Direct Preference Optimization Using Synthetic Data: Domain-Specific Model Training and Benchmarking for GPT models

Bohdan Pavlyshenko[1], Ivan Bulka[2]

[1] *Ivan Franko National University of Lviv, Lviv, Ukraine, Drahomanova St, 50, Lviv, Lviv Oblast, 79000*
[2] *Ivan Franko National University of Lviv, Lviv, Ukraine, Drahomanova St, 50, Lviv, Lviv Oblast, 79000*

Abstract
Large Language Models (LLMs) have revolutionized natural language processing, but their usage in specialized domains such as finance is complicated by limitations in domain-specific understanding and the use of specific terminology. This study investigates the adaptation of LLMs, focusing on Meta-Llama-3-8B-Instruct, for advanced financial NLP tasks through a combination of Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically LoRA and QLoRA, and Direct Preference Optimization (DPO) methods. Utilizing the open-source Sujet-Finance-Instruct-177k dataset, which covers six core financial NLP tasks, we demonstrate that PEFT approaches improve performance in tasks such as sentiment analysis and topic classification, while showing limited efficiency in complex generative tasks like question answering. To address this gap, we introduce DPO using synthetically generated preference pairs, enabling supervised alignment based on human-like feedback. Experimental results reveal that DPO enhances the model's performance in challenging question-answering tasks, as evidenced by increased LLM-based evaluation scores. Our findings highlight that while PEFT methods offer efficient domain adaptation, augmenting them with supervised preference optimization is crucial for optimal performance in financial applications.

Keywords
LLM, fine-tuning, DPO, Llama, LoRA, QLoRA, GPT 1

1. Introduction

Large Language Models (LLMs) made a big impact in the field of natural language processing by leveraging transformer-based architectures [1] to achieve near-human text generation and comprehension. However, while these general-purpose models demonstrate impressive capabilities, they often have limitations when applied to specialized domains such as finance, where deep domain expertise and the accurate use of technical terminology are essential. Effectively adapting LLMs to bridge the gap between general linguistic proficiency and domain-specific requirements remains a key challenge.

Furthermore, beyond the generation of text, there is an increasing demand for LLMs to produce outputs that are not only accurate but also optimally aligned with specific user objectives and domain-specific constraints. This underscores the necessity of providing explicit instructions or guidance to steer the models toward generating outputs that best fulfill the intended tasks..

2. Problem and Scientific Novelty

While Retrieval-Augmented Generation (RAG) [2] can help language models become more knowledgeable by allowing them to look at extra information, it has some drawbacks. RAG can

[1] Corresponding authors
bohdan.pavlyshenko@lnu.edu.ua (B. Pavlyshenko); ivan.bulka@lnu.edu.ua (I. Bulka)
0000-0001-9515-3488 (B. Pavlyshenko); 0009-0003-2962-7931 (I. Bulka)

not efficiently handle some domain-specific tasks, understand terminology, or have a good understanding of the domain.

To address these issues, we unified Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA and QLoRA [3]. These techniques allow us to adapt large language models like Meta-Llama-3-8B-Instruct to financial tasks without making the training process too costly or complicated. PEFT methods make it easier to customize models for specialized areas by updating only a small part of the model's parameters [4, 5].

In addition to these approaches, we used Direct Preference Optimization (DPO) [6]. DPO helps models learn directly from human feedback by showing them which responses are preferred. This technique is especially useful when fine-tuned models still make mistakes or give unsatisfactory answers. By using DPO, we aim to further improve the model's performance in financial applications, ensuring it provides more accurate and useful outputs even in difficult cases [7]. We will explore how much DPO can enhance results when standard fine-tuning does not deliver strong enough performance.

3. Results

3.1. Model Training

The experimental framework used the Sujet-Finance-Instruct-177k dataset, an open-source collection from Hugging Face containing 177,000 records across six core financial NLP tasks: sentiment analysis, question answering (standard, with context, yes/no, conversation-based), and topic classification. The dataset was divided into training (85%), validation (10%), and test (5%) splits, ensuring even class distribution.

Initial model fine-tuning was performed on Meta-Llama-3-8B-Instruct utilizing PEFT approaches, specifically LoRA and QLoRA, with standardized configurations to optimize training efficiency on A100 GPUs. Detailed model parameters are described in Table 1.

Table 1
Model Training Config

Parameter	Description	Value
lora_alpha	LoRA scaling factor	16
lora_dropout	Dropout parameter to reduce overfitting	0.1
r	Matrix rank relates to the number of trainable parameters	64
learning_rate	Learning rate	1e-6
optimizer	Optimizer	adam

To further enhance performance on complex tasks, particularly question answering, we introduced Direct Preference Optimization (DPO) during subsequent fine-tuning stages. DPO enables the model to explicitly learn from preference signals by contrasting positive and negative response samples.

3.2. DPO Dataset Construction

Since DPO requires pairs of positive (preferred) and negative (dispreferred) outputs for each training instance, we generated these pairs synthetically. The original dataset answers were treated as positive samples.

To source negative samples, we employed GPT-4o to systematically rewrite the original answers according to specific criteria:

- Simplification: Important information was omitted, thereby reducing answer completeness.
- More Context: Extra, often irrelevant, information was inserted.
- Tone Change: The answer tone was intentionally altered (e.g., shifting between formal or informal), which could impact clarity or appropriateness.

Prompts to GPT-4o included explicit editing instructions and answer modifications, with rewritten outputs collected as corresponding negative samples. Only the modified answers were retained, ensuring a clean dataset for DPO training. With the Synthetic dataset generation, we generated 33k samples for the question-answering user case and used them to tune a model [8].

3.3. DPO training configuration

DPO fine-tuning used the paired positive and negative samples to instruct the model to distinguish high-quality responses from suboptimal alternatives. This preference-based training directly optimizes the model to prefer more appropriate, contextually accurate, and informative outputs for financial question-answering tasks. DPO training configuration provided in Table 2.

Table 2
DPO training configuration

Parameter	Description	Value
lr_scheduler_type	Scheduler type	constant
warmup_steps	Number of warmup steps	150
learning_rate	Learning rate	1e-6
optimizer	Optimizer	adam

3.4. Evaluation

Model evaluation employed LLM-based scoring (using GPT-4 as an evaluator) [9] over the six task categories and demonstrated that, while previous fine-tuning approaches yielded improvements for simpler tasks, they provided limited gains for complex question-answering scenarios (Table 3). Following DPO-based tuning, a marked increase in LLM evaluation scores (Table 4) was observed specifically in the question-answering category, highlighting the effectiveness of supervised preference optimization in specialized financial NLP applications.

Table 3

Metrics evaluation (on test data, metric-score 1 …5 using GPT-4 model)

Task type	Base model	Tuned model
Question-answering	4.77	4.67
QA in conversation	4.43	4.18
QA with context (RAG)	4.75	4.9
Sentiment analysis	4.01	4.03
Topic classification	3.89	4.26
Yes / No questions	2.9	3.14
Avarage	4.13	4.2
Weighed average	4.27	4.33

Table 4
DPO metrics evaluation

Task type	Base model	Tuned model	DPO tuning
Question-answering	4.77	4.67	4.89

4. Conclusion

Our study examined the adaptation of large language models (LLMs) for specialized financial language processing tasks. Initial experiments leveraging parameter-efficient fine-tuning (PEFT) approaches such as LoRA and QLoRA achieved moderate improvements across a range of financial NLP benchmarks, particularly for tasks like sentiment analysis and topic classification. However, our findings also exposed the limitations of PEFT-based adaptation - particularly for complex generative scenarios such as question answering, where conventional fine-tuning did not confer substantial performance gains.

To address these limitations, we used Direct Preference Optimization (DPO) as an additional supervised alignment step, generating preference-labeled training pairs via controlled outputs from GPT-4o. The integration of DPO showed better results, evidenced by clear improvements in LLM-based evaluation scores for challenging question-answering tasks. These outcomes validate the hypothesis that preference-driven optimization can enhance LLM performance in sensitivity-critical, expert-knowledge domains.

Overall, our results suggest that while efficient fine-tuning methods remain invaluable for bridging generalist models into niche sectors, their efficiency is further amplified when coupled with instruction-level supervision that aligns outputs with domain-relevant user expectations. Future

research should explore hybrid approaches, leveraging both retrieval augmentation and advanced tuning, to maximize LLM reliability and utility in highly specialized professional contexts such as finance.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[2] Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. (2024). A Survey on RAG with LLMs. Procedia computer science, 246, 3781-3790.

[3] Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. (2024). A Survey on RAG with LLMs. Procedia computer science, 246, 3781-3790.

[4] Pavlyshenko, B., & Bulka, I. METRIC-BASED COMPARISON OF FINE-TUNED LLAMA 2 AND MIXTRAL LARGE LANGUAGE MODELS FOR INSTRUCTION TASKS. Electronics and information technologies/Електроніка та інформаційні технології, (26).

[5] Pavlyshenko, B., & Bulka, I. (2025). PARAMETER EFFICIENT FINE-TUNING AND OVERFITTING IN GPT LARGE LANGUAGE MODELS: A METRIC-BASED COMPARISON. Electronics and information technologies/Електроніка та інформаційні технології, (30), 33-42.

[6] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems, 36, 53728-53741.

[7] Zhou, Z., Liu, J., Yang, C., Shao, J., Liu, Y., Yue, X., ... & Qiao, Y. (2023). Beyond one-preference-for-all: Multi-objective direct preference optimization.

[8] Dandekar, A., Zen, R. A., & Bressan, S. (2018, August). A comparative study of synthetic dataset generation techniques. In International Conference on Database and Expert Systems Applications (pp. 387-395). Cham: Springer International Publishing.

[9] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., ... & Guo, J. (2024). A survey on llm-as-a-judge. arXiv preprint arXiv:2411.15594.