Proxy Tuning for Financial Sentiment Analysis: Overcoming Data Scarcity and Computational Barriers

Yuxiang Wang[1*], **Yuchi Wang**[1*], **Yi Liu**[1], **Ruihan Bao**[2], **Keiko Harimoto**[2], **Xu Sun**[1]
[1]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[2]Mizuho Securities Co., Ltd.
{yuxiangwang, wangyuchi}@stu.pku.edu.cn, imliuyi@pku.edu.cn
{ruihan.bao, keiko.harimoto}@mizuho-sc.com, xusun@pku.edu.cn

Abstract

Financial sentiment analysis plays a pivotal role in the financial domain. However, the task remains challenging due to the nuanced nature of financial sentiment, the need for high interpretability, and the scarcity of high-quality datasets. To address these issues, we leverage recent advancements in large language models (LLMs) and propose to adapt proxy tuning for financial sentiment analysis. Proxy tuning efficiently transfers knowledge from a pre-trained expert model to a controllable base model by incorporating logit differences, steering the base model toward the desired sentiment representation. Our method offers significant advantages: (1) it is training-free, reducing computational demands and data dependency; (2) it achieves promising performance, with a 36.67% improvement over the base model and over 90% of the tuned model's performance; and (3) it is highly adaptable, functioning in a plug-and-play manner without requiring access to model architectures or weights. These results demonstrate the potential of proxy tuning as an efficient and practical solution for financial sentiment analysis in data-scarce scenarios.

1 Introduction

Financial sentiment analysis (Smailovic et al., 2014; Cortis et al., 2017; Du et al., 2024) is a critical task in the financial domain with significant practical applications. For investors, it serves as a barometer of market trends, aiding in predicting fluctuations, formulating strategies, and assessing risks. For financial institutions, it provides valuable signals for algorithmic trading and quantitative investment, enabling strategy innovation and improved asset pricing. For regulators, it helps identify risks such as fraud, market manipulation, and systemic instability by reflecting market participants' decision-making tendencies.

However, this task remains nontrivial and continues to be a significant challenge in both academic research and practical industrial applications. We identify the primary challenges in addressing this problem as follows: **(1) Inherent complexity of the problem:** Compared to traditional sentiment analysis in the NLP community (Medhat et al., 2014), financial sentiment is more nuanced, often expressed in subtle ways and laden with specialized terminology, requiring a higher level of model comprehension. Furthermore, the relationship between sentiment fluctuations and market behavior may be nonlinear or even non-causal. Financial models must exhibit high interpretability to gain the trust of investors and institutions. Therefore, addressing this complex relationship with interpretability remains a significant challenge. **(2) Difficulty in acquiring high-quality datasets:** One fundamental requirement of modern machine learning is access to large-scale, high-quality datasets. However, financial sentiment analysis faces several challenges in this regard: (i) Difficult to collect: Finance is a sensitive domain, and many organizations are reluctant to share their data. (ii) Rapidly changing and time-sensitive: Financial data changes quickly, becoming outdated and unusable in a short period. (iii) Noisy and complex: Financial sentiment analysis often involves data with complex formats, such as tweets related to stock symbols. These non-natural language texts are difficult to interpret. Moreover, the data is often noisy, containing substantial amounts of irrelevant information.

To address these challenges, we turn to recent advancements in large language models (LLMs) (Brown et al., 2020; Ouyang et al., 2022). With large-scale data and training, LLMs have demonstrated emergent capabilities in text understanding and generation (Wei et al., 2022), making them well-suited for understanding complex financial texts and providing relatively reasonable and explainable analysis—both critical in the financial

[*] Equal contribution

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 169–174
January 19–20, 2025. ©2025 Association for Computational Linguistics

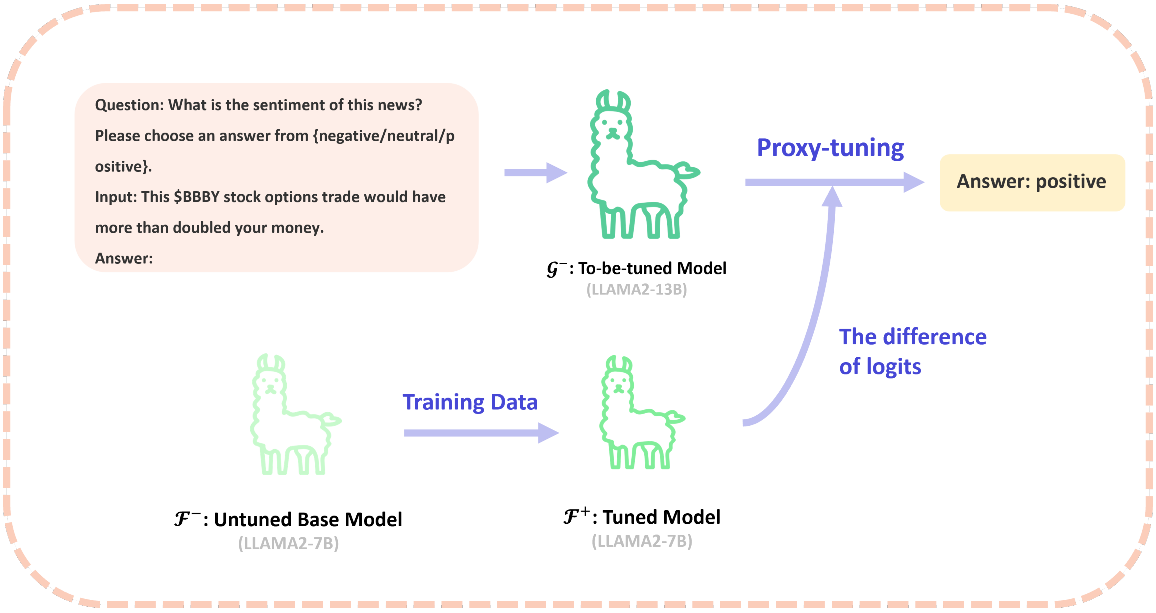
Question: What is the sentiment of this news?
Please choose an answer from {negative/neutral/positive}.
Input: This $BBBY stock options trade would have more than doubled your money.
Answer:
$\mathcal{G}^-$: To-be-tuned Model
(LLAMA2-13B)
Proxy-tuning
Answer: positive
The difference of logits
Training Data
$\mathcal{F}^-$: Untuned Base Model
(LLAMA2-7B)
$\mathcal{F}^+$: Tuned Model
(LLAMA2-7B)

Figure 1: Illustration of adapting proxy tuning into financial sentiment analysis. Instead of directly finetuning LLAMA2-13B, we finetune a LLAMA2-7B to get an expert model. Then we use the expert model to steer the outputs of LLAMA2-13B.

domain. While some studies have explored training financial LLMs for specific tasks (Wu et al., 2023; DeLucia et al., 2022), they often restrict access to the code and training datasets. Moreover, the high cost of training and fine-tuning these models, combined with the difficulty of acquiring the necessary datasets, makes them unaffordable for small companies. Therefore, the idea of **efficiently transferring knowledge from a well-trained black-box expert model to a model that can be fully controlled is both fascinating and important.** To this end, we propose adapting proxy tuning (Liu et al., 2024) for financial sentiment analysis. Specifically, we compute the difference in logits between an expert model and a base model, and incorporate this difference into our untrained model, steering it toward the desired direction.

Through experiments, we summarize the merits of adapting proxy tuning for financial sentiment analysis as follows: **(1) Training-free**: This approach significantly reduces computational resource requirements, while also circumventing the dilemma of needing large-scale, high-quality datasets. **(2) Promising generation results**: Our method demonstrates an average improvement of 36.67% over the base model, achieving over 90% of the performance of the tuned model, which requires substantially more computational resources. **(3) Plug-and-play and easy to adapt**: By manipulating only the logits space without accessing the specific model architecture or weights, our approach is highly adaptable to other models or even different model architectures.

2 Methodology

2.1 Preliminary

2.1.1 Large Language Models

Large Language Models (LLMs) (Ouyang et al., 2022; OpenAI, 2024) are a class of neural network models designed to generate or predict sequences of text. These models are often autoregressive, meaning that they predict the next token in a sequence based on the previous tokens. The autoregressive nature of LLMs can be formalized as follows:

$$P(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} P(x_t | x_1, x_2, \ldots, x_{t-1})$$

Where x_t represents the token at time step t, and T is the length of the sequence.

The vocabulary in LLMs consists of a fixed set of tokens, each corresponding to a unique index in a discrete space. The model's output logits, denoted as z_t, represent unnormalized log-probabilities of tokens in the vocabulary. The softmax function is typically applied to convert logits into a probability distribution:

$$P(x_t = w_i | x_1, \ldots, x_{t-1}) = \frac{\exp(z_{t,i})}{\sum_{j=1}^{V} \exp(z_{t,j})}$$

Where V is the size of the vocabulary and $z_{t,i}$ is the logit for token w_i at time step t. This probability distribution is used to sample or select the next token in the sequence.

2.1.2 Finetune a LLM

If we aim to enable a model to perform a specific task, fine-tuning it with a task-specific dataset is a crucial approach. Fine-tuning LLMs has emerged as an important area of research due to the high computational cost of training these models from scratch. Traditional fine-tuning involves updating all parameters of a pre-trained model, which is both computationally expensive and resource-intensive. To address these challenges, several efficient fine-tuning methods have been proposed, such as Low-Rank Adaptation (LoRA) (Hu et al., 2022) and Prefix Tuning (Li and Liang, 2021).

In essence, fine-tuning adjusts a model $\mathcal{M}^-$ to obtain a fine-tuned version $\mathcal{M}^+$. This process modifies the model's parameters, resulting in changes to its output logits, from $z_t^{\mathcal{M}^-}$ to $z_t^{\mathcal{M}^+}$. Consequently, this also alters the distribution of the output space, effectively transforming the probability distribution P of generating the next token into a new distribution P'.

2.2 Applying Proxy Tuning to Financial LLM

Though there are several methods to efficiently fine-tune large models, the lack of large-scale, high-quality datasets for many financial participants remains a significant challenge hindering the widespread adoption of these methods. To address this, we turn to the emerging proxy tuning (Liu et al., 2024) paradigm, which has recently gained attention in the NLP community. Proxy tuning suggests that, instead of directly tuning a large model, we tune a smaller proxy model and use the difference in predictions between the small tuned model and the untuned model to shift the original predictions of the larger, untuned model in the desired direction. This approach can be particularly useful in scenarios where data scarcity or computational resources are limiting factors.

We follow the basic philosophy of proxy tuning and adapt it for the financial sentiment analysis task. As illustrated in Figure 1, suppose we have a well-finetuned financial sentiment analysis expert model $\mathcal{F}^+$, which has been fine-tuned from a base model $\mathcal{F}^-$. We aim to fine-tune an open, pre-trained base model $\mathcal{G}^-$. A natural idea is that the difference in logits between the tuned expert model and the untuned base model reflects the desired direction for fine-tuning. Specifically, the logit difference, $z_t^{\mathcal{M}+} - z_t^{\mathcal{M}-}$, indicates the expected change in the model's prediction. To push the to-be-tuned base model $\mathcal{G}^-$ in the desired direction, we add this difference to the logit of $\mathcal{G}^-$. Thus, the logit for the tuned base model $\mathcal{G}^-$ can be computed as:

$$z_t^{\mathcal{G}+} = z_t^{\mathcal{G}^-} + (z_t^{\mathcal{F}^+} - z_t^{\mathcal{F}^-})$$

After adjusting the logits, we convert them to probabilities using the softmax function. In this way, the proxy model's adjustments guide the base model $\mathcal{G}^-$ toward the desired fine-tuned behavior, allowing us to leverage the knowledge embedded in $\mathcal{F}^+$ without requiring direct access to large-scale fine-tuning data.

3 Experiments

3.1 Settings

We use the LLAMA2 model family (Touvron et al., 2023) in our experiments. Specifically, we employ the 7B-BASE model as the anti-expert base model, denoted as $\mathcal{F}^-$, and fine-tune it using the LoRA method (Hu et al., 2022) on the datasets mentioned above to obtain the expert model, $\mathcal{F}^+$. We then use the difference between the anti-expert model, $\mathcal{F}^-$, and the expert model, $\mathcal{F}^+$, to steer the 13B-BASE model.

For evaluation, we largely follow the setup of FinGPT (Liu et al., 2023). We apply zero-shot prompting across all datasets and use greedy decoding for our generation strategy. The models are allowed to generate up to 128 tokens. All experiments are conducted on two 48GB A40 GPUs.

3.2 Datasets

We evaluate financial sentiment analysis using four datasets:

Financial Phrasebank (FPB) (Malo et al., 2014): This dataset consists of sentences from English-language financial news about all listed companies in OMX Helsinki, collected from the LexisNexis database. The labels are "positive", "negative", and "neutral".

FiQA-SA (FiQA-2018, 2018): This dataset contains sentences from English-language microblog headlines and financial news. FiQA-SA was first published as part of the 2018 challenge on financial question answering and opinion mining. Although the original dataset is annotated on a con-

Model	FPB	FiQA-SA	TFNS	NWGI	Average
7B					
(1) Directly tuned (expert)	84.81	77.45	87.23	61.06	77.64
13B					
(2) Base (untuned)	38.78	27.64	54.40	40.97	40.45
(3) Proxy-tuned (Ours)	82.43	76.72	88.02	61.30	77.12
(4) Directly tuned	85.15	82.91	88.15	63.92	80.03
Performance Gain	+43.65	+49.08	+33.62	+20.33	+36.67
Closed Gap	94.13%	88.80%	99.61%	88.58%	92.78%

Table 1: **Results for financial sentiment analysis.** For each model size, **Base** refers to the pretrained LLAMA2 model, **Directly tuned** refers to LLAMA2 model finetuned with LoRA, and the **Proxy-tuned** model uses LLAMA2-7B finetuned with LoRA as the expert and LLAMA2-7B as the anti-expert. **Performance Gain** refers to the accuracy gain of Proxy-tuned LLAMA2-13B over LLAMA2-13B untuned. **Closed Gap** refers to the difference in performance between Proxy-tuned LLAMA2-13B and LLAMA2-13B-BASE, divided by the difference between Directly tuned LLAMA2-13B and LLAMA2-13B-BASE.

tinuous scale, we discretize it into a classification task, categorizing it into negative, neutral, and positive classes following the methodology in BloombergGPT's paper (Wu et al., 2023).

Twitter Financial News Sentiment (TFNS) (Zeroshot, 2022): This dataset consists of an annotated corpus of English-language finance-related tweets. The labels are "Bearish" (negative), "Bullish" (positive), and "Neutral".

News With GPT Instructions (NWGI) (Oliverwang, 2023): This dataset contains financial news with ChatGPT-generated labels. It includes seven classification labels: "strongly / moderately / mildly negative", "neutral", "strongly / moderately / mildly positive". We convert these labels into negative, neutral, and positive classes to maintain consistency with the other datasets.

3.3 Results

We evaluate the original untuned model, the proxy-tuned model, and the directly tuned model on the four benchmark datasets listed above. The results are shown in Table 1.

As shown, Model 2, the untuned LLAMA2-13B-BASE model, achieves only 40.45% accuracy on average, which is just slightly better than random guessing (33% for 3 classes), indicating that it has limited knowledge of finance. When using proxy-tuning (Model 3), the accuracy improves by 36.67% on average. On FiQA-SA, the improvement is even more pronounced, reaching up to 50%, suggesting that the model effectively captures financial expert knowledge through proxy tuning.

We then compare the two tuning methods: proxy-tuning and direct tuning. To measure the relative effectiveness of proxy-tuning, we introduce the concept of "closed gap," which quantifies the improvement achieved by proxy-tuning compared to direct tuning. The "closed gap" is calculated as the difference in performance between the proxy-tuned LLAMA2-13B and the untuned LLAMA2-13B-BASE, divided by the difference between the directly tuned LLAMA2-13B and LLAMA2-13B-BASE. On average, proxy-tuning closes 92.78% of the performance gap between the proxy-tuned and directly tuned LLAMA2-13B models across the four benchmarks. This demonstrates that our proxy-tuned model achieves performance comparable to the directly tuned model, while significantly reducing the need for extensive training resources and high-quality datasets, which are common obstacles in financial sentiment analysis.

4 Conclusion

In this paper, we propose a framework leveraging large language models and proxy-tuning to address financial sentiment analysis, overcoming common challenges such as limited data and high computational costs. Our method achieves performance comparable to directly tuned models while being resource-efficient. We hope this work provides insights into efficiently transferring knowledge from expert black-box models to controllable ones and inspires broader applications in the financial domain.

5 Acknowledgement

We thank all the anonymous reviewers for their valuable suggestions. This work is supported by a Research Grant from Mizuho Securities Co., Ltd. We sincerely thank Mizuho Securities for valuable domain expert suggestions. Ruihan Bao and Xu Sun are the corresponding authors.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada. Association for Computational Linguistics.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56:1 – 42.

FiQA-2018. 2018. Fiqa-sa.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024. Tuning language models by proxy. *Preprint*, arXiv:2401.08565.

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

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Oliverwang. 2023. News with gpt instructions.

OpenAI. 2024. Gpt-4 system card.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Jasmina Smailovic, Miha Grcar, Nada Lavra, and Martin Žnidari. 2014. Stream-based active learning for sentiment analysis in the financial domain. *Inf. Sci.*, 285:181–203.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

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. *Preprint*, arXiv:2303.17564.

Zeroshot. 2022. Twitter financial news sentiment.