Received 29 November 2025, accepted 14 August 2026, date of publication 30 September 2026
www.computingonline.net/ computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.25.3.4844

Fine-Tuning Large Language Models for Financial Applications Using Direct Preference Optimization

BOHDAN PAVLYSHENKO[1], IVAN BULKA[2]
[1]System Design Department, Ivan Franko National University of Lviv, Lviv, Ukraine (e-mail: bohdan.pavlyshenko@lnu.edu.ua)
[2]System Design Department, Ivan Franko National University of Lviv, Lviv, Ukraine (e-mail: ivan.bulka@lnu.edu.ua)

Corresponding author: Ivan Bulka (e-mail: ivan.bulka@lnu.edu.ua).

ABSTRACT Large language models have changed natural language processing across many fields, but their general-purpose design limits performance in specialized domains like finance. This study examines the adaptation of Meta-Llama-3-8B-Instruct for financial NLP tasks through parameter-efficient fine-tuning methods and direct preference optimization. We use LoRA and QLoRA techniques with the Sujet-Finance-Instruct-177k dataset covering six financial task categories. While parameter-efficient fine-tuning improves performance on structured tasks like sentiment analysis and topic classification, complex question-answering tasks show limited improvement. To address this limitation, we considered direct preference optimization using synthetically generated preference pairs from GPT-4o. A Task Completion Score was created that uses an LLM as a judge and a GPT-4 model and defines model performance based on instructions in the domain. Results demonstrate that combining parameter-efficient fine-tuning with direct preference optimization increases model performance in challenging question-answering scenarios. Our findings show that preference-based optimization is essential for optimal performance in domain-specific applications requiring both accuracy and alignment with user expectations.

KEYWORDS LLM; PEFT; Direct Preference Optimization; DPO; LoRA; QLoRA; domain adaptation; question answering; synthetic data generation.

I. INTRODUCTION

LARGE Language Models (LLMs) have improved how we work with natural language across different areas through their transformer-based architectures [1]. These models, trained on vast amounts of general text data, demonstrate prominent capabilities in understanding and generating text like a human [2]– [4]. However, their general-purpose nature often limits their effectiveness in specialized domains such as finance, where domain-specific knowledge and terminology are important for optimal performance [5].

The current state of LLM development in specialized domains demonstrates a significant gap between general model capabilities and domain-specific requirements. This gap highlights the need for targeted adaptation approaches that can bridge the divide between general linguistic proficiency and domain expertise. While these models show impressive capabilities in general text understanding, they often struggle with the accurate use of technical terminology and deep domain knowledge required in fields like finance.

This study aims to evaluate Parameter-Efficient Fine-Tuning (PEFT) methods, specifically LoRA and QLoRA techniques [6], [7], for adapting Meta-Llama-3-8B-Instruct (8 billion parameters) to financial domain tasks. Additionally, we explore the impact of vocabulary expansion through domain-specific token addition on model performance across various financial NLP tasks. To address the limitations of standard fine-tuning in complex generative scenarios, we consider Direct Preference Optimization (DPO) [8] as an additional supervised alignment step.

Previous research has shown the effectiveness of metric-based comparisons for evaluating fine-tuned language models across different architectures and tasks [9], [10]. Building on these foundations, this work makes three original contributions, beyond the empirical use of existing PEFT and

DPO techniques. First, we formalize the Task Completion Score (TCS), an LLM-as-a-judge evaluation construct with explicit per-example, per-category, and dataset-level aggregation rules, used throughout this study to measure domain-adaptation quality in a reproducible way. Second, we design a controlled three-criterion protocol, simplification, irrelevant-context insertion, and tone alteration, for synthetically generating dispreferred responses for DPO training, providing a systematic and reproducible alternative to manual preference annotation. Third, we formulate domain adaptation as a constrained selection problem over a set of candidate adaptation operators and use it to show that no single operator dominates across all task categories, which motivates the composite parameter-efficient fine-tuning and preference-optimization strategy evaluated in this study.

II. RELATED WORK

Prior work has documented that general-purpose LLMs, despite strong performance on broad benchmarks, require additional supervision to reliably follow instructions and align with user objectives [5], and that this alignment gap is particularly consequential in specialized domains with dense, evolving terminology, such as derivatives hedging [11] and financial sentiment analysis [12] [13]. A broader survey of reward-design approaches to LLM alignment is given by Ji et al. [14]. LLM-based analytics have also been applied to other high-stakes, fast-evolving domains outside finance, including geopolitical and news analytics [3] and agent-based social simulation [4], indicating that the domain-adaptation problem addressed here is representative of a wider class of applications rather than specific to finance alone.

Retrieval-Augmented Generation (RAG) [15] is a widely used alternative to fine-tuning for domain adaptation, and has been surveyed extensively in the context of LLMs [16], including applications to financial and cryptocurrency news analysis [17]. However, RAG has documented limitations: it can fail to convey terminology and domain understanding that are not explicit in retrieved passages [18], and it faces coherence challenges when integrating retrieved content with generated text, particularly in complex financial reasoning [19]. These limitations motivate parametric domain adaptation (fine-tuning) as a complementary, rather than competing, strategy, which is the focus of the present study.

Full fine-tuning of billion-parameter LLMs is often computationally prohibitive, motivating parameter-efficient fine-tuning (PEFT) methods that update only a small subset of parameters [20]. LoRA [6] and its quantized variant QLoRA [7] are the most widely adopted PEFT techniques for LLMs and are reviewed comprehensively by Yang et al. [21]; both are reproduced in this study.

Direct Preference Optimization (DPO) [8] reparameterizes the reward-modeling step of reinforcement learning from human feedback into a single supervised objective over preference pairs. Multi-objective extensions of preference optimization have been proposed to balance competing alignment criteria [22]. Since human-annotated preference pairs are costly to obtain at scale, synthetic preference-pair generation has been studied as a scalable alternative [23], which is the approach adopted here to construct DPO training data from GPT-4o-rewritten negative samples.

Traditional reference-based metrics such as BLEU and ROUGE have documented limitations in capturing semantic quality and correlate weakly with human judgment on generated text [24] - [26]. This has motivated the use of LLMs themselves as evaluators ("LLM-as-a-judge"), which offers a more scalable alternative to human evaluation while retaining reasonable correlation with human judgments [27] [28], and has been adopted in benchmarks such as MT-Bench and Chatbot Arena [29] and surveyed broadly in [30] [31]. At the same time, LLM-as-a-judge evaluation has documented limitations in expert-knowledge domains [32], which is precisely the setting of this study. The design of the evaluation prompt itself is also known to materially affect LLM outputs [33] - [35].

III. PROBLEM STATEMENT

Let θ_0 denote the parameters of the pretrained base model (Meta-Llama-3-8B-Instruct), and let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the labeled domain corpus (Sujet-Finance-Instruct-177k), partitioned into $K = 6$ task categories $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$ corresponding to sentiment analysis, topic classification, yes/no questions, question answering, question answering with context, and question-answering conversations. Standard supervised fine-tuning seeks parameters θ minimizing the negative log-likelihood

$$\mathcal{L}_{FT}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log \pi_\theta(y_i \mid x_i). \tag{1}$$

A. PARAMETER-EFFICIENT REPARAMETERIZATION

For a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA restricts the update ΔW to a low-rank factorization $\Delta W = BA$, with $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$. The forward pass becomes

$$h = W_0 x + \frac{\alpha}{r} BA\, x, \tag{2}$$

where only $\theta_{adapt} = \{B, A\}$ is trainable and α is the scaling factor (`lora_alpha`), so that $|\theta_{adapt}| \ll |\theta_0|$. QLoRA applies the same update after quantizing W_0 to a lower-precision representation, reducing the memory footprint of θ_0 without altering the objective in Eq. (2).

B. PREFERENCE OPTIMIZATION

To address the degradation of complex generative tasks under objective (1), we require, for each prompt x, a preferred response y_w and a dispreferred response y_l, forming a preference dataset $\mathcal{D}_{pref} = \{(x_i, y_{w,i}, y_{l,i})\}_{i=1}^{M}$. Under the Bradley–Terry model of preference, $p(y_w \succ y_l \mid$

$x) = \sigma\big(r(x, y_w) - r(x, y_l)\big)$, and following the reward-reparameterization of Rafailov et al. [8], the policy θ is optimized directly by minimizing

$$\Delta_\theta(x, y_w, y_l) = \beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{ref}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{ref}(y_l \mid x)}, \tag{3}$$

$$\mathcal{L}_{DPO}(\theta; \pi_{ref}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{pref}}\big[\log \sigma\big(\Delta_\theta(x, y_w, y_l)\big)\big], \tag{4}$$

where π_{ref} is the PEFT-tuned policy prior to preference optimization, σ is the logistic function, and β controls the deviation from π_{ref}.

C. TASK COMPLETION SCORE

Let J denote the judge model (GPT-4), which maps an instruction x_i, a golden response y_i^*, and a generated response $\hat{y}_i = \pi_\theta(x_i)$ to an integer score

$$s_i(\theta) = J(x_i, y_i^*, \hat{y}_i) \in \{1, 2, 3, 4, 5\}. \tag{5}$$

Because J is non-deterministic, each example is scored over R independent evaluation runs ($R = 3$ in this study), giving the per-example Task Completion Score

$$TCS_i(\theta) = \frac{1}{R} \sum_{r=1}^{R} s_i^{(r)}(\theta). \tag{6}$$

For task category k with test subset $\mathcal{D}_k^{test}$ of size $n_k = |\mathcal{D}_k^{test}|$, the category-level score is

$$TCS_k(\theta) = \frac{1}{n_k} \sum_{i \in \mathcal{D}_k^{test}} TCS_i(\theta). \tag{7}$$

We report two aggregates over the K categories: the unweighted (macro) average

$$TCS_{macro}(\theta) = \frac{1}{K} \sum_{k=1}^{K} TCS_k(\theta), \tag{8}$$

and the sample-size-weighted (micro) average

$$TCS_{weighted}(\theta) = \frac{\sum_{k=1}^{K} n_k \, TCS_k(\theta)}{\sum_{k=1}^{K} n_k}. \tag{9}$$

D. FORMAL PROBLEM STATEMENT

Given the base policy θ_0 and the set of candidate adaptation operators studied in this work $\Phi = \{\text{PEFT}, \text{PEFT}_{vocab}, \text{PEFT+DPO}\}$ that produce, respectively, θ_{PEFT}, $\theta_{PEFT_{vocab}}$, and $\theta_{PEFT+DPO}$, this paper addresses the constrained selection problem. Because Table 5 shows that the operator ranking is not consistent across categories, selection is posed at the category level rather than as a single global choice:

$$\theta_k^* = \underset{\theta \in \{\theta_0\} \cup \Phi(\theta_0)}{\operatorname{argmax}} \; TCS_k\big(\theta; \mathcal{D}_k^{test}\big), \quad k = 1, \ldots, K \tag{10}$$

$$\text{s.t.} \quad |\theta_{adapt}| \le \rho \, |\theta_0|, \tag{11}$$

with $\rho \ll 1$ enforcing parameter efficiency. TCS_{macro} and $TCS_{weighted}$ (Eqs. (8)–(9)) are retained as dataset-level diagnostics for comparing operators overall, but are not used as the sole selection criterion in Eq. (10), since a single global argmax over $TCS_{weighted}$ would discard the category-level information that Section V shows to be decision-relevant. Sections IV–V instantiate Φ, report TCS_k, TCS_{macro}, and $TCS_{weighted}$ for every $\theta \in \{\theta_0\} \cup \Phi(\theta_0)$, and show that the operator maximizing TCS_k (Table 5) is not the same across categories, which is the formal basis for evaluating DPO specifically on the category where PEFT and PEFT_{vocab} underperformed (Section V) rather than deploying a single operator globally.

IV. MATERIALS AND METHODS

A. DATASET

The experimental setup utilized the Sujet-Finance-Instruct-177k open-source dataset from Hugging Face, comprising 177,000 training records balanced across six financial NLP task categories: sentiment analysis, question answering with context, question answering, yes-no questions, topic classification, and question-answering conversations. The dataset was partitioned into training (85%), validation (10%), and test (5%) sets while maintaining class balance across splits.

This dataset was chosen for several reasons. First, it provides comprehensive coverage of common financial NLP tasks, allowing us to evaluate model performance across different types of challenges. Second, the size of the dataset is sufficient for fine-tuning while remaining computationally manageable for conducting multiple experiments with different configurations.

B. PARAMETER-EFFICIENT FINE-TUNING

Two variants of the Meta-Llama-3-8B-Instruct model were fine-tuned using PEFT approaches [20]: standard LoRA and QLoRA implementation with rank r = 64 (Table 1), and LoRA and QLoRA with vocabulary expansion (approximately 3,000 finance-specific tokens).

Training was conducted on A100 GPU instances (40GB VRAM) for 2 epochs with a batch size of 4. The Low-Rank Adaptation (LoRA) technique was implemented to reduce computational requirements while maintaining model performance. Quantized Low-Rank Adaptation (QLoRA) was applied to further optimize memory usage during training [21].

The choice of LoRA rank of 64 represents a balance between model expressiveness and computational efficiency. Lower ranks would reduce the number of trainable parameters but might limit the model's ability to capture domain-specific patterns, while higher ranks would increase computational costs without necessarily improving performance.

The learning rate of 1e-6 was selected based on preliminary experiments to ensure stable training without catastrophic forgetting of the base model's capabilities.

Model training parameters are described in Table 1.

Table 1. Model Training Configuration

Parameter	Description	Value
lora_alpha	LoRA scaling factor	16
lora_dropout	Dropout parameter to reduce overfitting	0.1
quantization	Quantization method	4 bit
r	Matrix rank relates to the number of trainable parameters	64
learning_rate	Learning rate	1e-6
optimizer	Optimizer	adam

The vocabulary expansion approach involved identifying and adding approximately 3,000 finance-specific tokens to the model's vocabulary. These tokens were selected based on their frequency in financial texts and their importance for representing domain-specific concepts. This expansion allows the model to more efficiently encode and process financial terminology without relying on subword tokenization, which can sometimes fragment important financial terms into less meaningful components. The new token embeddings were randomly initialized and updated jointly with the LoRA adapters under the same training budget (Table 1).

C. DIRECT PREFERENCE OPTIMIZATION

To further enhance performance on complex tasks, particularly question answering, we utilized 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. This technique is especially useful when fine-tuned models still make mistakes or give unsatisfactory answers.

Since DPO requires pairs of positive (preferred) and negative (dispreferred) outputs for each training instance, we generated these pairs synthetically [23]. 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. This synthetic generation approach offers several advantages over manual annotation, including scalability, consistency, and the ability to control specific types of degradation in answer quality.

The negative samples were generated according to three specific criteria:

- Simplification: Important information was omitted, thereby reducing answer completeness. This type of negative sample helps the model learn to include all relevant details in its responses rather than providing overly brief or incomplete answers.
- More Context: Extra, often irrelevant, information was inserted. This teaches the model to maintain focus and avoid including unnecessary information that could confuse or distract from the main point.
- Tone Change: The answer tone was intentionally altered (shifting between formal or informal), which could impact clarity or appropriateness. This helps the model learn to maintain an appropriate professional tone for financial contexts.

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 33,000 samples for the question-answering use case and used them to tune the model. This sample size was chosen to provide sufficient training data while avoiding overwhelming the model with preference signals that might interfere with the knowledge already acquired during PEFT training.

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 is 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

The constant learning rate scheduler was chosen for DPO training to maintain stable optimization throughout the process. The warmup steps help prevent large gradient updates at the beginning of training that could destabilize the model. The same learning rate as PEFT training (1e-6) was maintained to ensure that preference optimization does not drastically alter the base model's behavior but rather refines it toward better alignment with desired output characteristics.

D. EVALUATION METHODOLOGY

Model evaluation used LLM-based scoring using GPT-4 as an evaluator, motivated by the LLM-as-a-judge literature reviewed in Section II. This approach uses a 1-5 scale with a specific prompt to assess model outputs across the six task categories. Considering the non-deterministic nature of LLMs, three evaluation runs were conducted per Eq. 6, and average metrics are reported.

We designed the Task Completion Score (TCS), formalized in Section III-C, using an LLM as a judge that defines model performance based on instructions in the domain. The evaluation prompt provided to GPT-4 was:

I will provide you with instructions, a golden response, and a generated response. Please take a look at the instructions, golden response, and generated response. You should

return the score, which evaluates how good the generated response is regarding instructions and regarding the golden response. The score should be only an integer number from 1 to 5. The higher the number, the better the score. The more enhanced the response, the higher the score assigned. Return only the score.

This prompt was designed to be clear and unambiguous, ensuring that the evaluator focuses on comparing the generated response to the golden response while considering the original instructions, consistent with prompt-design considerations reviewed in Section II. The 1-5 scale provides sufficient granularity to distinguish between different quality levels while avoiding the complexity of more fine-grained scales that might introduce additional variance in scoring.

V. RESULTS

A. TRAINING EFFICIENCY

Training efficiency analysis revealed minimal difference in computational requirements between standard and vocabulary-expanded models (Table 3). Both approaches demonstrated convergence within the allocated training timeframe. The small difference in training time (approximately 9 minutes over 22 hours) suggests that vocabulary expansion does not significantly increase computational overhead, making it a viable approach for domain adaptation without substantial additional cost.

Table 3. Training Time Comparison

Model	Hours
Llama-3-8B (PEFT)	22:12
Llama-3-8B (PEFT + vocab.)	22:21

Validation loss comparison is important for understanding the training efficiency (Table 4). Both models show similar validation loss trajectories, with slight decreases from epoch 1 to epoch 2, indicating that the models are learning effectively without overfitting. The very small difference in validation loss between the standard and vocabulary-expanded models suggests that both approaches achieve comparable convergence behavior.

Table 4. Validation Loss Comparison

Model/Epoch	1	2
Llama-3-8B (PEFT)	0.79	0.75
Llama-3-8B (PEFT + vocab.)	0.8	0.76

B. PARAMETER-EFFICIENT FINE-TUNING RESULTS

Performance evaluation using TCS across six task categories demonstrated varying effectiveness of domain-specific adaptations (Table 5). The experimental results demonstrate that domain-specific fine-tuning of the Llama-3-8B-based model yields some improvements in financial NLP tasks, with an overall TCS increase from 4.13 to 4.20 (+1.7%) for standard tuning and 4.27 for vocabulary-expanded models (+3%).

The results reveal several important patterns. The vocabulary-expanded model showed particular strength in simpler classification tasks such as yes-no questions (2.90 to 5.00, +72%) and sentiment analysis (4.02 to 4.17, +3.7%). This improvement can be attributed to the fact that these tasks benefit from having more precise representations of financial terms, which vocabulary expansion provides. The dramatic improvement in yes-no questions is particularly notable, suggesting that the model gains a much better understanding of how to formulate clear binary responses when it has access to domain-specific tokens.

The standard tuned model excelled in RAG applications (4.75 to 4.91, +3%), demonstrating that parameter-efficient fine-tuning without vocabulary expansion can still improve the model's ability to integrate retrieved information with generated responses. This suggests that the model learns better strategies for contextual understanding through PEFT alone.

However, complex reasoning tasks like question-answering showed signs of overfitting, with TCS actually decreasing from 4.78 to 4.67 for the standard tuned model and more dramatically to 3.73 for the vocabulary-expanded model. This decrease indicates that the model may be memorizing training patterns rather than learning generalizable reasoning strategies.

These mixed results highlight the need for careful hyperparameter optimization and additional training strategies, particularly for complex generative tasks that require nuanced reasoning rather than pattern recognition.

C. DIRECT PREFERENCE OPTIMIZATION RESULTS

Initial experiments leveraging parameter-efficient fine-tuning approaches such as LoRA and QLoRA achieved moderate improvements across a range of financial NLP benchmarks, particularly for structured 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 TCS increase.

Following DPO-based tuning, a marked increase in LLM evaluation scores (Table 6) was observed specifically in the question-answering category, highlighting the effectiveness of supervised preference optimization in specialized financial NLP applications.

The integration of DPO showed better results, evidenced by clear improvements in LLM-based evaluation scores for challenging question-answering tasks. The TCS increase from 4.67 (after PEFT) to 4.89 (after DPO) represents a 4.7% improvement, which more importantly brings the score above the base model's TCS of 4.78. This recovery and improvement beyond the base model demonstrates that DPO successfully addresses the overfitting issues introduced by PEFT training.

DPO was applied and evaluated specifically for the question-answering category because Table 5 identified it as the category where PEFT degraded performance relative to θ_0 under both the standard and vocabulary-expanded op-

Table 5. PEFT Evaluation on Test Data with TCS

Task Type	Base Model	Tuned Model	Tuned with Embeddings
Question-answering	4.78	4.67	3.73
QA in conversation	4.44	4.19	4.69
QA with context (RAG)	4.75	4.91	4
Sentiment analysis	4.02	4.04	4.17
Topic classification	3.89	4.26	4.05
Yes/No questions	2.90	3.14	5
Average	4.13	4.20	4.27
Weighted average	4.27	4.33	4.14

Table 6. DPO Evaluation with TCS

Task Type	Base	Tuned	DPO
Question-answering	4.78	4.67	4.89

erators; the other five categories, where PEFT or PEFT$_{vocab}$ already satisfied Eq. (10), were not re-evaluated with DPO. A full six-category re-evaluation of $\theta_{PEFT+DPO}$, needed to report $TCS_{weighted}$ for this operator, is left to future work.

The success of DPO in this context can be attributed to several factors. First, by providing explicit examples of what constitutes a good versus a bad answer, the model learns to avoid common failure modes such as being too brief, including irrelevant information, or using inappropriate tone. Second, the preference-based training helps the model develop better judgment about answer quality rather than simply memorizing input-output patterns. Third, the synthetic generation approach ensures that the model sees a diverse range of answer quality levels, helping it generalize better to new questions.

These outcomes validate the hypothesis that preference-driven optimization can materially enhance LLM performance in sensitivity-critical, expert-knowledge domains. The results suggest that combining PEFT with DPO provides a powerful approach for domain adaptation that leverages the efficiency of parameter-efficient methods while addressing their limitations through preference-based refinement.

VI. DISCUSSION

Given the relatively limited size of the dataset for large language model fine-tuning, we expect that increasing the dataset size and optimizing hyperparameters would result in improved performance metrics (TCS). The results demonstrate that while PEFT methods offer efficient domain adaptation, they have inherent limitations when applied to complex generative tasks.

The success of vocabulary expansion for simpler classification tasks suggests that domain-specific tokens can improve model understanding of specialized terminology. However, the decrease in TCS for complex question-answering tasks indicates that vocabulary expansion alone is insufficient for handling nuanced reasoning requirements. This finding has important implications for practitioners: vocabulary expansion should be considered primarily for tasks that benefit from precise term representation rather than as a universal improvement strategy.

Utilizing DPO addresses this gap by providing explicit preference signals that guide the model toward generating more appropriate responses. By learning from contrasting positive and negative examples, the model develops a better understanding of what constitutes a high-quality answer in the financial domain. This approach is particularly effective for tasks where the distinction between adequate and optimal responses is subtle and context-dependent.

One important observation from our experiments is that the base Meta-Llama-3-8B-Instruct model already performs reasonably well on financial NLP tasks, achieving an average score of 4.13 across all task categories. This suggests that large general-purpose models have absorbed considerable financial knowledge during pre-training. However, the variation in TCS across task types (ranging from 2.90 for yes-no questions to 4.78 for question-answering) indicates that some tasks benefit more from the base model's capabilities than others.

The effectiveness of synthetic data generation for DPO training raises interesting questions about the role of human annotation in model training. While our approach relies on GPT-4o to generate negative examples, this automated process offers advantages in terms of scalability and consistency. However, it also raises concerns about whether the synthetic degradation patterns adequately capture the full range of errors that models might make in real-world applications. Future work should investigate the correlation between model performance on synthetic preference pairs and performance on human-annotated preference data.

The computational efficiency of our approach is worth noting. By combining PEFT methods with DPO, we achieve significant performance improvements while training only a small fraction of the model's parameters. This makes the approach accessible to researchers and practitioners who may not have access to the computational resources required for full model fine-tuning.

VII. CONCLUSIONS

Our study examined the adaptation of large language models for specialized financial language processing tasks. Initial experiments leveraging parameter-efficient fine-tuning approaches such as LoRA and QLoRA achieved moderate improvements across a range of financial NLP benchmarks,

particularly for structured 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 shortcomings, we considered Direct Preference Optimization 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, with Task Completion Score increasing from 4.67 to 4.89. These outcomes validate the hypothesis that preference-driven optimization can materially enhance LLM performance in sensitivity-critical, expert-knowledge domains.

The vocabulary expansion experiments revealed task-specific benefits, with dramatic improvements for simple classification tasks (yes-no questions improving from 2.90 to 5.00) but decreased TCS for complex reasoning tasks. This finding suggests that domain adaptation strategies should be tailored to the specific requirements of different task types rather than applied uniformly.

Overall, our results suggest that while efficient fine-tuning methods remain invaluable for bridging generalist models into niche sectors, their efficacy 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.

Several directions for future work emerge from this study. First, investigating the optimal balance between PEFT training duration and DPO training could help prevent overfitting while maximizing the benefits of both approaches. Second, exploring different negative sample generation strategies for DPO could reveal whether more targeted degradation patterns lead to better model performance. Third, evaluating the generalization of our approach to other specialized domains beyond finance would help establish its broader applicability. Finally, comparing synthetic preference data with human-annotated preferences would provide valuable insights into the quality and effectiveness of automated preference generation.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. https://doi.org/10.48550/arXiv.1706.03762.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al., "Language models are few-shot learners," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[3] B. M. Pavlyshenko, "AI approaches to qualitative and quantitative news analytics on NATO unity," arXiv:2505.06313, 2025. https://doi.org/10.48550/arXiv.2505.06313.

[4] L. Zhou, Y. Xu, Z. Wang, and D. Wang, "FlockVote: LLM-empowered agent-based modeling for simulating US presidential elections," arXiv:2512.05982, 2025. https://doi.org/10.48550/arXiv.2512.05982.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al., "Training language models to follow instructions with human feedback," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 27730–27744, 2022. https://doi.org/10.52202/068431-2011

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022. https://doi.org/10.48550/arXiv.2106.09685

[7] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 10088–10115, 2023. https://doi.org/10.52202/075280-0441

[8] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 53728–53741, 2023. https://doi.org/10.52202/075280-2338

[9] B. Pavlyshenko and I. Bulka, "Metric-based comparison of fine-tuned Llama 2 and Mixtral large language models for instruction tasks," *Electron. Inf. Technol.*, no. 26. https://doi.org/10.30970/eli.26.2.

[10] B. Pavlyshenko and I. Bulka, "Parameter efficient fine-tuning and overfitting in GPT large language models: A metric-based comparison," *Electron. Inf. Technol.*, no. 30, pp. 33–42, 2025. https://doi.org/10.30970/eli.30.3.

[11] J. Yang, Y. Tang, Y. Li, L. Zhang, and H. Zhang, "Dynamic hedging strategies in derivatives markets with LLM-driven sentiment and news analytics," arXiv:2504.04295, 2025. https://doi.org/10.48550/arXiv.2504.04295.

[12] G. Iacovides, T. Konstantinidis, M. Xu, and D. Mandic, "FinLlama: LLM-based financial sentiment analysis for algorithmic trading," in *Proc. 5th ACM Int. Conf. AI in Finance*, 2024, pp. 134–141.

[13] F. Xing, "Designing heterogeneous LLM agents for financial sentiment analysis," *ACM Trans. Manage. Inf. Syst.*, vol. 16, no. 1, pp. 1–24, 2025. https://doi.org/10.1145/3688399.

[14] M. Ji, Y. Wu, Z. Wu, S. Wang, J. Yang, M. Dras, and U. Naseem, "A survey of progress in LLM alignment from the perspective of reward design," *IEEE Trans. Artif. Intell.*, 2026. https://doi.org/10.1109/TAI.2026.3654921

[15] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.

[16] M. Arslan, H. Ghanem, S. Munawar, and C. Cruz, "A survey on RAG with LLMs," *Procedia Comput. Sci.*, vol. 246, pp. 3781–3790, 2024. https://doi.org/10.1016/j.procs.2024.09.178.

[17] B. M. Pavlyshenko, "Multilevel analysis of cryptocurrency news using RAG approach with fine-tuned Mistral large language model," arXiv:2509.03527, 2025. https://doi.org/10.48550/arXiv.2509.03527.

[18] H. Essafi, "GenAI, LLM/MLLM, RAG, and their impacts on hallucination, reliability and trustworthiness," in *Proc. Int. Conf. Flexible Query Answering Systems (FQAS)*, Cham, Switzerland: Springer, 2025, pp. 12–15. https://doi.org/10.1007/978-3-032-05607-8_3.

[19] J. Zhang, X. Wang, W. Ren, L. Jiang, D. Wang, and K. Liu, "RATT: A thought structure for coherent and correct LLM reasoning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 25, 2025, pp. 26733–26741. https://doi.org/10.1609/aaai.v39i25.34876.

[20] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," arXiv:2104.08691, 2021. https://doi.org/10.48550/arXiv.2104.08691.

[21] M. Yang, J. Chen, J. Tao, Y. Zhang, J. Liu, J. Zhang, et al., "Low-rank adaptation for foundation models: A comprehensive review," arXiv:2501.00365, 2024. https://doi.org/10.48550/arXiv.2501.00365.

[22] Z. Zhou, J. Liu, C. Yang, J. Shao, Y. Liu, X. Yue, et al., "Beyond one-preference-for-all: Multi-objective direct preference optimization," 2023.

[23] A. Dandekar, R. A. Zen, and S. Bressan, "A comparative study of synthetic dataset generation techniques," in *Proc. Int. Conf. Database and Expert Systems Applications (DEXA)*, Cham, Switzerland: Springer, 2018, pp. 387–395.

[24] N. Schluter, "The limits of automatic summarisation according to ROUGE," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 41–45. https://doi.org/10.18653/v1/E17-2007.

[25] Y. Graham, "Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 128–137.

[26] D. Deutsch, R. Dror, and D. Roth, "On the limitations of reference-free evaluations of generated text," arXiv:2210.12563, 2022. https://doi.org/10.48550/arXiv.2210.12563.

[27] D. Trautmann, A. Petrova, and F. Schilder, "Legal prompt engineering for multilingual legal judgement prediction," arXiv:2212.02199, 2022. https://doi.org/10.48550/arXiv.2212.02199.

[28] N. Lee, J. Hong, and J. Thorne, "Evaluating the consistency of LLM evaluators," in *Proc. 31st Int. Conf. Computational Linguistics*, 2025, pp. 10650–10659.

[29] L. Zheng, W. L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, et al., "Judging LLM-as-a-judge with MT-bench and chatbot arena," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 46595–46623, 2023. https://doi.org/10.52202/075280-2020

[30] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, et al., "A survey on LLM-as-a-judge," *The Innovation*, 2024.

[31] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, et al., "From generation to judgment: Opportunities and challenges of LLM-as-a-judge," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2025, pp. 2757–2791. https://doi.org/10.18653/v1/2025.emnlp-main.138.

[32] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J. J. Li, M. Jiang, and R. A. Metoyer, "Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks," in *Proc. 30th Int. Conf. Intelligent User Interfaces*, 2025, pp. 952–966. https://doi.org/10.1145/3708359.3712091

[33] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, and S. Hasan, "Does prompt formatting have any impact on LLM performance?," arXiv:2411.10541, 2024. https://doi.org/10.48550/arXiv.2411.10541.

[34] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, et al., "The prompt report: A systematic survey of prompt engineering techniques," arXiv:2406.06608, 2024. https://doi.org/10.48550/arXiv.2406.06608.

[35] A. El Amri, *LLM Prompt Engineering for Developers: The Art and Science of Unlocking LLMs' True Potential*. Birmingham, U.K.: Packt Publishing, 2024.

BOHDAN PAVLYSHENKO DSc (AI), professor of System Design Department, Electronics and Computer Technologies Faculty at Ivan Franko National University of Lviv, Ukraine. His current scientific interest lies in quantitative linguistics,machine learning, reinforcement learning, predictive analytics, computer vision, social network mining, time series analytics, numeric modeling,risk assessment, reliability theory, and financial modeling.

IVAN BULKA AI Engineer and PhD student in the System Design Department at Ivan Franko National University of Lviv, Ukraine. His current research focuses on large language models, particularly agentic behavior and decision-making with LLMs. His research interests include parameter-efficient fine-tuning techniques, preference optimization methods, and the application of LLMs in domain-specific contexts. He has published several papers on fine-tuning techniques for large language models and their practical applications in specialized domains.