'Finance Wizard' at the FinLLM Challenge Task: Financial Text Summarization

Meisin Lee and **Soon Lay-Ki**

Monash University, Bandar Sunway, Selangor Malaysia

{lee.meisin, soon.layki}@monash.edu

Abstract

This paper presents our participation under the team name 'Finance Wizard' in the FinNLP-AgentScen[1] 2024 shared task 2: Financial Text Summarization. It documents our pipeline approach of fine-tuning a foundation model into a task-specific model for Financial Text Summarization. It involves (1) adapting Llama3 8B, a foundation model, to the Finance domain via continued pre-training, (2) multi-task instruction-tuning to further equip the model with more finance-related capabilities, (3) finally fine-tuning the model into a task-specific 'expert'. Our model, FinLlama3_sum, yielded commendable results, securing the third position in its category with a ROUGE-1 score of 0.521.

1 Introduction

Since the release of Large Language Models (LLMs), they have been swiftly fine-tuned into specialized LLMs in various domains such as biomedical, legal, finance and economics. For the Finance domain, a number of FinLLMs have been released: **FinMA** (also known as PIXIU) (Xie et al., 2023a), **InvestLM** (Yang et al., 2023), **FinGPT** (Wang et al., 2023) and **BloombergGPT** (Wu et al., 2023). As part of expanding the capabilities of Financial LLM, organizers of the FinLLM challenge prepared three subtasks, namely financial classification, financial text summarization and stock trading prediction. For subtask 2 (financial text summarization), participants are given a dataset of financial news article and the task is to produce abstractive summaries for each piece of news.

While all the FinLLMs listed above are trained to be multi-task models capable of a wide variety of finance-related tasks, this paper differs from these FinLLMs where we aim to produce a task-specific model to achieve the best possible score for Financial Text Summarization. The training approach we took is influenced by the findings in (Jang et al., 2023; Lee et al., 2024) that *expert* Language Models (LM) fine-tuned on just a single task can outperform a multi-task LM trained with numerous different tasks. We aim to produce a task-specific model, similar to FinPythia(Xie et al., 2023b) - a model trained specifically for financial sentiment analysis.

This paper documents the considerations and planning that went into producing the task-specific model for the financial text summarization task. This includes the selection of the Foundation Model (both evaluating existing FinLLMs and new LLMs), the selection of training corpus and datasets, as well as the design of the end-to-end fine-tuning approach.

2 Related Work

To the best of our knowledge, there are two survey papers written on Financial Large Language Models (FinLLMs): (1) *Large Language Models in Finance: A Survey*(Li et al., 2023) and (2) *A survey of Large Language Models in Finance (FinLLMs)* (Lee et al., 2024). Based on their findings, the list of FinLLMs and their properties are captured in Table 1. For brevity, only LLMs[2] are listed, finance PLMs (Pretrained Language Models) such as Fin-BERT(Araci, 2019; Yang et al., 2020; Liu et al., 2021) and FLANG(Chung et al., 2024) are omitted.

These FinLLMs are trained on multiple tasks, which includes Sentiment Analysis, Text Classification, Structure Boundary Detection, Named Entity Recognition, Question Answering and more. Just like the foundation LLMs (Radford et al., 2019), these multi-task FinLLMs are also used in in-context learning and in unseen tasks via few-shot

---

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

[2] According to (Zhao et al., 2023), Large LM are models that have 7 billion or more parameters

Type	Model	Foundation	Parameters	Techniques	Task(s)	Context Len.
Multi-task	BloombergGPT (Wu et al., 2023)	BLOOM	50B	PT, PE	SA, TC NER,QA	2048
	InvestLM (Yang et al., 2023)	LLaMa	65B	IFT, PE,	SA,TC QA,Summ	2048
	FinMA (Xie et al., 2023a)	LLaMa2	7B, 30B	IFT, PE	SA,TC,NER QA,SMP	2048
	FinGPT (Wang et al., 2023)	LlaMa2, Falcon, MPT, BLOOM ChatGLM2, QWEN	7B and various sizes	IFT, PE	SA,TC NER,RE	4096
1-Task	Fin-Pythia-1.4B (Rodriguez Inserte et al., 2023)	Pythia	1.4B	PT, PE	SA	2048

Table 1: A summary of the list of Instruction Fine-tuned FinLLMs. List of tasks that these models are fine-tuned on are: SA (Sentiment Analysis), TC (Text Classification), SBD (Structure Boundary Detection), NER (Named Entity Recognition), QA (Question Answering), SMP (Stock Movement Prediction), Summ (Text Summarization), and RE (Relation Extraction).

and zero-shot prompting.

3 Approach

We designed a pipeline approach where it first starts with a foundation LLM, followed by adapting it to the finance domain, and then instruction-tuning it into a task-specific 'expert'. This approach is illustrated in Figure 1, and also described below:

1. Continual pre-training: in the first stage, we further pre-trained a foundation LLM on a collection of Finance-related corpus on the text completion task (causal language modeling) via unsupervised training. This produced **FinLlama3**.
2. Multi-task instruction-tuning: the second stage involves instruction-tuning FinLlama3 on multiple financial tasks via supervised fine-tuning (SFT) to obtain **FinLlama3_mt**.
3. Task-specific Instruction-tuning: in the last stage, the model is instruction-tuned specifically on just one-task - text summarization to produce **FinLlama3_sum**.

3.1 Choice of Foundation Model

For the task of Financial Text Summarization, an LLM's contextual window length plays an important role because the context window need to be large enough to fit the entire news passage so that crucial information is not lost. The open-source Multi-task FinLLMs (FinMA, InvestLM, FinGPT) listed in Table 1 act as good candidates as 'foundation model' for instruction-tuning for Financial Text Summarization. However, their context window length is constrained to be at a maximum of

4,096 (see column 'Context Len' in Table 1 for information). At the point of writing, Llama3 (8B, 70B) is the newest open-source LLM that has the largest context window (8,192 tokens), it is boasted to outperform all of its predecessors. Hence, we chose Llama3 8B as the foundation model.

3.2 Continual Pre-training

Continual pre-training of LLMs aims to adapt an existing general LLM, which is also known as Foundation Models, to a new domain and be enriched with new domain knowledge, in the same way that FinBERT was trained. With Llama3 8B identified as the foundation model, we further pre-trained it on a Finance Corpus we have identified. The corpus is made up of financial texts such as financial news, financial statements, company annual reports, financial research reports, financial literature, market data, etc. Specifically, the corpus is made up of:

1. News from Reuters[3]: The datasets contains 55,700 records of financial news stories from Reuters.
2. News from CNBC, Reuters, WSJ[4]: This dataset contains 18,400 records of financial news from various sources,.
3. Finance and economic glossary (Investopedia)[5]: This dataset contains 6,310 records of financial and economic glossary

---

[3] https://huggingface.co/datasets/rjjan/reuters21578
[4] https://huggingface.co/datasets/Lettria/financial-articles
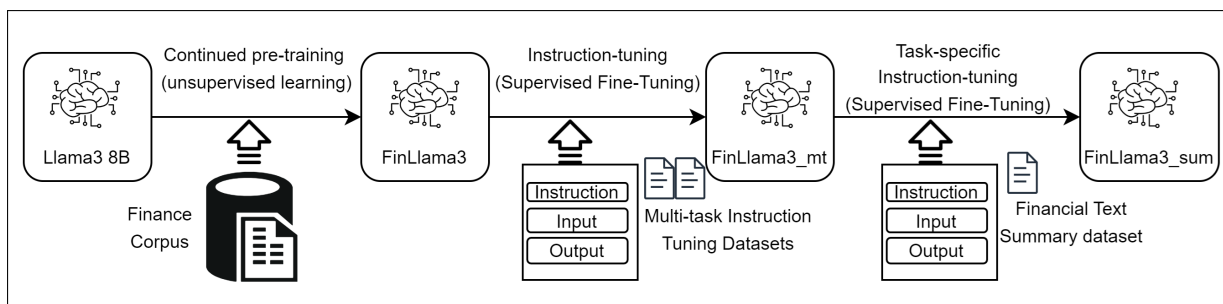[5] https://huggingface.co/datasets/infCapital/investopedia_terms_en

Figure 1: The design of our end-to-end fine-tuning approach. This shows the evolution of a foundation model to the final task-specific 'expert' for financial text summarization.

4. Edgar SEC filings (1993 - 2020)[6]: This dataset comprises financial reports (10-K) submitted to the U.S. Securities and Exchange Commission (SEC). We only used reports for the year 2019.

5. FinWeb[7]: The dataset is a website that provides economic knowledge and information on finance, loans and products.

3.3 Instruction Fine-tuning

Instruction Fine-tuning fine-tunes LLMs on a stream of supervised instruction-following data, this step aims to empower LLMs to follow users' instruction. Instruction-tuning dataset comes in a specific template, consisting of (1) instruction, (2) text/input and (3) answer/output as shown in Table 3.

3.3.1 Multi-task Instruction Tuning

For multi-task instruction tuning, we aligned closely with established financial benchmarks such as the pioneering FLUE benchmark (Shah et al., 2022) and the Fundamental Tasks of the FinBen (Xie et al., 2024) benchmark. We used a public dataset Sujet-Finance-Instruct-177k[8] containing about 177,000 records that covers the following tasks in the finance domain:

1. Sentiment Analysis: Financial Phase Bank (FPB)(MALO et al., 2014), FiQA-SA(Maia et al., 2018)

2. Text classification: News Headlines Classification(Sinha and Khandait, 2021)

3. Named Entity Recognition (NER) (Alvarado et al., 2015)

4. Question Answering FiQA(FiQA), ConvFinQA(Chen et al., 2022)

For more information about this multi-task dataset refer Appendix B.1.

The Multi-Task Instruction-tuning phase aims to train the LLMs on multiple concurrent Finance-NLP tasks, thus improving their versatility and usability on a variety of financial tasks.

3.3.2 Task-Specific Instruction Tuning

In this stage, we fine-tuned the model on a single, specific task - Abstractive Text Summarization. Apart from solely using the training dataset (EDTSUM) provided as part of the shared task challenge, we have also identified other text summarization datasets in the Finance domain. These datasets are combined for the instruction-tuning step:

- Training set (EDTSum) (8,000 records) of financial news summarization
- Earnings call summarization (2,424 records)[9]
- Long earnings calls bullet point summarization (ECTSum)(Mukherjee et al., 2022)[10]
- Edgar Annual Reports (10k) SEC filings summarization (1,000 records)[11]
- Financial news summarization (27,000 records)[12]

4 Experiments and Results

4.1 Training Details

1. **Stage 1: Continual pre-training on Financial Corpus** - we further pre-train Llama3 8B[13] on the financial corpus for 2 epochs to

---

[6] https://huggingface.co/datasets/eloukas/edgar-corpus

[7] https://github.com/deep-over/FiLM/tree/main/pretraining/document/train_dataset/finweb

[8] https://huggingface.co/datasets/sujet-ai/Sujet-Finance-Instruct-177k

[9] https://huggingface.co/datasets/soumakchak/earnings_call_dataset

[10] https://github.com/rajdeep345/ECTSum

[11] https://huggingface.co/datasets/wyx-ucl/SUM-DATASET-BASED-EDGAR-CORPUS

[12] https://huggingface.co/datasets/kdave/Indian_Financial_News

[13] https://huggingface.co/unsloth/llama-3-8b-Instruct-bnb-4bit

	Model	Approach	ROUGE-1	ROUGE-2	ROUGE-L	BertScore	BartScore
(A)	Llama3 8B	zero-shot	0.2020	0.0921	0.1572	0.6506	-4.0342
	FinLlama3	zero-shot	0.2020	0.0923	0.1600	0.6701	-4.0012
(B)	FinMA-7B	zero-shot	0.2021	0.0942	0.1621	0.5935	-4.0941
	FinGPT(LlaMa2)	zero-shot	0.0005	0.0001	0.0005	0.5128	-4.9231
	FinLlama3_mt	zero-shot	0.2044	0.0945	0.1604	0.6517	-3.6981
(C)	**FinLlama3_sum**	*instruction-tuned*	**0.5210**	**0.3406**	**0.4735**	**0.9084**	**-3.4980**

Table 2: Text Summarization results tested on the EDTSum test set. (A): foundation models, (B): multi-tasks FinLLMs, (C): task-specific model. Access to InvestLM model was not available, hence not listed in the table above.

produce **FinLlama3**.

2. **Stage 2: Multi-task instruction-tuning** - FinLlama3 is fine-tuned on multiple financial tasks for another 15 epochs to produce **FinLlama3_mt**. Results and evaluation metrics for each tasks are found in Table 4.

3. **Stage 3: Task-specific Instruction-tuning** - FinLlama3_mt is then fine-tuned on text summarization task for 9 epochs. We utilize three metrics, such as ROUGE (1, 2, and L) (Lin, 2004), BERTScore (Zhang et al., 2019) and BARTScore (Yuan et al., 2021), to evaluate the quality of the generated summaries.

All training was done using Unsloth's Fast-LanguageModel[14] that uses the PEFT (Parameter Efficient Fine-Tuning) library, 4-bit quantization and QLoRA (Quantized Low-Rank Adaptation). The execution was two times faster and used 60% less memory compared to Huggingface transformer library, significantly lower the computational requirements. As for hardware, we used one A100 GPU with 80GB memory, and with `auto-find-batch-size` set to true.

4.2 Results and Discussion

Apart from testing our model using the test set (EDTSUM_test), we also ran other multi-task FinLLMs listed in Table 1 against the same test, the results are captured in Table 2. Llama3 8B is the foundation model that serves as a baseline comparison with the other FinLLMs. Among the models, only two, InvestLM (Yang et al., 2023) and our FinLlama3_sum were specifically instruction-tuned on the financial text summarization task. The rest of the models were evaluated for their text summarization ability via zero-shot prompting.

As expected, multi-task financial models in Group B show better performance compared to Group A's (foundation models) baseline performance. This shows that instruction-tuning a model in multiple Finance-related tasks somewhat improves the model's overall ability to generalize to unseen tasks via zero shot prompting. However, only marginal improvement is observed here. Instead, significant improvement is observed for FinLlama3_sum that was specifically instruction-tuned for the financial text summarization task.

The results also prove that our pipeline approach is successful as the models (FinLlama3 → FinLlama3_mt → FinLlama_sum) show progressive improvements in financial text summarization. While the final model, FinLlama3_sum, performs well in text summarization, it might suffer from catastrophic forgetting on earlier fine-tuned tasks. As part of a rigorous testing, this model should be subjected to benchmark tests such as the FinBen benchmark (Xie et al., 2024), which consist of 35 datasets across 23 financial tasks. It is important to determine if the model suffers from any catastrophic forgetting, and also to test its ability to generalize to a wider suite of financial tasks.

5 Conclusion

Our pipeline approach of turning a foundational model Llama3 8B into a FinLLM (LLM adapted to the Finance domain) and finally to a task-specific 'expert' in Text Summarization has proven to be effective, as the final model is capable of generating summaries with a Rouge-1 score of 0.521 and is ranked 3rd in the challenge. As part of continuous improvement, we intend to explore other open-source LLMs such as Llama3 70B, and other newer models such as Gwen2 72B.

6 Availability

We made our model available here: https://huggingface.co/meisin123/llama3_FinLLM_textsum and codes here: https://github.com/meisin/IJCAI_FinLLMChallenge.

---

[14]Unsloth is a lightweight library for faster LLM fine-tuning, which is fully compatible with the Huggingface ecosystem (Hub, transformers, PEFT, TRL),https://unsloth.ai/

References

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

FiQA. https://sites.google.com/view/fiqa.

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2023. Exploring the benefits of training expert language models over instruction tuning. In *International Conference on Machine Learning*, pages 14702–14729. PMLR.

Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2024. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Pekka MALO, Ankur SINHA, Pekka KORHONEN, Jyrki WALLENIUS, and Pyry TAKALA. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, et al. 2022. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pau Rodriguez Inserte, Mariam Nakhlé, Raheel Qader, Gaetan Caillaut, and Jingshu Liu. 2023. Large language model adaptation for financial sentiment analysis. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, Bali, Indonesia. Association for Computational Linguistics.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335.

Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 589–601. Springer.

Neng Wang, Hongyang Yang, and Christina Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

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 fin-ben: 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. 2023a. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2023b. Efficient continual pre-training for building domain

specific large language models. *arXiv preprint arXiv:2311.08545*.

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

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

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Prompt template

Template:
Instruction: [task prompt]
Text: [input context]
Answer: [output]
Example for Text Summarization
Instruction: "You are given a text that consist of multiple sentences. Your task is to perform abstractive summarization on this text. Use your understanding of the content to express the main ideas and crucial details in a shorter, coherent, and natural sounding text."
Text: *financial news article*
Answer: *abstractive summary of* Text

Table 3: Instruction-tuning Template for Text Summarization

B Multi-task Instruction-Tuning

B.1 Dataset

This dataset, `sujet-ai/Sujet-Finance-Instruct-177k`, can be accesed here: https://huggingface.co/datasets/sujet-ai/Sujet-Finance-Instruct-177k. This dataset is made up of

- **Sentiment Analysis**: 44,209 entries. This involves analyzing financial texts to categorize sentiments as positive, negative, neutral, bearish, or bullish.
- **QA (Question Answering)**: 38,801 entries. Direct-answer finance questions that don't require additional context.
- **QA with Context**: 40,475 entries. Finance questions necessitating context for answers.
- **QA Conversation**: 15,613 entries. This category includes questions following a conversational history between a user and an LLM assistant.
- **Yes/No Question**: 20,547 entries. Questions necessitating a simple yes or no answer.
- **Topic Classification**: 16,990 entries. Tasks requiring the classification of financial texts into specific finance-related categories.
- **NER (Named Entity Recognition) Sentiment Analysis**: 962 entries. This involves conducting sentiment analysis at the entity level within texts.

B.2 Multi-task Instruction-tuning results

Dataset	Metrics	Results
Sentiment Analysis	F1	0.69
	Acc	0.71
Text Classification	AvgF1	0.77
NER	Entity F1	0.57
QA	EmmAcc	0.46

Table 4: Multi-task Instruction Tuning Results based on each tasks.