CatMemo at the FinLLM Challenge Task: Fine-Tuning Large Language Models using Data Fusion in Financial Applications

Yupeng Cao*, Zhiyuan Yao*, Zhi Chen*, Zhiyang Deng*
*Equal Contribution
Stevens Institute of Technology, Hoboken, NJ
{ycao33,zyao9,zchen100,zdeng10}@stevens.edu

Abstract

The integration of Large Language Models (LLMs) into financial analysis has garnered significant attention in the NLP community. This paper presents our solution to IJCAI-2024 FinLLM challenge, investigating the capabilities of LLMs within three critical areas of financial tasks: financial classification, financial text summarization, and single stock trading. We adopted Llama3-8B and Mistral-7B as base models, fine-tuning them through Parameter Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA) approaches. To enhance model performance, we combine datasets from task 1 and task 2 for data fusion. Our approach aims to tackle these diverse tasks in a comprehensive and integrated manner, showcasing LLMs' capacity to address diverse and complex financial tasks with improved accuracy and decision-making capabilities.

1 Introduction

In recent years, FinTech research has increasingly focused on using textual information to aid investment decisions by analyzing various financial textual data (Allen et al., 2021). However, the complexity of financial documents makes it difficult to classify and summarize market information. Additionally, the intricate and volatile nature of financial markets poses significant challenges for making informed, sequential investment decisions. To address these challenges, advanced natural language processing techniques and models are necessary to process and interpret vast amounts of financial data accurately (Fisher et al., 2016). Lately, Large Language Models (LLMs) have demonstrated impressive capabilities in the field of finance (Bubeck et al., 2023; Li et al., 2023). These models excel in understanding and generating human-like text, making them ideal candidates for tackling complex financial tasks.

Although LLMs demonstrate significant promise in the financial sector, their efficacy in specific financial tasks requires deeper investigation. The FinLLM challenge @ IJCAI-2024 initiative, as introduced in Xie et al. (2024), seeks to investigate the potential of LLMs in analyzing financial documents and enhancing decision-making processes. By leveraging the power of LLMs, the initiative aims to improve the accuracy and efficiency of financial information processing, ultimately aiding in improved investment strategies and a better market understanding.

This paper describes our technical solution for three diverse tasks provided by the FinLLM challenge: financial classification (Sy et al., 2023), text summarization (Zhou et al., 2021), and single stock trading (Yu et al., 2024). The classification task involves distinguishing between claims and premises in financial texts, the summarization task aims to distill extensive financial narratives into succinct summaries, and the trading task focuses on formulating predictive trading decisions based on algorithmic insights.

The core idea of our solution is to fine-tune pre-trained LLMs using PEFT (Mangrulkar et al., 2022) and LoRA (Hu et al., 2021) techniques, leveraging data fusion strategy on the provided datasets from task 1 & 2 in the FinLLM challenge. Specifically, we select Llama3-8B (AI@Meta, 2024) and Mistral-7B (Jiang et al., 2023) as the pre-trained base models due to their large number of parameters, which enable them to capture complex patterns and nuances in financial text data—essential for the three tasks in the challenge. Additionally, these models are pre-trained on vast and diverse datasets, providing a broad understanding of language that can be fine-tuned for financial domains, enhancing their versatility and adaptability to specific financial tasks. Furthermore, both models support PEFT and LoRA techniques, allowing efficient and effective specialization for the financial domain, even with limited labeled data.

Proceedings of the Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen), Jeju, South Korea, August 3, 2024

Our extensive experiments conducted on the three shared tasks have yielded significant findings: 1) Mistral-7B outperforms Llama3-8B in terms of both overall performance and its ability to generate well-structured outputs; 2) the fine-tuned model by using the fused data, showed enhanced results on Task 1 and Task 2; 3) however, this fine-tuned model did not demonstrate improvement in the more complex single-stock trading task (Task 3). For this, we do a more detailed analysis of the results in Section 4.

2 Shared Task Description

The FinLLM challenge consists of three shared tasks: financial classification (task 1), text summarization (task 2), and single stock trading (task 3). Datasets description can be found in: https://huggingface.co/datasets/TheFinAI/flare-finarg-ecc-auc_test and https://huggingface.co/datasets/TheFinAI/flare-edtsum_test.

Task 1 in the FinLLM challenge focuses on the **financial classification**, specifically categorizing sentences within financial documents as either claims or premises. A claim is a statement that asserts a point of view or opinion, while a premise provides the supporting information or evidence for that claim. This task is fundamental for understanding and analyzing financial narratives, as it helps in structuring the information into coherent arguments, which is essential for various downstream applications such as sentiment analysis, risk assessment, and investment decision-making. The evaluation metric for Task 1 is the **F1 score**, which provides a balanced measure of the model's precision and recall.

Task 2 in the FinLLM challenge focuses on **financial texts summarization**. The objective is to condense lengthy financial documents into concise summaries that capture the essential information and key insights while omitting redundant or less important details. This task is crucial for enabling quick and effective information processing, allowing stakeholders to make informed decisions without wading through extensive reports. Task 2 utilizes three metrics, namely ROUGE (1, 2, and L) and BERTScore, to evaluate generated summaries in terms of relevance, with the **ROUGE-1 score** serving as the final ranking metric.

Task 3 in the FinLLM challenge focuses on the application of LLMs to **single stock trading**, aiming to make informed and predictive trading decisions. The primary goal of this task is to develop a model that can analyze various financial texts and other relevant data to predict the future price movements of a single stock and make trading decisions based on these predictions. The evaluation metric includes Sharpe Ratio (SR), Cumulative Return (CR), Daily (DV) and Annualized Volatility (AV), and Maximum Drawdown (MD), with the **Sharpe Ratio (SR)** used as the final ranking metric.

3 Proposed Method

The success of large language models like GPT-4 (Achiam et al., 2023) and Llama3 demonstrates the benefits of integrating diverse data sources during pre-training, enhancing their capabilities and generalizability across various real-world applications. This approach not only broadens the model's understanding of different data forms but also significantly boosts performance on specialized tasks through fine-tuning (Nguyen-Mau et al., 2024; Huang et al., 2024). Inspired by these advancements, our work employs a cross-task data fusion strategy for LLM fine-tuning, aiming to enhance the model's effectiveness by combining insights from different financial tasks. Figure 1 illustrates the proposed fine-tuning method.

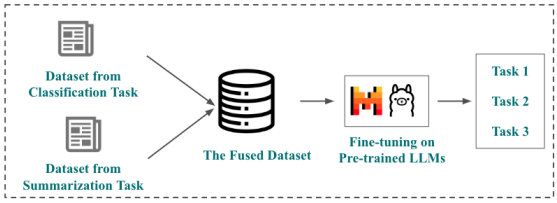


Figure 1: Schematic of proposed fine-tuning method.

We curated and preprocessed a robust training set from two tasks: financial text classification and financial text summarization, to cover a wide range of real-world financial scenarios. We excluded the dataset for task 3, which focuses on texts related to three specific stocks, due to its narrow company-specific content. This selective integration forms the basis for fine-tuning a pre-trained LLM, equipping it to effectively understand and generate nuanced financial texts. After fine-tuning, we applied the enhanced model to each of the three tasks to evaluate its practical utility and performance across various financial applications.

4 Experiment and Discussion

In this section, we present technical details of our implementation and numerical results of our fine-tuned models on tasks 1, 2, and 3. We also compare the performance of these models on different tasks and present our observations on discrepancies between the two base models.

4.1 Experiment Setup

Mistral-7B and Llama3-8B are employed as the base LLM in this study. Due to the limit of computational resources, we perform fine-tuning using Low-Rank Adaptation (LoRA, Hu et al. (2021)) with LoRA-α 16 and 4-bit quantization (Jacob et al., 2018) to reduce the usage of GPU memory and to accelerate training. The models were trained and inferenced on two NVIDIA RTX-A6000 GPUs (each has 48GB DRAM) with one epoch. Our implementation employs PEFT, Quantization libraries and other pipelines provided in Huggingface[1]. We divided the training set portion of the validation set in the ratio of 80:20 for performance evaluation. The models are further tested and compared using the provided testing data sets.

4.2 Experiment Results on Validation Set

In preliminary experiments, we observed a significant difference in performance between the fine-tuned Mistral-7B and Llama3-8B models. Mistral-7B demonstrated superior predictive capabilities and produced well-formatted outputs that could be easily parsed to yield final predictions. In contrast, Llama3-8B required additional processing of its outputs through specific prompting, which could potentially alter the original outputs. Consequently, we decided to conduct all subsequent experiments using Mistral-7B.

4.2.1 Task 1

Table 1 illustrates that the fine-tuned LLMs have significantly improved reasoning for downstream-specific tasks. Furthermore, the LLMs, fine-tuned using the fused dataset, exhibit significant performance enhancements, where it achieves a 0.5634 F1 score. This evidence supports the notion that integrating different tasks can substantially enhance the reasoning capabilities of LLMs.

4.2.2 Task 2

Table 2 also demonstrates that the fine-tuned LLMs,by using the fused dataset, achieved significant performance gains in the text summarization task. This reinforces the idea that integrating various tasks can notably enhance the generalization capabilities of LLMs across different applications.

[1]https://huggingface.co/

Dataset	ACC	F1
No Fine-tune	0.4997	0.1581
Task 1	0.3490	0.3913
Task 1 + Task 2	0.6259	0.5634

Table 1: The performance for two models tasked with classifying sentences as either "premise" or "claim". It includes two key metrics: Accuracy (ACC) and F1 Score (F1). Model "Task 1" was fine-tuned using only the dataset from Task 1, while Model Task 1 + Task 2 used datasets from both Task 1 and Task 2 for fine-tuning.

Dataset	Rouge-1	Rouge-2	BertScore
Task 1	0.4847	0.2921	0.6904
Task 1 + Task 2	0.4920	0.3015	0.6946

Table 2: The performance results for two models tasked with summarizing. It includes metrics for evaluating summarization: Rouge and Bert Score. Model "Task 1" was fine-tuned using only the dataset from Task 1, while Model Task 1 + Task 2 used datasets from both Task 1 and Task 2 for fine-tuning.

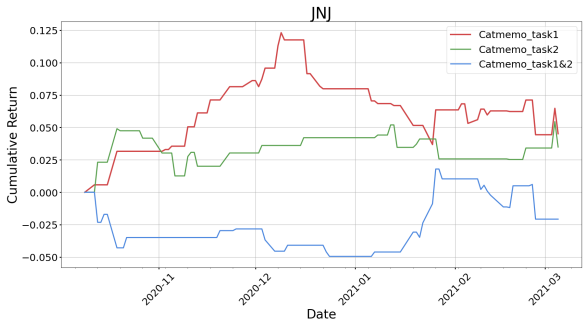
4.2.3 Task 3

We compare the three fine-tuned models based on Mistral-7B in Task 3. We exclude the models fine-tuned from Llama3-8B in this comparison because Llama3-based models cannot consistently produce trading decisions in the correct format. We fine-tuned three models:

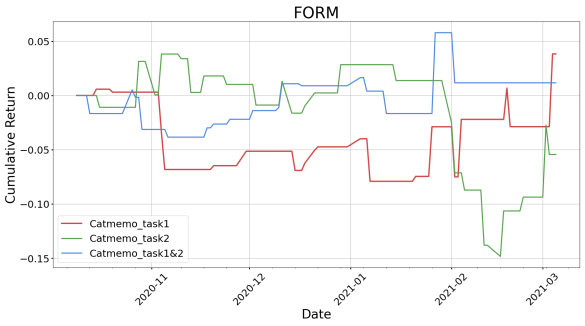
1. Model 1 is fined-tuned only using the training data from Task 1,
2. Model 2 is fined-tuned only using the training data from Task 2,
3. Model 3 is fined-tuned using the training data from Task 1 and Task 2.

The three models are implemented in the FinMem framework as described in Yu et al. (2024) to generate trading decisions.

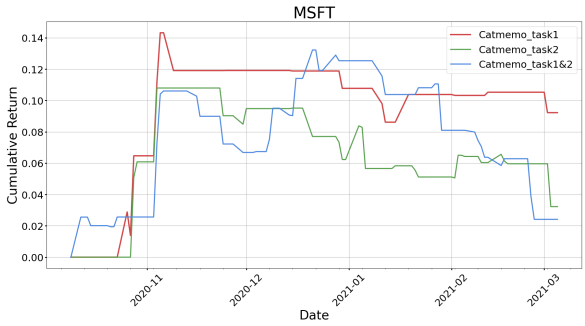
We are interested in the performance discrepancies of these models trained on different datasets. Figure 2 shows the return changes of the three models across four stocks during the testing period. Table 3 details the performance metrics of the models on different stocks. The models generate distinct strategies for all four assets, indicating sensitivity to



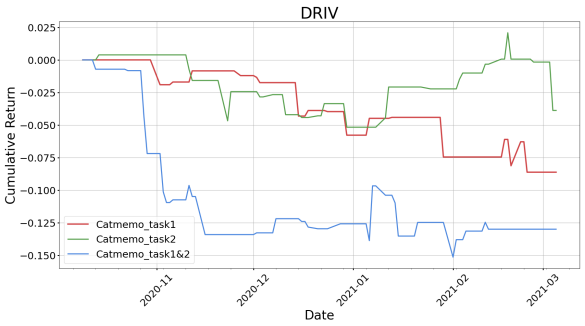
(a) JNJ



(b) FORM



(c) MSFT



(d) DRIV

Figure 2: Comparison of Cumulative Returns in 4 Stocks

	FORM			JNJ			MSFT			DRIV		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
CR↑	0.038	-0.054	0.012	0.045	0.035	-0.021	0.092	0.032	0.024	-0.086	-0.039	-0.130
SR↑	0.440	-0.574	0.176	0.927	0.898	-0.506	1.594	0.564	0.418	-2.139	-0.834	-2.291
SD↓	0.014	0.015	0.010	0.006	0.006	0.009	0.009	0.009	0.009	0.006	0.007	0.009
AV↓	0.217	0.237	0.165	0.101	0.097	0.102	0.144	0.143	0.144	0.101	0.116	0.142
MD↓	0.084	0.175	0.046	0.084	0.059	0.144	0.056	0.074	0.104	0.084	0.059	0.144

Table 3: Performance Metrics Comparison Across Different Models and Datasets.

the fine-tuning datasets. However, none of the models consistently produce profitable strategies. The Mistral-7B model is relatively small compared to state-of-the-art LLMs like OpenAI GPT-4 (Achiam et al., 2023) and Google Gemini (Team et al., 2023), limiting its ability to solve complex tasks such as trading decisions. This aligns with the reported performance of other LLMs in Xie et al. (2024). Additionally, Model 3, trained on both datasets, does not outperform the models trained on each dataset individually. This could be due to the introduction of noise or conflicting information from combining datasets. Given that tasks 1 and 2 are not directly related to trading, it is reasonable that all three models perform poorly in this task.

4.3 Experiment Results on Test Set

Based on the above analysis, we selected the Mistral-7B model, fine-tuned through data fusion, for the final challenge testing. In Task 1, the model achieved an ACC of 0.711, an F1 score of 0.4199, and a Matthews correlation coefficient (MCC) of 0.6818. In Task 3, the integrated Sharp Ratio (SR) was -0.6199. These results are consistent with those observed in our validation set.

5 Conclusion

In this study, we fine-tuned LLMs using datasets that span multiple tasks, resulting in performance improvements in classification and summarization tasks. However, our approach did not yield positive results for the stock trading task. This outcome suggests that more complex financial tasks may require advanced data fusion steps. Furthermore, it underscores the need to explore the impact of incorporating larger datasets on the model's performance after fine-tuning.

Limitation

Our work relies on the pre-trained large language model at 7B/8B level with 4-bit quantization, we have not considered other parameter-level pre-trained models like Llama3-70B which will be explored in the future.

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*.

AI@Meta. 2024. Llama 3 model card.

Franklin Allen, Xian Gu, and Julapa Jagtiani. 2021. A survey of fintech research and policy discussion. *Review of Corporate Finance*, 1:259–339.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. 2016. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hui Huang, Bing Xu, Xinnian Liang, Kehai Chen, Muyun Yang, Tiejun Zhao, and Conghui Zhu. 2024. Multi-view fusion for instruction mining of large language model. *Information Fusion*, page 102480.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 374–382.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Toan Nguyen-Mau, Anh-Cuong Le, Duc-Hong Pham, and Van-Nam Huynh. 2024. An information fusion based approach to context-based fine-tuning of gpt models. *Information Fusion*, 104:102202.

Eugene Sy, Tzu-Cheng Peng, Shih-Hsuan Huang, Heng-Yu Lin, and Yung-Chun Chang. 2023. Fine-grained argument understanding with bert ensemble techniques: A deep dive into financial sentiment analysis. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 242–249.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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*.

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. *arXiv preprint arXiv:2105.12825*.