L3iTC at the FinLLM Challenge Task: Quantization for Financial Text Classification & Summarization

Elvys Linhares Pontes[1] and **Carlos-Emiliano González-Gallardo**[2]
and **Mohamed Benjannet**[1] and **Caryn Qu**[1] and **Antoine Doucet**[2]
[1] Trading Central Labs, Trading Central, Paris, France
[2] University of La Rochelle, L3i, La Rochelle, France
{elvys.linhares_pontes,mohamed.benjannet,caryn.qu}@tradingcentral.com
{carlos.gonzalez_gallardo,antoine.doucet}@univ-lr.fr

Abstract

This article details our participation (L3iTC) in the FinLLM Challenge Task 2024, focusing on two key areas: Task 1, financial text classification, and Task 2, financial text summarization. To address these challenges, we fine-tuned several large language models (LLMs) to optimize performance for each task. Specifically, we used 4-bit quantization and LoRA to determine which layers of the LLMs should be trained at a lower precision. This approach not only accelerated the fine-tuning process on the training data provided by the organizers but also enabled us to run the models on low GPU memory. Our fine-tuned models achieved third place for the financial classification task with an accuracy of 75.44% and sixth place in the summarization task on the official test datasets.

1 Introduction

Financial markets are characterized by their complexity and the vast volume of unstructured data they generate daily. The use of Large Language Models (LLMs) in finance has brought significant focus to tasks involving the analysis, generation, and decision-making related to financial texts. Indeed, LLMs have demonstrated remarkable performance in a large range of applications, from conversational agents to complex decision-making systems. Despite the advances, their potential for thorough analysis and decision-making in finance is still unexplored.

The Financial Challenges in Large Language Models (FinLLM)[1] aims to investigate and enhance the role of LLMs in advancing financial analysis and decision-making processes (Xie et al., 2024). More precisely, it focuses on three applications: financial classification of sentences (Sy et al., 2023), financial news summarization (Zhou et al., 2021), and single stock trading (Yu et al., 2023).

[1] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-agentscen/shared-task-finllm

Motivated by these challenges, we participated (L3iTC) on the financial text classification and financial text summarization tasks. We proposed a fine-tuning process that combine 4-bit quantization and LoRA to optimize several LLMs for each task. This approach accelerated the fine-tuning process on the training data provided by the organizers but also enabled us to run the models on low GPU memory. Our results secured the third place for the financial classification task with an accuracy of 75.44% and sixth place in financial text summarization on the official test datasets.

2 FinLLM Challenge Task

With the advent of LLMs in finance, tasks related to financial text analysis, generation, and decision-making have garnered increasing attention. Key applications in this domain include financial classification, financial text summarization, and single-stock trading. While several approaches utilizing LLMs have demonstrated remarkable performance in these areas, their capabilities for comprehensive analysis and decision-making in finance remain largely unexplored.

FinLLM aims to investigate and enhance the role of LLMs in advancing financial analysis and decision-making processes (Xie et al., 2024).

2.1 Task 1: Financial Classification

The first task aims to evaluate the capabilities of LLMs in identifying and categorizing texts as either premises or claims (Sy et al., 2023). This task is particularly challenging due to financial texts' nuanced and complex nature, where distinguishing between these concepts (claims and premises) requires sophisticated understanding and contextual analysis. The organizers provided a training dataset with 7.75k data examples and the official test dataset composed of 969 examples.

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

2.2 Task 2: Financial Text Summarization

This task is designed to test the capabilities of LLMs in generating coherent and concise summaries (Zhou et al., 2021). The challenge lies in the ability to accurately capture the essential points and nuances of complex financial news, ensuring that the summary remains both informative and coherent (Li et al., 2023). The organizers provided a training dataset with 8k data examples and the official test dataset composed of 2k examples.

2.3 Model Leakage Detection

To measure the risk of data leakage from the test set during model training, organizers have developed a new metric called the Data Leakage Test (DLT), building on existing research (Wei et al., 2023). DLT assesses the risk of data leakage by calculating the difference in perplexity between training and test data for large language models (LLMs). A larger DLT value indicates a lower likelihood of the LLM having seen the test set during training, suggesting a lower risk of model cheating, while a smaller DLT value suggests a higher risk of data leakage and model cheating.

2.4 Evaluation Metrics

For the financial text classification task, the organizers employed two primary evaluation metrics to gauge the performance of the participants' models: F1-score and accuracy. F1-score considers both precision and recall, providing a balanced measure of a model's accuracy. Accuracy represents the ratio of correctly predicted instances to the total instances.

For the financial text summarization task, organizers used ROUGE (1, 2, and L), BERTScore, and BARTScore metrics. ROUGE-n measures the overlap of n-grams between the generated summaries and the reference summaries. BERTScore calculates the similarity between the generated and reference summaries using sentence representation. Finally, BARTScore compares the generated summaries against a reference summary to determine how well the generated summaries capture the reference summaries' meaning, fluency, and coherence.

3 L3iTC Approaches

We participated in the first two tasks. We developed the following architecture to address these tasks to generate our fine-tuned LLM for the FinLLM shared task (Figure 1).

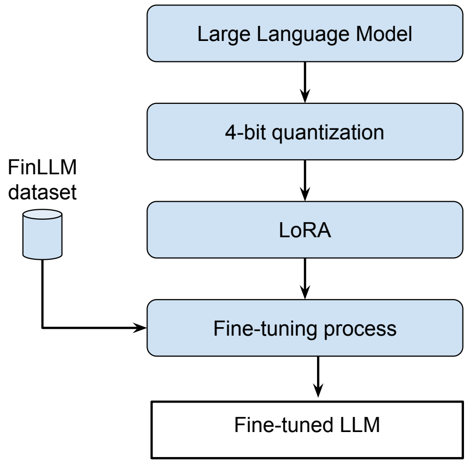

Figure 1: Fine-tuning methodology

3.1 Large Language Models

LLMs can perform a variety of natural language processing tasks such as translation, summarization, and conversational dialogue (Chang et al., 2024). They are trained on diverse datasets encompassing a wide range of topics, enabling them to generate coherent and contextually relevant responses. Among prominent LLMs available today, we selected the following Instruct models due to their high performance and relative small size: Mistral-7B-Instruct-v0.2[2], Mistral-7B-Instruct-v0.3[3], and Meta-Llama-3-8B-Instruct[4].

3.2 Fine-tuning

In the classic fine-tuning of LLMs, a significant portion of model weights is typically modified, necessitating substantial computational resources. To alleviate GPU memory requirements during fine-tuning, we employed quantization techniques as proposed by Dettmers et al. (2022). Specifically, we utilized 4-bit quantization to reduce the memory footprint of LLMs prior to fine-tuning.

To make fine-tuning more efficient, LoRA (Hu et al., 2021) improves efficiency by using two smaller matrices, known as update matrices, to represent weight updates via low-rank decomposition. These matrices are trained to adjust to new data while minimizing the total number of changes. The original weight matrix stays unchanged and is not further modified. The final results are

[2]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[3]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[4]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

achieved by combining the original weights with the adapted ones. In our study, we focused on training parameters within specific modules, including “q_proj”, “k_proj”, “v_proj”, “o_proj”, “gate_proj”, “up_proj”, “down_proj”, and “lm_head” employing a dropout rate of 0.05.

For both tasks undertaken, we partitioned the training dataset into three subsets: train, validation, and test. The validation and test subsets each comprised 10% of the examples, with the remaining 80% constituting the training data. We set the learning rate to 5×10^{-5} and the batch size to 4. The models underwent fine-tuning over 2,000 steps.

4 Preliminary results

Table 1 summarizes the performance of fine-tuning each LLM for Task 1. Initially, the models predicted more than just the target word, attempting to choose the correct class and justify their selection. This approach led to poor results. The best model without fine-tuning was Mistral-7B-Inst-v0.2, which achieved an accuracy of 54% and an F1-score of 0.39.

Despite the limitations encountered during the fine-tuning process, particularly those related to LoRA configuration and 4-bit quantization, all fine-tuned models showed improved performance by generating only the target predicted class. Notably, the best fine-tuned model was FT-Clas-Mistral-7B-Inst-v0.3, which achieved an accuracy of 78% and an F1-score of 0.78. Therefore, we selected it to compete on the first task on the official test dataset.

Team	Accuracy	F1
Mistral-7B-Inst-v0.2	54%	0.39
Mistral-7B-Inst-v0.3	46%	0.36
Meta-Llama-3-8B-Inst	52%	0.48
FT-Clas-Mistral-7B-Inst-v0.2	76%	0.76
FT-Clas-Mistral-7B-Inst-v0.3	**78%**	**0.78**
FT-Clas-Meta-Llama-3-8B-Inst	67%	0.67

Table 1: Preliminary fine-tuning results for financial classification task.

Table 2 summarizes the performance of fine-tuning each LLM for Task 2. Unfortunately, the fine-tuning process did not yield significant improvements in the ROUGE score and even resulted in a decline in BERTScore performance. The main reasons for the poor results are mainly related to our finetuning process. More precisely, the quantization process of 4-bits indeed reduces the amount of GPU memory necessary to fine-tune the model; however, this quantization limited the precision of the learning process which also affected the quality of our models.

Although FT-Sum-Mistral-7B-Inst-v0.2 obtained the best ROUGE-1 score, which is used as the final ranking metric, we found that BertScore better correlates summary quality and human judgment (Table ?? lists some summaries generated by FT-Sum-Mistral-7B-Inst-v0.2 and Mistral-7B-Inst-v0.3). Thus, we selected the Mistral-7B-Inst-v0.3 model for the second task.

Team	ROUGE-1	BertScore
Mistral-7B-Inst-v0.2	22.45	0.5373
Mistral-7B-Inst-v0.3	22.48	**0.5374**
Meta-Llama-3-8B-Inst	22.40	0.5333
FT-Sum-Mistral-7B-Inst-v0.2	**23.12**	0.5097
FT-Sum-Mistral-7B-Inst-v0.3	22.50	0.502
FT-Sum-Meta-Llama-3-8B-Inst	22.31	0.488

Table 2: Preliminary fine-tuning results for financial text summarization task.

5 Official Results

The organizers created a test dataset consisting of 969 test cases for the first task and 2,000 test cases for the second task. The official results are listed in Tables 3 and 4 for tasks 1 and 2, respectively. For the financial classification task, our fine-tuned model achieved notable results, with an accuracy of 75.44% and an F1-score of 0.5149. Remarkably, we secured third place, trailing the first place by less than one percentage point in accuracy.

Team	Accuracy	F1	MCC	DLT
Team Barclays	**76.26%**	**0.5237**	**0.7427**	38.9
Albatross	75.75%	0.5174	0.7555	–
L3iTC	*75.44%*	*0.5149*	*0.7581*	2.2
Wealth Guide	75.13%	0.5018	0.7406	–
Finance Wizard	72.86%	0.4554	0.7008	–
CatMemo	71.10%	0.4199	0.6818	–
Upaya	70.90%	0.4166	0.6941	–

Table 3: Official results for the Financial classification task on the test dataset (Task 1).

As expected, our models did not achieve the best results, primarily due to limitations encountered during the fine-tuning process. Specifically, the use of 4-bit quantization constrained the learning process and adversely affected the quality of summary generation. Consequently, our model ranked sixth for the second task, with a ROUGE-1 score of 0.3661 and a BERTScore of 0.875.

Team	ROUGE-1	ROUGE-2	ROUGE-L	BertScore	BartScore	DLT
LBZ	**0.5346**	**0.3581**	**0.4921**	**0.9117**	**-3.41**	
Upaya	0.5294	0.3582	0.4860	0.9106	-3.45	–
Finance Wizard	0.5210	0.3406	0.4735	0.9083	-3.49	0.8332
Revelata	0.5004	0.3330	0.4643	0.9070	-3.80	1.7346
Albatross	0.3691	0.2010	0.3227	0.8720	-3.93	–
L3iTC	*0.3661*	*0.1872*	*0.3046*	*0.8750*	*-4.25*	–
Wealth Guide	0.3089	0.1794	0.2819	0.8595	-4.96	–

Table 4: Official results for the Financial text summarization task on the test dataset (Task 2).

Id	Mistral-7B-Inst-v0.3	FT-Sum-Mistral-7B-Inst-v0.2
edtsum0	The all-season tire market in Europe is projected to grow by USD 3.42 billion during 2020-2024, with a CAGR of nearly 9% ...	All-Season Tire Market in Europe 2020-2024 | Evolving Opportunities with Bridgestone Corp. and Continental AG | Technavio ...
edtsum2	The chocolate market is projected to grow by USD 41.15 billion from 2020 to 2024, with a CAGR of over 5% ...	Chocolate Market to Grow by USD 41.15 Billion During 2020-2024 | Increasing Premiumization of Chocolates to boost the Market Growth | Technavio Report | English USA - English USA - English USA - ...
edtsum15	Aon PLC has experienced dealings with Farallon Capital Management, L.L.C. on behalf of funds managed jointly with Farallon Partners, L.L.C. ...	Form 8.3 - Aon PLC - 12 April 2021 - Farallon Capital Management, L.L.C. on behalf of funds managed jointly with Farallon Partners, L.L.C. - Amended - 13 April 2021 - 10:00 am - 10:00 am

Table 5: Examples of the generated summaries on the official test dataset. These examples demonstrate that the readability and informativeness are superior in the Mistral-7B-Inst-v0.3 model compared to the FT-Sum-Mistral-7B-Inst-v0.2 model.

6 Conclusion

This article presents our participation (L3iTC) in the FinLLM Challenge Task 2024, concentrating on two primary tasks: Task 1, financial text classification, and Task 2, financial text summarization. To tackle these challenges, we fine-tuned several LLMs to enhance their performance for each specific task.

For Task 1, our fine-tuning efforts led to a third-place finish, achieving an accuracy of 75.44%, just 0.82 percentage points behind the first place. In Task 2, we secured sixth place on the official test datasets. These outcomes demonstrate that combining LoRA configuration and 4-bit quantization allows for the efficient fine-tuning of LLMs minimizing GPU memory and processing time, yielding notable results for tasks that do not require the generation of numerous tokens. In addition, combining quantization and LoRA enables the possibility of fine-tuning LLMs in smaller infrastructures that demand less energy thus reducing their carbon footprint (Samsi et al., 2023). However, when the number of tokens generated increases, as in the case of financial text summarization, this approach reveals its limitations. The quality of the generated summaries declines compared to those produced by the original LLMs without fine-tuning, highlighting the trade-offs involved in using this combination for tasks requiring extensive text generation.

Future work will focus on enhancing the fine-tuning process by employing 8-bit or 16-bit quantization and evaluating their performance on complex tasks such as summarization. Additionally, we aim to perform data augmentation for these datasets and integrate datasets from different tasks. This approach aims to develop a more robust model capable of handling various tasks without compromising the quality of generation.

Acknowledgments

This work has been supported by the ANNA (2019-1R40226), TERMITRAD (2020-2019-8510010), Pypa (AAPR2021-2021-12263410), and Actuadata (AAPR2022-2021-17014610) projects funded by the Nouvelle-Aquitaine Region (France); as well as the France Relance (ANR-21-PRRD-0010-01) project funded by the French National Recherche Agency (ANR). We would like to also thank José G. Moreno for the insightful discussions.

References

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

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. *Preprint*, arXiv:2106.09685.

Haozhou Li, Qinke Peng, Xu Mou, Ying Wang, Zeyuan Zeng, and Muhammad Fiaz Bashir. 2023. Abstractive financial news summarization via transformer-bilstm encoder and graph attention-based decoder. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3190–3205.

Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. 2023. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9. IEEE.

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, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023. Skywork: A more open bilingual foundation model. *Preprint*, arXiv:2310.19341.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. 2024. The finben: An holistic financial benchmark for large language models. *Preprint*, arXiv:2402.12659.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W. Suchow, and Khaldoun Khashanah. 2023. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *Preprint*, arXiv:2311.13743.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the event: Corporate events detection for news-based event-driven trading. *Preprint*, arXiv:2105.12825.