLM

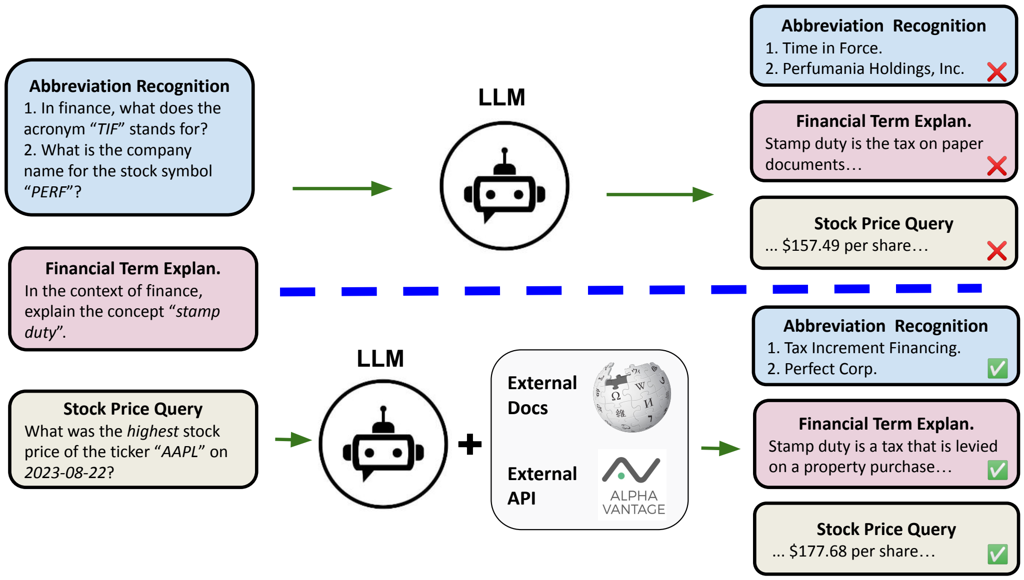

Figure 1: An overview and sample outputs from our empirical analysis of three financial tasks. The sample outputs for the abbreviation recognition task are generated by GPT4, while the others are generated by Llama2-7B-Chat.

responses on three financial tasks: recognizing financial abbreviations, explaining financial terminologies, and fetching stock prices. When provided with standalone questions, the models sometimes hallucinate with wrong facts, but the performance significantly improves when integrated with external data sources, such as financial documents and a function to generate a query command.

3.1 Large Language Models (LLMs)

Throughout our experiments, we use the HuggingFace weights[4][5] of the pretrained Llama2-7B model [21] and its instruction-tuned+RLHF version Llama2-7B-chat [21]. Also, we utilize the OpenAI API to call the models of GPT3.5-turbo[6], and GPT4[7]. We use the greedy decoding method to generate texts for all models. In addition to the general-purpose LMs, we investigate the performance of FinMA-7B-NLP [25], a multi-task fine-tuned LLaMA-1-7B model [20] with instruction data on five finance tasks, including question answering.[8] This evaluation includes a comparison with its base model, LLaMA-1-7B [20], to highlight differences on hallucination after domain-specific finetuning.

Budget. For the inference process with the Llama2-7B and Llama2-7B-chat models, we utilized an A100 GPU. The cumulative GPU time for our entire project amounted to roughly 40 hours. Regarding the generation tasks with the GPT3.5-turbo and GPT4 models, the associated API costs were approximately $200. Moreover, the API costs for employing FactScore with GPT3.5-turbo were approximately $250.

3.2 Financial Tasks

When selecting representative tasks in finance, we considered factors such as their relevance to real-world financial activities and the potential risks of inaccurate outputs. Based on these considerations, we identified three primary tasks, which together provide an assessment of LLM performance. Examples of prompt and outputs for these three tasks are gave in Appendix A.

[4]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[5]https://huggingface.co/meta-llama/Llama-2-7b-hf
[6]https://openai.com/blog/chatgpt
[7]https://openai.com/research/GPT4
[8]To optimize FinMA's ability to follow instructions effectively, we employ the same prompt template that the original authors used during the finetuning process.

3.2.1 Task I: Financial Abbreviation Recognition

Task description. In this task, we measure a LLM model's ability to recognize financial acronyms and stock symbols. We randomly select 192 financial acronyms from Wikipedia [9] and 1215 stock symbols from an online list[10]. This selection is designed to cover a wide range of financial contexts and complexities. The primary objective is to evaluate a LLM mode's capability to accurately expand financial acronyms or provide the full company names corresponding to specific stock symbols. See Table 3 and Table 6 in the Appendix for example outputs and the prompt template used in this task.

Metric. The accuracy for both tasks is measured using a consistent approach. The accuracy score is calculated as the ratio of accurately identified acronyms to the total number, which is 192 or 1000, depending on the task. An example is considered to be *correctly recognized* if the actual full name's string is a substring of the predicted full name.

3.2.2 Task II: Financial Term Explanations

Task description. In this task, a LLM is prompted to give an explanation of a financial terminology. We randomly select 160 infrequently-visited financial terminologies from Wikidata API[11], specifically targeting those with the lowest page views between 2021-01-01 and 2023-01-01. This selection process aim to emphasize more obscure financial concepts that are less commonly encountered in typical discussions. See Table 4 and Table 7 in the Appendix for example outputs and the prompt templates used in this task.

Metric. We employ the FactScore (ChatGPT+Retrieval) metric [15] to measure the factuality of our generated content. This method quantifies the ratio of the number of correct atomic facts to the total number of atomic facts within a given response, benchmarking against the content of each term's explanation (on the Wikipedia full page).

3.2.3 Task III: Stock Price Query

Task description. In this task, we query LLMs for a historical stock price with 560 examples in total. We randomly select 70 stock tickers. For these tickers, we determine a set of four dates that is in the period covered by Llama2's pretraining data (i.e., before September 2022) [21], specifically on 2022-05-23, 2022-06-22, 2022-07-22, and 2022-08-22. See Table 5 and Table 8 in the Appendix for example outputs and the prompt templates used in this task.

Metric. In scenarios where no external tool is utilized, our assessment metrics include 1) accuracy, which is the percentage of the predicted prices that are exactly the same to their corresponding actual prices [12], 2) mean absolute error (MAE), and 3) the percentage of queries where the integer parts of the predicted values are the same as that of the actual price. In the prompt-based tool learning scenario, a response is considered to be correct only if the generated query code of a function call exactly matches one of the predefined set of expected function calls.

3.3 Methods for Mitigating Hallucination

Few-shot prompting. We hand crafted the few-shot prompts for each task, utilizing them to guide the models in their learning process.

Decoding by contrasting layers (DoLa). DoLa [3] is designed to enhance the factual accuracy of LLMs by contrasting the outputs from different layers of the model. This approach assumes that higher layers of the model contain more factual knowledge, which can be leveraged to reduce hallucinations and improve the accuracy of response. In our implementation of DoLa, we utilized the same contrasting layer configurations as those specified by the original authors[13].

Retrieval augmentation generation (RAG). In the tasks of acronym recognition and long-form generation of explanation of financial terms, to further improve the relevance and factuality of the

[9] https://en.wikipedia.org/wiki/List_of_business_and_finance_abbreviations

[10] https://eoddata.com/symbols.aspx. We randomly select symbols that are in the stock exchanges of NASDAQ , NYSE, and AMEX.

[11] https://query.wikidata.org/

[12] Obtained from the Alpha Vantage API: https://www.alphavantage.co/

[13] https://github.com/voidism/DoLa

generated content, we integrate RAG [11] that sources content directly from Wikipedia. By leveraging the vast informational expanse of Wikipedia, we ensure that our generated outputs are grounded in factual and current knowledge. To make this retrieval process efficient and accurate, we employ the FAISS vector store [9], which ensures that the most relevant content is extracted from Wikipedia and seamlessly incorporated into our model's outputs.

Prompt-based tool learning. For the stock price query task, our aim is to equip the model with the capability to use external tools. We employ a prompt-based learning technique. In each query, we assess the model's ability in generating the correct Python function call for a tailored wrapper of the Alpha Vantage API based on given natural language instructions. The model is briefed about the function parameters, encompassing the ticker, date, and price type. With this information, the model must produce the function name and the parameters correctly, adhering to Python's syntax. A response is considered as correct only if it precisely matches a set of the expected query strings.

4 Empirical Results for Hallucination

4.1 Quantifying Hallucinations

Our main results across the three benchmark tasks are given in Table 1 and Table 2, and our main findings are as follows.

General-purpose LLMs generate factually incorrect content in finance. An initial evaluation of relatively smaller, open-source models, specifically Llama2-7B and Llama2-7B-chat, reveals their comparatively limited capacity to assimilate extensive knowledge within the finance domain. Delving into more sophisticated models, an examination of the GPT4 model, as presented in Table 1, demonstrates a 82.5% and 90.4% accuracy when we directly query it in the acronym and stock symbol test. Also, in the long-form generation task that the employed FactScore metric [15] delineates an 81.11% score for GPT4 in a direct query setting. While these performances are notable, they inherently suggest a room for error. Taking a deeper look, upon examining the incorrect responses, we notice that certain generated answers are outdated, stemming from obsolete information. For instance, as illustrated in Figure 1, the GPT4 model incorrectly references "PERF" as the stock symbol for "Perfumania Holdings," not accounting for its delisting, highlighting a lapse in updating its knowledge base. In the sensitive and dynamic finance domains, these deviations and outdatedness can manifest into significant consequences. Consequently, it is essential for researchers and professionals to adopt a discerning approach when utilizing outputs from large language models and emphasize the necessity of autonomous validation for pivotal data they generate.

Multi-task domain-specific finetuning could diminish LMs' general instruction-following abilities. As demonstrated in Table 1, FinMA-7B—a model fine-tuned for specific tasks within the finance domain—underperforms its base model, Llama1-7B, in various tasks under both zero-shot and few-shot settings. This trend indicates that while multi-task domain-specific finetuning aims to bolster a model's domain-specific capabilities, it might also lead to a decrease in its overall ability to accurately follow instructions and adapt to new tasks. Such a decrease could result in more frequent occurrences of instruction-inconsistent hallucinations 2. This observation serves as a cautionary note for the utilization of multi-task finetuning in the development of future domain-specific language models.

General-purpose LLMs generate seriously unreliable real-world financial predictions. The application of LLMs to real-world tasks in the finance domain, particularly in predicting stock prices, raises significant concerns regarding the reliability of their outputs, as highlighted by the results presented in Table 2. In the zero-shot setup, models such as Llama2-7B and Llama2-7B-chat exhibit alarmingly high Mean Absolute Errors (MAE) of 6357.6 USD and 6380.5 USD, respectively. Furthermore, while the utilization of few-shot prompting allows models to generate responses that are closer to correct answers, there still exists a gap in terms of MAE and accuracy. This high deviation from the true values suggests substantial inaccuracies in their predictions.

Interestingly, the models of GPT3.5-turbo and GPT4 opt for a more cautious approach, abstaining from generating responses to some of stock symbol recognition questions and all stock-price-related questions in the absence of external tools. This restraint is praiseworthy as it prevents the propagation of potentially erroneous and misleading financial predictions to users, fostering user trust and minimizing the risk of misinformation in such a critical domain.

Table 1: Results of the acronym recognition task and terminology explanation task. The GPT-3.5 Turbo and GPT-4 models abstain from answering 56.5% and 42.3% of questions, respectively. Their accuracy are evaluated only on the subset of questions to which they provide answers.

Model	Method	Acronym Accuracy (%)	Stock Symbol Accuracy (%)	Term explanation FactScore (%)
Llama1-7B	Zero-Shot	40.5	14.3	46.34
Llama2-7B	Zero-Shot	50.1	16.0	51.94
Llama2-7B-chat	Zero-Shot	74.8	16.5	64.68
Llama2-13B	Zero-Shot	58.3	24.2	56.08
Llama2-13B-chat	Zero-Shot	75.0	12.2	66.72
GPT3.5-turbo	Zero-Shot	76.8	87.7*	78.54
GPT4	Zero-Shot	82.5	90.4*	81.11
FinMA-7B	Zero-Shot	30.4	7.4	25.56
Llama1-7B	Few-Shot	54.3	22.7	49.27
Llama2-7B	Few-Shot	57.3	23.0	51.60
Llama2-7B-chat	Few-Shot	77.1	20.0	67.02
Llama2-13B	Few-Shot	62.7	25.4	63.28
Llama2-13B-chat	Few-Shot	76.1	23.7	67.97
FinMA-7B	Few-Shot	36.5	9.8	27.64
Llama2-7B	DoLa	41.5	17.3	56.39
Llama2-7B-chat	DoLa	61.8	17.6	65.42
Llama2-13B	DoLa	42.8	23.4	67.06
Llama2-13B-chat	DoLa	62.4	13.7	67.50
Llama2-7B	RAG	86.6	61.5	75.07
Llama2-7B-chat	RAG	90.8	**69.3**	90.48
Llama2-13B	RAG	87.4	62.8	78.56
Llama2-13B-chat	RAG	**93.4**	68.5	**90.67**

4.2 Mitigating Hallucination

RAG significantly improves the factuality in finance. Evaluating the impact of the RAG on both foundational and finetuned models provides compelling evidence of its efficacy. As can be seen in Table 1, integrating RAG consistently elevates the performance of both Llama-2 and Llama-2-chat models. Table 1 further underscores the advantage of RAG, showing substantially higher FactScores in the long-form generation task for both models when RAG is implemented. Such consistent improvements across multiple metrics and settings confirm that RAG serves as a significant enhancement to both pretrained and instruction-tuned models.

Prompt-based tool learning helps significantly on the time-sensitive task. As discerned from Table 2, the smaller models, Llama2-7B and Llama2-7B-chat, without the integration of an external tool, exhibit negligible accuracy in handling stock queries. Conversely, the application of the prompt-based tool learning leads to a transformative elevation in performance, with Llama2-7B+tool and Llama2-7B-chat+tool achieving remarkable accuracies of 100.00% with only one training example. This enhancement in reliability and precision is critical, especially within the rapidly evolving landscape of financial domains. It's notable that the sophisticated models, GPT3.5-turbo and GPT4, refrain from addressing stock price queries in a zero-shot setting. However, when augmented with prompt-based tool tool learning, these models are adeptly optimized to provide accurate answers. The integration of zero-shot and few-shot tool learning thus emerges as an effective strategy bridging the knowledge gap and enhancing the reliability of language models, especially in the dynamic domain of finance.

Few-shot learning better improves the ability to follow the question-answering format than factuality. As shown in Table 1, we observe that the pretrained Llama2-7B and Llama2-13B models improve significantly in their question-answering capabilities under few-shot learning, compared to their zero-shot performance. However, this improvement does not markedly surpass the performance of their zero-shot, instruction-tuned counterparts, Llama2-7B-chat and Llama2-13B-chat. Furthermore, for the chat variants, few-shot learning's impact is more limited, yielding only modest

Table 2: The comparative performance of different models for the stock price query task, with and without an external tool. *The OpenAI models abstain to answer any question related to stock prices.

Models	Valid (%) ↑	MAE ($) ↓	int_correct (%) ↑	Accuracy (%) ↑
		Zero-Shot		
Llama2-7B	85.3	6357.6	2.0	0.0
Llama2-7B-chat	100.0	6380.5	2.7	0.0
GPT3.5-turbo*	0.0	-	-	-
GPT4*	0.0	-	-	-
Llama2-7B+DoLA	76.8	6857.9	0.9	0.0
Llama2-7B-chat+DoLA	97.4	6460.2	1.2	0.0
Llama2-7B+tool	47.5	-	-	76.4
Llama2-7B-chat+tool	89.8	-	-	91.2
GPT3.5-turbo+tool	100.0	-	-	98.4
GPT4+tool	100.0	-	-	100.0
		One-Shot		
Llama2-7B	90.2	489.3	2.1	0.0
Llama2-7B-chat	100.0	234.5	3.4	0.0
Llama2-7B+DoLA	86.6	521.6	2.1	0.0
Llama2-7B-chat+DoLA	97.4	234.9	2.4	0.0
Llama2-7B+tool	97.4	-	-	100.0
Llama2-7B-chat+tool	99.8	-	-	100.0
GPT3.5-turbo+tool	100.0	-	-	100.0
GPT4+tool	100.0	-	-	100.0

improvements. This trend suggests that while few-shot learning aids in adapting to question formats, its role in enhancing factual precision is less substantial.

DoLa has limitations in enhancing models with knowledge gaps in training data. The DoLa decoding method is designed to enhance the factual accuracy of LMs by contrasting the outputs from different layers of the model. This approach assumes that higher layers of the model contain more factual knowledge, which can be leveraged to reduce hallucinations and improve the accuracy of responses. However, the effectiveness of DoLa is inherently dependent on the breadth and depth of knowledge encoded in the language model's pretraining dataset. As shown in Table 1 and Table 2, while DoLa can improve factual accuracy in some instances, its effectiveness is limited when the underlying model lacks comprehensive knowledge in its pretraining dataset, such as some stock ticker and stock prices. This limitation is particularly noticeable in scenarios where the model is expected to provide responses based on information that might not be well-represented or current in its training data. Since DoLa relies on amplifying the knowledge already present in the model, its ability to compensate for gaps in the model's foundational knowledge is constrained.

5 Conclusion, Limitations and Future Works

In this study, we conduct an empirical analysis to assess the hallucination problem of LLMs in the financial domain. Our work demystifies the LLM's reliability and ability of explaining financial terminologies and concepts. Furthermore, a performance analysis reveals the practical viability and performance of these models in querying the historical stock prices. To mitigate hallucinations, we show the effectiveness of the RAG method and prompt-based tool learning to generate correct function calls, thereby ensuring the provision of factually correct and up-to-date information in the finance domain.

While our research provides insights into the capabilities of LLMs in the finance domain, it is essential to acknowledge certain limitations. Firstly, our tasks, although representative, cannot encompass the full spectrum of real-world tasks in the vast and varied domain of finance. This implies that our results may not generalize to all possible financial tasks and scenarios. Secondly, the mitigation strategies we introduced and tested are task-specific. Their effectiveness might vary when applied to different tasks or in diverse financial contexts. Future research might need to adapt or expand these strategies to ensure their relevance and effectiveness across a broader range of financial applications.

Moving forward, potential avenues for future works include refining the hallucination mitigation techniques for broader financial applications, exploring ways to further increase the accuracy and reliability of LLMs in dynamic financial settings, and understanding the interplay between LLM outputs and financial decision-making. Our findings highlight the critical issue of hallucinations, establishing a groundwork for advancing responsible and reliable LLM deployment in the financial domain.

References

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[3] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

[4] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*, 2023.

[5] Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*, 2023.

[6] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

[7] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[8] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023.

[9] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[10] Hyungjin Ko and Jaewook Lee. Can chatgpt improve investment decision? from a portfolio management perspective. *From a Portfolio Management Perspective*, 2023.

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[12] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

[13] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022.

[14] Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. FinGPT: Democratizing internet-scale data for financial large language models. *Workshop on Instruction Tuning and Instruction Following at NeurIPS*, 2023.

[15] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

[16] OpenAI. GPT-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[17] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.

[18] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[19] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*, 2023.

[20] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[21] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[22] Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.

[23] Hongyan Wang, WeiZhen Wu, Zhi Dou, Liangliang He, and Liqiang Yang. Performance and exploration of chatgpt in medical examination, records and education in chinese: Pave the way for medical ai. *International Journal of Medical Informatics*, 177:105173, 2023.

[24] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[25] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*, 2023.

[26] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models. *Symposium on FinLLM, IJCAI*, 2023.

[27] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020.

[28] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. *Symposium on FinLLM at IJCAI*, 2023.

[29] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models. *ACM ICAIF*, Nov., 2023.

[30] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

A Output Examples

Examples of Correct Output

Llama 2 7B:
Prompt: In finance, the acronym "MICR" stands for
Generation: Magnetic Ink Character Recognition.
Prompt: In finance, the stock symbol "AAPL" stands for
Generation: Apple Inc.

Llama 2 7B Chat:
Prompt: In finance, what does the acronym "MICR" stand for?
Generation: MICR stands for Magnetic Ink Character Recognition. It is a technology ...
Prompt: In finance, what does the stock ticker "AAPL" stand for?
Generation: It stands for Apple Inc., which is the stock symbol used to represent ...

Examples of Wrong Output

Llama 2 7B:
Prompt: In finance, the acronym "PPI" stands for
Generation: Price To Book Ratio.
Prompt: In finance, the stock symbol "ZVIA" stands for
Generation: "Zions Bancorporation".

Llama 2 7B Chat:
Prompt: In finance, what does the acronym "PPI" stand for?
Generation: Personal Protection Insurance.
Prompt: In finance, what does the stock ticker "ZVIA" stand for?
Generation: "ZVIA" stands for Zvia Food & Beverage, Inc.

Table 3: Examples of zero-shot output for the Llama2 models on the two abbreviation tasks.

Llama 2 7B:
Prompt: In the context of finance, explain the concept of "Security Commission.\n\n Security Commission
Generation: is a regulatory body that monitors the financial market in India. It was established in 1987 to regulate the capital markets and protect investors' interests. The commission has several functions, including:\n\n Regulating the securities market by issuing rules and regulations governing the activities of brokers, traders, and other participants in the market.\n\n Monitoring compliance with these rules and regulations.\n\n* Taking action against violators.\n\n* Promoting transparency and fairness in the market.\n\nThe Security Commission also provides guidance on best practices for investment and risk management. This includes providing educational materials and hosting seminars and workshops to help investors make informed decisions about their investments.

Llama 2 7B Chat:
Prompt: In the context of finance, explain the concept of "Security Commission".
Generation: Security Commission is a term used in the financial industry to refer to an organization that oversees and regulates the securities market. The main function of a Security Commission is to protect investors by ensuring that securities firms operate fairly and honestly, and that they comply with relevant laws and regulations.\nIn many countries, the Security Commission is responsible for issuing licenses to securities firms, approving their financial statements, and monitoring their activities to ensure that they are acting in the best interests of their clients. The Security Commission may also investigate and take action against firms that violate securities laws or engage in unethical practices.\nThe Security Commission plays a critical role in maintaining the integrity of the securities markets and protecting investors from fraudulent activities. By setting and enforcing strict standards for securities firms, the Security Commission helps to ensure that investors have access to a fair and transparent marketplace where they can buy and sell securities with confidence.

Table 4: Examples of zero-shot output for the Llama2 models on the finance terminology task.

Vanilla Zero-Shot

Llama 2 7B:
Prompt: The highest stock price for ticker symbol "CC" on the date of 2022-05-23 was
Generation: 1. nobody knows what will happen in the future, but it is a good idea to be prepared for any scenario.\nThe stock price of ’CC’ has been on a downward trend over the past year and has recently fallen below its ...

Llama 2 7B Chat:
Prompt: What was the highest stock price for ticker symbol "CC" on the date of 2022-05-23?
Generation: The highest stock price for ticker symbol "CC" on the date of 2022-05-23 was $149.99.

Zero-Shot Tool Learning

Prompt:
Instruction:
For each natural language question below, generate the corresponding Python function call using the get_stock_price_on_date(ticker, date, price_type) function.
Documentation:
Use the ticker parameter for the stock’s ticker symbol mentioned in the question.
Use the date parameter for the specified date in the question.
The price_type parameter should be either "highest" or "lowest".

Question:
What was the highest stock price for ticker symbol ‘CC’ on the date of 2023-09-22?
Answer of the Python function call:

Generation: \n“‘python\nget_stock_price_on_date(‘CC’, ‘2023-09-22’, ‘highest’)

Llama 2 7B Chat:
Prompt:
Instruction:
For each natural language question below, generate the corresponding Python function call using the get_stock_price_on_date(ticker, date, price_type) function.
Documentation:
Use the ticker parameter for the stock’s ticker symbol mentioned in the question.
Use the date parameter for the specified date in the question.
The price_type parameter should be either "highest" or "lowest".

Question:
What was the highest stock price for ticker symbol ‘CC’ on the date of 2023-09-22?
Answer of the Python function call:

Generation: get_stock_price_on_date(‘CC’, ‘2023-09-22’, ’highest’)

Table 5: Examples of vanilla zero-shot and zero-shot tool learning output for the Llama2 models on the stock price query task.

B Few-Shot Prompt Templates

Financial Acronym Task

Question: In finance, what does the acronym "IPO" stand for?
Answer: Initial Public Offering.

Question: In finance, what does the acronym "ROI" stand for?
Answer: Return on Investment.

Question: In finance, what does the acronym "ETF" stand for?
Answer: Exchange-Traded Fund.

Question: In finance, what does the acronym "ESG" stand for?
Answer: Environmental, Social, and Governance.

Question: In finance, what does the acronym "APR" stand for?
Answer: Annual Percentage Rate.

Question: In finance, what does the acronym "[test acronym]" stand for?
Answer:

Stock Symbol Task

Question: In finance, what is the company name for the stock ticker "ADA"?
Answer: The company name for the stock symbol "ADA" is Adams Plc.

Question: In finance, what is the company name for the stock symbol "GBG"?
Answer: The company name for the stock symbol "GBG" is Gb Group Plc.

Question: In finance, what is the company name for the stock symbol "LGRS"?
Answer: The company name for the stock symbol "LGRS" is Loungers Plc.

Question: In finance, what is the company name for the stock symbol "AIG"?
Answer: The company name for the stock symbol "AIG" is American International Group.

Question: In finance, what is the company name for the stock symbol "RGR"?
Answer: The company name for the stock symbol "RGR" is Sturm Ruger & Company.

Question: In finance, what is the company name for the stock symbol "MAIN"?
Answer: The company name for the stock symbol "MAIN" is Main Street Capital Corp.

Question: In finance, what is the company name for the stock symbol "[stock symbol]"?
Answer:

Table 6: The few-shot demonstration templates used in the financial two abbreviation task.

Question: In the context of finance, explain the concept of "Compound Interest".
Answer: Compound interest is the interest calculated on the initial principal, which also includes all the accumulated interest from previous periods on a deposit or loan. Thought of as "interest on interest," it can significantly increase savings or costs over time, as the interest is calculated on a cumulative basis. Unlike simple interest, which is calculated only on the principal amount, compound interest is calculated on the principal amount plus any interest that's been added previously.

Question: In the context of finance, explain the concept of "Hedge Funds".
Answer: Hedge funds are alternative investments using pooled funds that employ different strategies to earn active return, or alpha, for their investors. Hedge funds may use a wide range of strategies, including short selling, leverage, swaps, derivatives, and arbitrage. They are typically open to a limited range of investors and require a large initial minimum investment. Hedge funds are known for their aggressive investment strategies aimed at yielding high returns.

Question: In the context of finance, explain the concept of "Equity".
Answer: In finance, equity refers to the ownership of assets after all debts associated with those assets are paid off. For example, in stock trading, equity represents the ownership interest held by shareholders in a corporation, measured in the unit of shares. In the context of real estate, equity means the difference between the property's current market value and the amount the owner still owes on the mortgage. It represents the amount that the property owner would receive after selling the asset and paying off the mortgage.

Question: In the context of finance, explain the concept of "[term]".
Answer:

Table 7: The few-shot demonstration template used in the financial terminology explanation task.

Vanilla Few-Shot

Question: What was the highest stock price for stock symbol "QS" on the date of 2022-05-23?
Answer: $12.1300.

Question: What was the lowest stock price for stock symbol "QS" on the date of 2022-05-23?
Answer: $11.2750.

Question: What was the highest stock price for stock symbol "QS" on the date of 2022-06-22?
Answer: $9.5400.

Question: What was the lowest stock price for stock symbol "QS" on the date of 2022-06-22?
Answer: $8.8100.

Question: What was the highest stock price for stock symbol "QS" on the date of 2022-07-22?
Answer: $12.0900.

Question: What was the lowest stock price for stock symbol "QS" on the date of 2022-07-22?
Answer: $10.8400.

Question: What was the highest stock price for stock symbol "QS" on the date of 2022-08-22?
Answer: $11.1391.

Question: What was the lowest stock price for stock symbol "QS" on the date of 2022-08-22?
Answer: $10.6600.

Question: What was the [lowest/highest] stock price for stock symbol "[stock symbol]" on the date of [date]?
Answer:

One-Shot Tool Learning

Instruction:
For each natural language question below, generate the corresponding Python function call using the get_stock_price_on_date(ticker, date, price_type) function.
Documentation:
Use the ticker parameter for the stock's ticker symbol mentioned in the question.
Use the date parameter for the specified date in the question.
The price_type parameter should be either "highest" or "lowest".

Question:
What was the highest stock price for ticker symbol 'CC' on the date of 2023-09-22?
Answer of the Python function call:
get_stock_price_on_date('CC', '2023-09-22', 'highest')

Question:
What was the [lowest/highest] stock price for stock symbol "[stock symbol]" on the date of [date]?
Answer of the Python function call:

Table 8: The few-shot demonstration template used in the stock price query task.