Wealth Guide at the FinLLM Challenge Task: A Sophisticated Language Model Solution for Financial Trading Decisions

Sarmistha Das[1], **R E Zera Marveen Lyngkhoi**[1], **Sriparna Saha**[1] and **Alka Maurya**[2]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna
[2] CRISIL Limited
sarmistha_2221cs21@iitp.ac.in, zeramarveenlyngkhoi@gmail.com

Abstract

Single stock trading (SST) is surging in the Fintech era, fueled by tech advances and enhanced trading platforms. Meanwhile, natural language processing (NLP) is revolutionizing finance, with advanced AI and large language models (LLMs) leading the charge. This paper represents our participation in FinNLP-AgentScen-IJCAI 2024 (Joint Workshop of the 8th Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning). The primary objective of this task is to assess the capacity of LLMs to execute sophisticated trading decisions ("buy", "sell", "hold") grounded on a fusion of open-source stock and ETF (exchange traded fund) data. Our methodology centers on integrating news articles with their sentiment scores and correlating them with the stock price on the corresponding day, leveraging LLaMA-2-13 billion. Furthermore, we have explored various LLMs, including Mistral, Gemma, subjecting them to knowledge transfer and additional fine-tuning procedures in zero-shot and few-shot settings. Our model secured the first position in the SST task with a 0.926 Sharpe ratio. Our resultant findings underscore, with sufficient context and information, LLMs can perform these tasks effectively without including historical data.

1 Introduction

Single stock trading (SST) is gaining significant traction due to its strong correlation with making money. Due to persistent inflation and the transition to higher trend inflation and interest rates, the global equity markets are expected to face challenges in 2024 [1]. Despite the challenges posed by persistent inflation and the transition to higher trend rates, the New York Stock Exchange (NYSE) maintains its position as the world's largest stock exchange, boasting an equity market capitalization exceeding 28 trillion U.S. dollars as of March 2024[2]. The proliferation of stock-related information disseminated through various channels, such as news outlets and Twitter, has been instrumental in helping investors analyze market trends. This abundance of data has spurred the adoption of Natural Language Processing (NLP) techniques to explore the intricate relationship between textual data and fluctuations in stock prices, as evidenced by the studies conducted by Xu et al. (Xu and Cohen, 2018) and Oliveira et al. (Oliveira et al., 2017). In 2021, Zhou et al. (Zhou et al., 2021) demonstrated the importance of textual features (e.g., bag-of-words) and sentiments to directly make stock predictions. The advancement of Large Language Models (LLMs) has revolutionized trading agents by addressing many concerns in NLP (Achiam et al., 2023). Models like GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023), and Flan-T5 (Chung et al., 2022) demonstrate unique architectures and diverse language pipelines (Raffel et al., 2020a; Zhuang et al., 2021). These LLMs, with carefully designed prompts, can articulate reasoning and outcomes, enabling immediate scrutiny and adjustment of their decision-making processes. LLMs enhance decision-making by integrating extensive pre-trained knowledge with insights from diverse textual and numerical data sources, overcoming the limitations of isolated environments (Wang et al., 2024). Prompt-guided reasoning

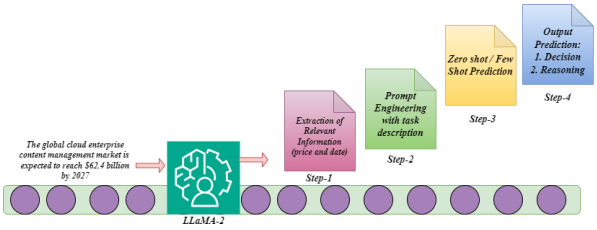

Figure 1: Workflow of our proposed approach with LLaMA-2 in Single Stock Trading prediction task

[1]https://www.troweprice.com

[2]https://www.statista.com

has been shown to significantly improve problem-solving across various domains (Wang et al., 2023). Financial LLMs like FinGPT (Liu et al., 2023), BloombergGPT (Wu et al., 2023), and InvestLM (Yang et al., 2023) are increasingly researched for informed trading decisions, interacting dynamically with financial data and leveraging large parameter configurations. FinMA (Xie et al., 2023), designed for financial instruction tuning, utilizes a dataset of 136K financial samples to enhance its performance in financial decision-making. FinBen (Xie et al., 2024) became the first open-source financial benchmark, encompassing 35 datasets across 23 financial tasks. Subsequently, FINMEM (Yu et al., 2024) is proposed, a novel LLM-based framework for financial decision-making comprising profiling, memory with layered message processing, and decision-making modules.

We participated in the FinLLM challenge, inspired by recent advancements in large language models (LLMs) for finance. It was organized by FinNLP-AgentScen-2024 IJCAI and supported by project JPNP20006 commissioned by NEDO; the competition featured three tasks: financial classification, financial text summarization, and single-stock trading. Our team secured the top position for Task 3, the single stock trading task with a Sharpe ratio of 0.926 by leveraging our model[3]. Subsequently, we secured 4th place in Task 1 based on the financial text classification task and 5th position in Task 2 focused on the financial text summarization task.

2 FinLLM Challenge-Shared Tasks

The FinLLM challenge focuses on evaluating the capabilities of large language models (LLMs) in the financial domain across three primary tasks: 1) Financial classification, which aims to categorize sentences as claims or premises; 2) Financial text summarization, which involves abstracting financial texts into concise summaries; and 3) Single stock trading, which aims to make predictable decisions regarding stock trading. This challenge builds upon the FinLLM (Xie et al., 2024) shared tasks. Our participation concentrated on all three tasks, starting with Task 1, which involved financial text prediction for identifying "premise" or "claim," Task 2 focused on concisely summarising abstract financial news articles with 8,000 training samples and 2,000 test samples and Task 3, which focused on single stock prediction with the primary objective of making informed decisions to "buy," "hold," or "sell" a stock. For Task 1, we were provided with 7,750 training samples and 969 testing samples. In contrast, Task 3 encompassed a dataset of 291 data points, integrating both open-source stock and ETF data. The dataset included separate training and test sets with attributes such as stock_id, price, date, query, and news. Table 1 depicts the instances of the dataset we received for Task 3.

[3]Codes are available here: https://github.com/sarmistha-D/Wealth_Guide-FinLLM2k24

3 Methodology

In this section, we briefly discuss two tasks, with our prime focus on Task 3, followed by Task 1 and Task 2.

3.1 Problem Statement for Task-3

Given a dataset D consisting of financial news texts and associated metadata attributes such as price, date, trading_id, and news as $D = \{(p_i, d_i, t_i, n_i)\}_{i=1}^{N}$; our objective is to predict the trading decision for a particular stock for each record; where, $p_i \in \mathbb{R}$ is the price of the stock at time i, $d_i \in \mathbb{D}$ is the date at time i, $t_i \in \mathbb{T}$ is the trading identifier at time i, $n_i \in \mathbb{N}$ is the financial news text associated with the stock at time i. The prime goal is to learn a function $f : \mathbb{N} \times \mathbb{R} \times \mathbb{D} \times \mathbb{T} \rightarrow \{\text{"buy"}, \text{"sell"}, \text{"hold"}\}$ using an LLM such that for a given set of inputs, the function f predicts the appropriate trading decision.

3.2 Approaches

For single stock prediction, we utilize summarized news N_i and their sentiment scores in conjunction with the stock price P_i. These elements are appended as context to the query Q_i, forming the prompt $\text{Prompt}_i = N_i + P_i + Q_i$. $Prompt_i$ is then fed into the model for zero-shot prediction. The model output Y is preprocessed to extract both the decision d and the reasoning r behind it. The textual decision is converted into a numerical representation with mappings: buy = 1, sell = -1, hold = 0.5, and no decision = 0. To evaluate performance, the score is calculated by multiplying the action taken on each day (action_i) with the return of that day (daily_i), formally expressed as $\text{returns}_i = \text{daily}_i \times \text{action}_i$, where returns_i represents the return on day i, daily_i is the daily return on day i, and action_i is the action taken on day i. The Sharpe ratio is then computed as

Table 1: Sample instance of the given Single Stock Trading dataset

id	jnj_test0
date	"2020-10-09"
price	{ "DRIV": 17.52210235595703 }
filing_k	"FORM": "null"
filing_q	{ "FORM": "null" }
news	"DRIV": ["The global cloud enterprise content management market is expected to reach \$62.4 billion by 2027, driven by a CAGR of 25.6% and significant growth in the U.S. and China. The positive score for this news is 2.3659735504111268e-08. The neutral score for this news is 0.9999990463256836. The negative score for this news is 9.6368637514387 96e-07.", "The global emergency lighting batteries market is expected to reach \$2.8 billion by 2027, growing at a CAGR of 10.8% despite the COVID-19 pandemic's impact. The positive score for this news is 1.166244 1465887241e-05. The neutral score for this news is 0.9995514750480652. The negative score for this news is 0.000436866597738117.", "Despite the impact of the COVID-19 pandemic, the global market for two-wheeler spark plugs is expected to reach 86.2 million units by 2027, growing at a CAGR of 4.9%. The positive score for this news is 1.1285221262369305e-05. The neutral score for this news is 0.9988551139831543. The negative score for this news is 0.0011336031602695584.", "Despite pandemic setbacks, the global market for two-wheeler upside-down forks is expected to reach 701.8 thousand units by 2027, driven by growth in China and the U.S. The positive score for this news is 9.909140175068387e-08. The neutral score for this news is 0.9999970197677612. The negative score for this news is 2.81238385468896e-06.", "The global embedded analytics market is expected to reach \$84.6 billion by 2027, driven by a 13% CAGR,"]}

Sharpe Ratio $= \frac{R_p - R_f}{\sigma_p}$, where R_p is the portfolio's average excess return, R_f is the risk-free rate, and σ_p is the portfolio's volatility, with a Sharpe ratio of 1 being considered good as mentioned in (Yu et al., 2024). For the competition, we tested the llama2-13B model, which was fine-tuned on a financial summarization dataset using zero-shot prompting, as well as other models such as Gemma-7b and Mixtral-7b.

3.3 Sentiment Consideration

In 2021, Zhou et al. (Zhou et al., 2021) showed that textual features (e.g., bag-of-words) and sentiments are crucial for stock predictions, leveraging corporate events as key drivers of stock movements to profit from temporary mispricing. Inspired by this notion, we incorporated sentimental consideration into our model. For the sentiment score of the news the model used was FinBert (Yu et al., 2024) where we have summarized news N_i as input and output is positive score pos_i, negative score neg_i and neutral score neu_i, which are appended to the summarized news $News_i$ to get $N_i = N_i + pos_i + neu_i + neg_i$ as final news input to the model.

3.4 Definition of Task-1

In the FinLLM challenge, we actively participated in Task 1, which entailed a financial text classification task. Our objective was to determine whether a given financial text constitutes a $premise$ or a $claim$. For a given input text T_i, we aim to learn a function, $C : \mathbb{N} \times \mathbb{T} \rightarrow \{\text{"premise"}, \text{"claim"}\}$. We received 7,750 training data samples (Please refer to the appendix section D) To this end, we fine-tuned several state-of-the-art language models across 5-10 epochs each to perform this task.

Table 2: Ablation studies among different generative based language models on Financial Text classification(Task-1) task

MODEL	F1	Accuracy
BERT	72.42	**56.75**
T5-Small	68.43	52.01
Bart	64.29	47.37
DistilBert	**75.13**	50.14

3.5 Definition of Task-2

Our participation in Task 2 explores the capabilities of large language models (LLMs) in summarizing financial documents. Using a specific prompt template, we framed the input text as multi-sentence financial news and the output as its abstractive summary. Our goal is to learn a function $S : \mathbb{T} \rightarrow \mathbb{S}$, mapping input texts $\mathbb{T}$ to their summaries $\mathbb{S}$. We conducted fine-tuning of a leading-edge language model over 3-5 epochs to achieve this objective.

3.6 Experimental Setup

For the Task-3 experiments, we utilized three language models: Using a zero-shot and few-shot setting, we combined news articles with sentiment scores and stock prices to generate decision predictions using reasoning. These experiments were conducted on an NVIDIA GeForce RTX 3090 24GB, operating with 4-bit precision. All the experiments were conducted using the same hyperparameter settings: temperature =1, top_k=5,do_sample=True, and max_new_token=350. For the financial text classification task(Task-1), the learning rate was set to 2e-05, and the optimizer was Adam, with a weight decay of 0.01.

4 Experimental Results and Discussions

In this section, we briefly discuss the experiments and research findings of Task 1 and Task 3; Task

Table 3: Ablation studies among different language models on SST (Single Stock Trading) prediction task (Task-3); DRIV, JNJ, MSFT are stock names and SR- stands for Sharpe Ratio

Experimental Settings		Models			
		Phi-1.3b	Gemma-7b	Mistral-7b	LLaMA2-13b
Zero shot settings	SR	0.1643	0.1252	0.1037	**0.6976**
	SR- DRIV	-0.1038	0.1329	0.07715	**0.48563**
	SR- JNJ	-0.3207	**0.1176**	0.09508	0.078737
	SR- MSFT	-0.0685	0.1250	0.13916	**1.55556**
Few shot settings	SR	0.89321	0.1132	0.9132	0.6756
	SR- DRIV	0.13109	0.1129	0.06715	**0.42563**
	SR- JNJ	0.0867	**0.1147**	0.09528	0.077437
	SR- MSFT	0.1210	0.1390	0.14976	**1.23556**

Input	Gemma-7B	Mistral-7B	LLaMA-2-13B (Wealth Guide)
Price = 11.237374305725098 News= ['The global recloser control market is expected to grow significantly, driven by increasing demand for power quality and reliability, especially in the electric segment and emerging economies like China. The positive score for this news is 1.1491235518690246e-08. The neutral score for this news is 0.9999998807907104. The negative score for this news is 6.358970239261907e-08.']	Decision: Reasoning: ```python buy ```	{'sentiment': 0.14141012581704443, 'sentiment_score': {'buy': {'value': 0.14141012581704443, 'label': 'buy'}, 'hold': {'value': 0.0, 'label': 'hold'}, 'sell': {'value': 1.0, 'label': 'sell'}}, 'score': {'value': 1.1491235518690246e-08, 'label': 'buy'}, 'text': "The global recloser control market is expected to grow significantly, driven by increasing demand for power quality and reliability, especially in the electric segment and emerging economies like China. The positive score for this news is 1.1491235518690246e-08. The neutral score for this news is 0.9999998807907104. The negative score for this news is 6.358970239261907e-08.", 'text_score': {'value': 0.9999998807907104, 'label': 'neutral'}, 'text_type': 'news'}	Decision: Hold Reasoning: The news indicates that the global recloser control market is expected to grow significantly, which could lead to an increase in demand for power quality and reliability. However, the positive score is relatively low, indicating that the market may not grow as much as expected. Additionally, the neutral score is close to the positive score, indicating that the market may be relatively stable. Therefore, it is best to hold onto the investment and monitor the market trends before making any decisions.

Figure 2: Qualitative Analysis between our proposed LLaMA-2, Gemma-7b and Mistral-7b

2 results are mentioned in appendix section E. We conducted the experiments for Task 1, a binary classification to classify "premise" and "claim" class labels with F1-score and accuracy metrics. We compared the results with four generative language-based models, for instance, BERT (Devlin et al., 2018), BART (Lewis et al., 2019), T5-small(Raffel et al., 2020b) and DistilBERT (Sanh et al., 2019). To conduct Task 3 of the SST task experiment, we compared the results of our model, LLaMA-2-13B (Touvron et al., 2023), with three other LLM models, Gemma-7B (Team et al., 2024), Phi-1.3B (Gunasekar et al., 2023) and Mixtral-7B (Jiang et al., 2023) under two conditions: "few-shot" and "zero-shot" settings.

4.1 Resultant Discussion

Evidently, in the financial text classification task (Task-1) detailed in Table 2, the DistilBERT model demonstrated superior performance compared to other models fine-tuned under identical configurations. Notably, BART exhibited the poorest performance, which can be attributed, in part, to its propensity for accelerated overfitting relative to DistilBERT. This overfitting issue is less pronounced in DistilBERT, likely due to its more compact architecture. Subsequently, Table 3 illustrates the performance comparison between DistilBERT, popular LLM models, and our model, LLaMA-2-13B. Our model significantly outperformed the other models across most metrics, except for the Sharpe Ratio for JNJ in both zero-shot and few-shot settings. In these metrics, Gemma-7B outperformed LLaMA-2-13B in both settings. However, Mistral-7B performed lower than the other LLMs but still better than Phi-1.3b. Conclusively, it is evident that LLMs are a better choice for large datasets than generative language models due to their superior ability to make context-aware inferences. Additionally, Mistral-7B and Gemma-7B exhibited inconsistent outputs when tested multiple times on the same data inputs. In contrast, LLaMA-2-13B consistently provided accurate decisions and reasoning for all test inputs, achieving an overall Sharpe ratio of 0.6976. This consistency and higher performance can be attributed to LLaMA-2-13B's larger model size and greater number of parameters. Figure 2 depicts two output instances where our model LLaMA-2 performed significantly better compared to Mistral-7b and Gemma-7b.

5 Conclusion

In this paper, we presented our approach for the FinNLP-2024 shared task, FinNLP-AgentScen-IJCAI 2024, evaluating LLM capabilities in financial tasks. Using 291 news & price data points without historical data, we leveraged the LLaMA-2-13B model, known for generating coherent and contextually accurate text. We developed sentiment-score-based trading (SST) prediction model to take "hold", "sell", and "buy" decisions by adapting LLaMA-2 to predict news sentiment and determine decisions based on the sentiment score. This approach helped us to secure the top position in Task -3 of the FinLLM-2024 challenge.

Acknowledgement

This work is a collaborative effort between the Indian Institute of Technology Patna and CRISIL. We thank Dr. Sriparna Saha for her thorough review throughout the competition and the writing process. Additionally, the authors acknowledge the assistance provided by CRISIL Limited.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. 2023. Fingpt: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485*.

Nuno Oliveira, Paulo Cortez, and Nelson Areal. 2017. The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with applications*, 73:125–144.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. 2023. Making large language models better reasoners with alignment. *arXiv preprint arXiv:2309.02144*.

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

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.

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

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. 2024. Finmem: A performance-enhanced llm trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 595–597.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the event: Corporate events detection for news-based event-driven trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2114–2124, Online. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th chinese national conference on computational linguistics*, pages 1218–1227.

A Ethical Consideration

This paper includes limited information about various stock names, values, and financial organization names, primarily provided by the FinLLM challenge organizers. The authors do not endorse or recommend investing in any specific stock. The primary objective of this research is to explore the capabilities of large language models (LLMs) in the financial domain and to contribute to the advancement of research in the FinTech sector. The authors have deliberately refrained from offering specific trading advice.

B Limitations and Future work

LLM tend to hallucinate in zero-shot tasks and even few shot task when no proper context are given.Using bigger LLM require a lot of resources to run and train them as opposed to smaller llm. Using smaller llm though has its limitation like sacrificing better performance in terms of output as compared to bigger model.Also diffrent LLM responds differently to the same prompts so Prompt engineering might re required for zero shot and few shot task. Also for task like classification Bigger llm tend to overfit faster as oppose to smaller model during finetuning.

Future Work: Our future endeavours encompass enhancing Task 3 by incorporating historical data, as such data provides crucial insights. Additionally, we aim to extend the task to support multilingualism, thereby creating new opportunities and avenues for research and application.

C Prompt Construction

For efficient results, we tested with multiple prompts, and we found the following two prompt variants to work for multiple models .

Prompt-1:

Instruction:

Given this context: {context} and price: {price}, output only one decision from the square bracket [buy, sell, hold] and provide reasoning on why.

Response:

Decision:

Reasoning:

Prompt-2:

Instruction:

Given this context: {context} and price: {price}, output only one decision from the square bracket [buy, sell, hold] and provide reasoning on why. The output of the Decision should be only one of [buy, sell, hold].

Response:

Decision:

Reasoning:

D Hidden details about Task-1 and Task-3

In the FinLLM challenge, The main focus of Task 1 is on classifying financial texts as either 'premise' or 'claim'. The mapping should be done from input features to the classes 'premise' or 'claim' for a given input text. We utilized 7,500 training samples for fine-tuning various advanced language models over 5-10 epochs each to achieve optimal performance on this task. Table 5 illustrates a sample instance of the given dataset. In Task 1, a total of 8 different teams participated. Our team, Wealth Guide, secured 4th place in this competition. Table 4 depicts the overall performance of the participated teams for task-1.

Table 4: Performance Comparison between the proposed model, Wealth Guide and other competitive teams.

Team Name	Accuracy	F1-Score	MCC
Team Barclays	0.7626	0.5237	0.7427
Albatross	0.7575	0.5174	0.7555
L3iTC	0.7544	0.5149	0.7581
Wealth Guide	0.7513	0.5018	0.7406
Finance Wizard	0.7286	0.4554	0.7008
CatMemo	0.711	0.4199	0.6818
Upaya	0.709	0.4166	0.6941
jt	0.4933	0.0141	0.590

Table 5: The sample instances received for Task-1; Financial text Classification task

Id	query	answer	text	choices
finargeccauc0	Analyze sentences from earnings conference calls and identify their argumentative function. Each sentence is either a premise, offering evidence or reasoning, or a claim, asserting a conclusion or viewpoint. Return only premise or claim. Text: I mean, sometimes it's not that you came up with some brilliant strategy, it's just like really good work consistently over a long period of time. Answer:	premise	I mean, sometimes it's not that you came up with some brilliant strategy, it's just like really good work consistently over a long period of time.	["premise", "claim"]
finargeccauc1	Analyze sentences from earnings conference calls and identify their argumentative function. Each sentence is either a premise, offering evidence or reasoning, or a claim, asserting a conclusion or viewpoint. Return only premise or claim. Text: Even while in International, we're continuing to invest in a lot of areas, we continue to frontload Prime benefits for the newer geographies, we continue to launch new countries as we launch Prime in Australia recently. Answer:	claim	Even while in International, we're continuing to invest in a lot of areas, \we continue to frontload Prime benefits for the newer geographies, we continue to launch new countries as we launch Prime in Australia recently.	["premise", "claim"]

Table 6: The performance comparison between the proposed model, Wealth Guide, and other competitive teams for the Summary Generation oriented Task-2

Team	**Metrics**					
	Rouge-1	**Rouge-2**	**Rouge-L**	**BertScore**	**BartScore**	**DLT**
Finance Wizard	0.521037018	0.34060938	0.473530112	0.90836845	-3.497988865	1.7346
Upaya	0.529459817	0.358203218	0.486046685	0.910644962	-3.45155009	0.8332
Wealth Guide	0.308893532	0.179468097	0.281924302	0.85959909	-4.961457408	-
Albatross	0.369077581	0.201058395	0.322684316	0.872049115	-3.933526929	-
LBZ	0.534616211	0.358105428	0.492179554	0.911732047	-3.407560172	-
L3iTC	0.366093426	0.187210467	0.304610677	0.875037043	-4.257126737	-
Revelata	0.500411369	0.333023818	0.464356474	0.907018743	-3.805486962	-

For task-3, we received the dataset of financial news articles along with information such as stock prices, dates, and trading IDs. Our goal is to predict whether to "buy," "sell," or "hold" a stock based on this data. Each record in our dataset includes a) The stock price at a certain time, b) The date, c) A unique trading identifier, and d)The financial news text related to the stock at that time. To achieve significant predictions, we used a model called FinBert to analyze the sentiment of the news articles. FinBert summarizes each news article and provides three scores: positive, negative, and neutral. We then combine these scores with the summarized news to create a comprehensive input for our model. This combined input enhances our model's ability to accurately predict the appropriate trading decision. In Task 3, a total of four teams participated, and our proposed model, Wealth Guide, secured the top position. The detailed results are presented in Table 7.

Table 7: For Task-3, the performance comparison between the proposed model, Wealth Guide and other competitive teams

Metrics	Team			
	Wealth Guide	Upaya	Albatross	CatMemo
Sharpe Ratio	0.926385228	-0.467489019	-0.48383204	-0.619939784
Sharpe Ratio- DRIV	0.485625528	-0.380232272	-0.251306057	-1.393291177
Sharpe Ratio- FORM	1.585611423	-0.108506918	-1.435471054	0.175932289
Sharpe Ratio- JNJ	0.078737051	-1.102831656	-1.558522674	-0.383243051
Sharpe Ratio- MSFT	1.555566911	-0.278385232	1.309971626	-0.879157198

E Task 2 in brief

Task 2 aims to evaluate the effectiveness of large language models (LLMs) in summarizing financial documents. We were provided with 8,000 training samples and 2,000 test samples to transform financial news articles into concise summaries Using a specific prompt template—"Instruction: [task prompt] Context: [input context] Response: [output].". For this task, we finetuned the LLAMA2-13B model. The evaluation metrics utilized were the ROUGE score and BERT score. In the competition, LLAMA2-13B achieved a ROUGE-1 (R1) score of 0.3088, which was used for the final competition ranking, as indicated in Table 6. The

LLAMA2-13B model was finetuned with a learning rate of 2e-03 using the AdamW optimizer on a 24GB NVIDIA RTX 3090 GPU.