ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
http://siit.ugatu.su 2026. Vol. 8, No. 3(27)
SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGIES
Scientific paper

UDC 004.8 DOI: 10.54708/SIIT-2026-no3-p105 EDN: BPUXND

Parameter-efficient fine-tuning of Gemma-3 for financial sentiment analysis

L. A. Gardashova • A. E. Gasimov

Азербайджанский государственный университет нефти и промышленности

Financial sentiment analysis plays a central role in market monitoring and risk assessment, yet many state-of-the-art approaches rely on fully fine-tuning large language models with high computational demands. In this work, we examine a resource-efficient alternative by applying Low-Rank Adaptation (LoRA) to the 270M-parameter Gemma-3 model and evaluating its performance on the FinancialPhraseBank (all-agree) dataset. Despite its compact size, the LoRA-tuned model achieves an accuracy of 0.9536, a macro-F1 score of 0.9362, and a weighted-F1 score of 0.9533, clearly improving over the Gemma-7B financial-sentiment baseline reported by Mo et al. while remaining competitive with published FinBERT-style systems on the same benchmark. Additional analysis shows consistent performance across sentiment classes, a macro ROC-AUC of 0.997, well-behaved confidence estimates, and stable calibration. From a practical standpoint, the parameter-efficient setup substantially reduces memory usage and trainable parameter count, enabling effective domain adaptation on widely accessible hardware. These results demonstrate that high-quality financial sentiment classification does not require full fine-tuning of multi-billion-parameter models, and that compact models paired with parameter-efficient methods offer an attractive balance between accuracy, efficiency, and deployability.

Financial sentiment analysis; Gemma-3; Low-Rank Adaptation; LoRA; parameter-efficient fine-tuning; FinancialPhraseBank.

INTRODUCTION

Sentiment analysis plays a central role in financial decision-making, where the emotional tone of news headlines and market commentary can influence investor behavior, asset prices, and market volatility. In domains driven by high-frequency information flow, even subtle sentiment cues can affect risk assessment and short-term allocation strategies. The FinancialPhraseBank dataset remains one of the most widely used benchmarks for this task, providing a controlled environment for assessing sentence-level sentiment classifiers for financial text [Mal14].

The emergence of large language models (LLMs) has reshaped the landscape of NLP. Modern systems are primarily built on the transformer architecture introduced by Vaswani et al. [Vas17], which replaced recurrence with parallelizable self-attention and formed the foundation for pre-trained models such as BERT [Dev19], GPT-style decoders, LLaMA [Tou23], and the Gemma family [Tea25]. These models demonstrate strong capabilities in capturing semantic nuance, contextual dependencies, and domain-specific patterns across diverse text genres, including financial reporting.

A recent study by Mo et al. [Mo24] fine-tuned several transformer-based models–including the 7B-parameter Gemma-7B–and reported strong accuracy, precision, recall, and F1-score on the FinancialPhraseBank dataset. Their reported Gemma-7B configuration uses parameter-efficient

Гардашова Л. А., Гасымов А. Э. Оптимизация параметров Gemma-3 для анализа финансовых настроений // СИИТ. 2026. Т. 8, № 3(27). С. 105-116. (In English). DOI: 10.54708/SIIT-2026-no3-p105. EDN: BPUXND.

Gardashova L. A., Gasimov A. E. “Parameter-efficient fine-tuning of Gemma-3 for financial sentiment analysis” // SIIT. 2026. Vol. 8, no. 3(27), pp. 105-116. DOI: 10.54708/SIIT-2026-no3-p105. EDN: BPUXND.

supervised fine-tuning, which makes it a useful large-model reference point for evaluating whether a much smaller adapter-tuned model can recover similar task performance.

However, many real-world financial applications–especially those involving on-premise analytics, embedded systems, or latency-sensitive decision pipelines–cannot rely on massive models due to hardware constraints, cost limitations, and energy considerations. In addition, the FinancialPhraseBank dataset is relatively small, raising the question of whether large-scale full fine-tuning is even necessary to achieve strong task performance.

Parameter-Efficient Fine-Tuning (PEFT) methods, such as Low-Rank Adaptation (LoRA) [Hu22], offer an efficient alternative by updating only a small number of task-specific parameters while keeping the base model frozen. Recent work has shown that PEFT-tuned small and medium-sized models can match–or even surpass–the performance of fully fine-tuned large models on domain-specific or low-resource tasks.

In this paper, we revisit financial sentiment analysis from an efficiency-centric perspective. We apply LoRA to the 270M-parameter Gemma-3 model [Tea25] and evaluate its performance on the FinancialPhraseBank dataset. Despite its compact size, the LoRA-tuned Gemma-3 model improves on the Gemma-7B baseline reported by Mo et al. [Mo24] in accuracy and F1-score, while requiring only a fraction of the compute, memory, and trainable parameters. Cross-paper comparisons further show that the model is competitive with published FinBERT-style systems, although the strongest domain-specialized encoders still retain an advantage on some 100%-agreement evaluations. These findings highlight the practical utility of compact PEFT-enabled models rather than treating parameter count alone as the main determinant of benchmark performance.

Related Work

Financial Sentiment Analysis and the FinancialPhraseBank

Early work on financial sentiment analysis relied on lexicon-based and classical ML methods applied to financial news and corporate disclosures. Malo et al. introduced the FinancialPhraseBank and demonstrated the importance of domain-specific phrase-level modeling for identifying semantic orientations in economic text [Mal14]. The dataset has since become a standard benchmark for evaluating sentence-level financial sentiment classifiers.

Recent surveys broaden this perspective by reviewing rule-based, machine-learning, and deep-learning approaches to financial sentiment analysis [Du24]. These works highlight both the importance of carefully curated labeled corpora and the growing trend toward transformer-based models capable of capturing linguistic subtleties in financial narratives.

Pre-trained Language Models for Financial Text

The introduction of BERT [Dev19] marked a significant shift in sentiment analysis, replacing hand-crafted features with contextualized embeddings learned from large-scale pre-training. Araci's FinBERT [Ara19], which further adapts BERT using financial-domain corpora, achieved substantial improvements over prior baselines on several financial sentiment tasks, including FinancialPhraseBank. Subsequent variants–such as ProsusAI FinBERT and EnhancedFinSentiBERT [Sun25] – demonstrated that domain-adapted encoders can significantly improve sentiment classification accuracy.

Comparative studies, such as that by Nasiopoulos et al. [Nas25], investigate fine-tuned deep learning models ranging from BERT and FinBERT to proprietary LLMs like GPT-4o, showing that model choice interacts strongly with dataset size and label granularity. While these works underscore the effectiveness of domain-adapted PLMs, they tend to focus on models in the hundreds-of-millions parameter range and rarely examine compact sub-billion models from an efficiency perspective.

Large Language Models for Financial Sentiment

With the rapid development of transformer-based autoregressive LLMs [Vas17, Tou23, Tea25], recent studies have begun exploring their use in financial sentiment classification. Mo et al. [Mo24]

fine-tune several generative models–including DistilBERT, LLaMA, Phi, and Gemma-7B – on the FinancialPhraseBank dataset, reporting that Gemma-7B achieves the strongest overall performance among the evaluated architectures. These results demonstrate that LLMs can effectively capture nuanced sentiment signals in financial text.

Other work has examined fine-tuning or prompting GPT-style LLMs for financial classification and forecasting [Du24, Nas25]. In parallel, FinBERT-style encoder models remain strong FinancialPhraseBank baselines: Araci [Ara19] reports 0.97 accuracy and 0.95 F1 on the 100% agreement subset, Atsiwo [Ats24] reports 0.97 accuracy and 0.96 F1 with synthetic-data augmentation and a long-context FinBERT variant, and Sun et al. [Sun25] report an F1 score of 0.98 for EnhancedFinSentiBERT on the same high-agreement setting. These results provide a useful benchmark for peak task performance, while the larger LLM studies emphasize the trade-off between model scale, fine-tuning strategy, and downstream accuracy.

In contrast, our work evaluates a substantially smaller open model (Gemma-3, 270M parameters) and focuses specifically on whether PEFT enables compact models to match or outperform previously published large-model baselines while remaining competitive with domain-specialized encoders.

Parameter-Efficient Fine-Tuning of Large Language Models

As transformer-based LLMs have grown in size, full fine-tuning has become increasingly resource-intensive. Parameter-Efficient Fine-Tuning (PEFT) methods mitigate this burden by updating only a small fraction of parameters. LoRA [Hu22] introduces trainable low-rank matrices into attention and feed-forward layers, achieving performance comparable to full fine-tuning while drastically reducing memory requirements. QLoRA [Det23] extends this approach by combining quantization with LoRA adapters, enabling efficient fine-tuning of models with tens of billions of parameters on a single GPU.

Other PEFT techniques include IA^3 [Liu22], which modulates internal activations using learned scaling vectors, and a variety of adapter- and prompt-based methods. Recent surveys [Han24, Wan25] provide comprehensive comparisons and highlight that PEFT often maintains downstream performance while drastically reducing training cost, VRAM usage, and total parameter updates.

While PEFT has been widely studied on general NLP benchmarks, relatively few works systematically examine its effectiveness in the financial domain, particularly on small and medium-sized open models. Our study extends this line of research by applying LoRA to a compact 270M-parameter Gemma model and evaluating both prediction quality and computational efficiency on the FinancialPhraseBank dataset.

Contributions

The main contributions of this study are as follows:

- Efficient Model Re-Evaluation: We re-examine financial sentiment classification using a compact LLM (270M parameters) and demonstrate that full fine-tuning of very large models is unnecessary for this task.
- Performance Improvement over a 7B Baseline: We show that a LoRA-tuned Gemma-3 270M model surpasses the reported Gemma-7B financial-sentiment baseline from prior work, despite being drastically smaller and computationally more efficient.
- Comprehensive Metric and Error Analysis: Using accuracy, macro-F1, ROC-AUC, calibration curves, and confusion matrices, we provide a detailed performance comparison that reveals where compact LoRA-based models are competitive and where specialized encoder baselines remain stronger.
- Resource-Efficiency Evaluation: By monitoring GPU memory usage, trainable parameter count, and compute requirements, we quantify the cost reductions achieved through LoRA and highlight its suitability for practitioners with limited computational resources.

Overall, this study demonstrates that parameter-efficient training on compact LLMs is viable and highly competitive for financial sentiment tasks, especially when compared to resource-heavy fine-

tuning of multi-billion-parameter models. These findings have meaningful implications for financial institutions and practitioners seeking efficient, deployable NLP systems without sacrificing predictive accuracy.

Experimental Setup

This section describes the dataset, model configuration, prompting strategy, tokenization pipeline, and training environment used to evaluate parameter-efficient fine-tuning on the Gemma-3 270M model. The goal is to define all experimental components clearly so that subsequent methodological decisions and performance comparisons can be interpreted in a controlled and reproducible manner.

Dataset

All experiments are conducted on the FinancialPhraseBank (FPB) corpus introduced by Malo et al. We use the all-agree subset, which contains financial news snippets annotated as negative, neutral, or positive by multiple domain experts who unanimously agreed on the label. This subset is widely used in financial sentiment research because its strict agreement criterion yields exceptionally high-quality ground-truth labels.

The rerun uses 1,629 training examples, 182 validation examples, and 453 held-out test examples. After filtering malformed or unusable rows, the effective training set contains 1,627 examples. The dataset consists of short, information-dense statements typical of financial journalism, making it a strong benchmark for evaluating both domain robustness and sentiment sensitivity in compact language models.

Model Specification

The base model for all experiments is Gemma-3 (270M parameters), a lightweight decoder-only transformer designed for instruction-following tasks. Despite its compact size, the model retains the architectural features of modern LLMs, including multi-head self-attention with rotary positional embeddings, feed-forward MLP sublayers, and pre-norm residual connections [Vas17].

The attention mechanism used in each transformer block is given by:

$$\text{Attention}(Q,K,V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right). \quad (1)$$

The feed-forward network is defined as:

$$\text{FFN}(x) = W_2\sigma(W_1x + b_1) + b_2. \quad (2)$$

These components provide the representational capacity needed for domain-specific adaptation while maintaining GPU-friendly memory consumption.

Prompting Strategy

Sentiment classification is framed as an instruction-following generation problem. Each training example is wrapped in a fixed prompt template:

```
You are a financial sentiment classifier.
Text: "<text>"
Answer with one word: negative, neutral, or positive.
Sentiment:
```

The model is trained to autoregressively produce exactly one of the three sentiment labels. This template mirrors prior Gemma prompting conventions and ensures comparability with existing baselines, including Gemma-7B.

Tokenization and Label Encoding

All text is processed using the official Gemma tokenizer. Because Gemma does not include a dedicated padding token, the end-of-sequence (EOS) token is reused for padding.

The full prompt–target sequence is tokenized, but only the label tokens contribute to the supervised loss. All prompt tokens are masked with -100 to prevent gradient flow into instruction components. Sentiment labels follow a consistent integer mapping:

0 → negative;
1 → neutral;
2 → positive.

This setup guarantees deterministic and consistent label alignment across the training and evaluation stages.

Training Environment

All fine-tuning runs are performed on a single 16 GB Nvidia GPU using mixed precision (FP16 or BF16 depending on hardware). Training proceeds for three epochs with the following hyperparameters:

Learning rate: 2×10^{-4};
Optimizer: AdamW;
Warmup ratio: 0.03;
Effective batch size: 8 (via gradient accumulation).

During training, we log training loss, validation performance, and GPU memory consumption for later analysis in the results section.

METHODOLOGY

This section details the parameter-efficient fine-tuning strategy, the training objective, and the evaluation protocol used to assess downstream performance, calibration, and reliability.

Parameter-Efficient Fine-Tuning

To adapt the Gemma-3 model without modifying its full parameter set, we employ Low-Rank Adaptation (LoRA) [Hu22]. Instead of updating the original weight matrix W, LoRA introduces a low-rank decomposition:

$$\Delta W = BA \tag{3}$$

where A in $R^{(r\times d)}$ and B in $R^{(r\times d)}$ with $r \ll d$. The effective weight during training becomes:

$$W' = W + \alpha\, BA. \tag{4}$$

LoRA Configuration:

Rank $r = 16$;
Scaling factor alpha = 32;
No dropout;
Target modules: W_Q, W_K, W_V, W_O.

This configuration significantly reduces the number of trainable parameters and enables efficient training on a single 16 GB Nvidia GPU while maintaining strong expressiveness in the attention mechanism.

Training Objective

We train the model using supervised causal language modeling. The model receives the complete instruction prompt and is required to generate only the target sentiment token. All non-target tokens are masked in the loss function, ensuring that optimization focuses exclusively on predicting the sentiment label. This approach is consistent with standard instruction-tuning procedures and simplifies extraction of calibrated probabilities during evaluation.

Evaluation Protocol

Model performance is assessed on a held-out test split using a comprehensive set of metrics:

Overall accuracy;
Macro-F1 and class-specific F1;
Per-class precision and recall;
Confusion matrices for error analysis;
Confidence distribution curves;
Expected Calibration Error (ECE);
Manual inspection of high-confidence misclassifications.

These metrics jointly capture both classification accuracy and confidence reliability, which are critical properties for financial sentiment applications.

Baseline Comparison

For quantitative comparison, we use two levels of external context. First, we compare directly against the Gemma-7B result reported by Mo et al. [Mo24], which is the closest large-model FinancialPhraseBank baseline in model family and task framing. Second, we compare against published FinancialPhraseBank results from FinBERT-style encoder systems [Ara19, Ats24, Sun25]. These comparisons are cross-paper rather than perfectly controlled: train/test splits, agreement subsets, augmentation strategies, and reported F1 variants differ across studies. They are therefore used as a performance-efficiency reference frame rather than as a strict leaderboard.

Results and Discussion

This section presents the empirical results of fine-tuning the 270M-parameter Gemma-3 model using LoRA and compares its performance with published FinancialPhraseBank baselines. We evaluate performance on the FinancialPhraseBank (all-agree) test set and analyze calibration, error behavior, computational efficiency, and cross-paper positioning. Results demonstrate that the compact LoRA-tuned Gemma-3 model clearly improves over the reported Gemma-7B baseline from Mo et al. [Mo24], while approaching the performance range of stronger domain-specialized FinBERT-style systems.

Overall Classification Performance

The LoRA-tuned Gemma-3 model achieves an overall accuracy of 0.9536 on the FinancialPhraseBank (all-agree) test set. The macro-averaged metrics provide a balanced, class-agnostic view of performance, yielding a macro precision of 0.9404, a macro recall of 0.9328, and a macro F1-score of 0.9362. The one-vs-rest macro ROC-AUC is 0.997, indicating strong separability even for the smaller negative and positive classes.

These values show that the model is not only accurate in aggregate but also performs consistently across the three classes. The small gap between macro precision and macro recall (0.0076) suggests that the model is neither overly conservative nor overly eager in assigning class labels. This balance is especially notable in financial sentiment analysis, where class boundaries can be linguistically subtle.

Weighted metrics further reinforce this conclusion. The weighted precision, recall, and F1-score are 0.9536, 0.9536, and 0.9533, respectively. Because these metrics incorporate class frequency, they reveal that performance on the majority neutral class does not mask deficiencies in the minority classes. The near-equivalence between weighted recall and overall accuracy also indicates that misclassifications are not driven by one dominant failure mode.

Taken together, these findings demonstrate that high-quality sentiment classification can be achieved with a compact model architecture when combined with parameter-efficient training. The performance is particularly noteworthy given that Gemma-3 (270M) operates at less than 4% of the parameter count of the 7B baseline.

Table 1

Improvement from zero-shot Gemma 3 270M to LoRA-tuned Gemma 3 270M on the FinancialPhraseBank all-agree test set. Lower ECE is better

Setup	Accuracy	Macro-F1	Weighted-F1	ECE
Base Gemma 3 270M (zero-shot)	0.5055	0.4027	0.4924	0.0760
LoRA-tuned Gemma 3 270M	0.9536	0.9362	0.9533	0.0178
Absolute gain	+ 0.4481	+ 0.5336	+ 0.4609	– 0.0582

The improvement table highlights the main practical effect of LoRA fine-tuning: accuracy rises by 44.81 percentage points, while macro-F1 increases by 0.5336, corresponding to a 132.5% relative gain over the zero-shot macro-F1 score. Calibration also improves substantially, with ECE decreasing from 0.0760 to 0.0178.

Cross-Paper Comparison with Published FPB Results

To contextualize these results, Table 2 compares the LoRA-tuned Gemma 3 270M model with external FinancialPhraseBank results. The Mo et al. row is taken from the supplied paper PDF and its arXiv version; the remaining rows are drawn from published FinancialPhraseBank studies that report results on the 100%-agreement or all-agree setting.

Table 2

Cross-paper comparison on FinancialPhraseBank all-agree or 100%-agreement settings. Reported F1 variants and split protocols differ by paper, so the table should be read as contextual positioning rather than a strict apples-to-apples leaderboard

Model / Setup	Params	Tuning	Accuracy	F1
Gemma 3 270M (this work)	0.274B	LoRA	0.9536	0.9362
Gemma-7B, Mo et al. [Mo24]	7B	PEFT/SFT	0.8740	0.8760
FinBERT, Araci [Ara19]	0.11B	domain PT + FT	0.9700	0.9500
finbert-lc, Atsiwo [Ats24]	0.11B	synthetic FT	0.9700	0.9600
EnhancedFinSentiBERT, Sun et al. [Sun25]	0.11B	dictionary + neutral features	N/R	0.9800

Relative to the larger Gemma-7B baseline, the 270M LoRA model improves accuracy by 7.96 percentage points (0.9536 vs. 0.8740) and F1 by 0.0602 (0.9362 vs. 0.8760), while using roughly 26 times fewer total parameters. Compared with the strongest FinBERT-style systems, the rerun is slightly behind peak 100%-agreement F1, especially EnhancedFinSentiBERT's reported 0.98 F1. This is expected because those systems use encoder architectures and additional financial-domain adaptation mechanisms tailored specifically to sentiment classification. The comparison therefore positions Gemma 3 270M LoRA as a compact generative-model alternative that is substantially stronger than the published Gemma-7B result, though not the absolute state-of-the-art among specialized encoders.

Per-Class Performance

A deeper inspection of precision, recall, and F1-scores for the individual sentiment classes highlights how the model handles the distinct linguistic characteristics of each category.

Table 3

Per-class precision, recall, F1-score, and support for the LoRA-tuned Gemma 3 270M model (3 s.f.)

Class	Precision	Recall	F1-score	Support
Negative	0.903	0.918	0.911	61
Neutral	0.965	0.986	0.975	278
Positive	0.953	0.895	0.923	114
Macro average	0.940	0.933	0.936	453
Weighted average	0.954	0.954	0.953	453

The neutral class achieves the strongest performance (F1 = 0.975), which is noteworthy because neutral statements in financial news often exhibit the smallest lexical and semantic cues. The very high recall for this class (0.986) indicates that the model rarely misses a neutral instance. This suggests that the LoRA-tuned model has learned a robust representation of domain-specific neutrality, which is often the largest and most ambiguous class.

In contrast, the positive class shows a more asymmetric profile: precision remains high at 0.953, while recall is lower at 0.895. This indicates that positive predictions are usually correct, but a subset of positive statements is still absorbed into the negative or neutral classes. Such behavior is consistent with financial text patterns, where optimistic sentiment tends to be expressed indirectly or hedged, leading to subtler cues.

The negative class displays a balanced profile (precision 0.903, recall 0.918), suggesting that the model has a stable understanding of negative sentiment without excessive over-prediction. Its F1-score of 0.911 reflects strong discrimination despite negative statements being the minority class.

Overall, the per-class distribution reveals that Gemma-3 excels in identifying neutral sentiment with high reliability, while maintaining balanced performance on the two polarity classes. The positive class remains the most difficult class by recall, but the error pattern is structured rather than random, which is desirable for downstream interpretability and calibration.

These findings align with recent work showing that small instruction-tuned models can form sharp decision boundaries when guided by explicit task structure and domain-specific prompting. The results suggest that LoRA fine-tuning effectively tailors the latent space of Gemma-3 to the financial domain without overfitting or collapsing minority-class boundaries.

Confusion Matrix Analysis

Fig. 1 presents the raw and normalized confusion matrices for the test set predictions. The model exhibits strong class separation overall, with 432 correct predictions out of 453 test examples. Most errors still occur near the neutral-polarity boundary, but the rerun also contains a small number of direct polarity reversals, especially positive statements classified as negative.

The neutral class shows the highest stability, with 274 out of 278 samples correctly identified, corresponding to a normalized recall of 0.99. The four neutral errors are all assigned to the positive class, suggesting that the model rarely treats neutral statements as negative but can interpret mildly optimistic corporate wording as positive.

Negative samples are correctly classified in 56 out of 61 cases, corresponding to a normalized recall of 0.92. Four negative samples are assigned to the neutral class and one is assigned to the positive class. This indicates that negative sentiment is usually well separated, although polarity inversion is no longer completely absent in the rerun.

The positive class displays the largest spread of errors: 102 out of 114 positive examples are correctly classified, while six are predicted as neutral and six are predicted as negative. This aligns

with the earlier precision-recall analysis: positive sentiment remains the most difficult category by recall, likely due to the more implicit and hedged nature of positive phrasing in financial discourse.

Overall, the confusion patterns demonstrate that model errors are systematic and linguistically interpretable rather than random. Neutral statements remain highly stable, while the remaining mistakes concentrate in cases where polarity cues are subtle or where positive financial language can be read as weak, conditional, or risk-bearing.

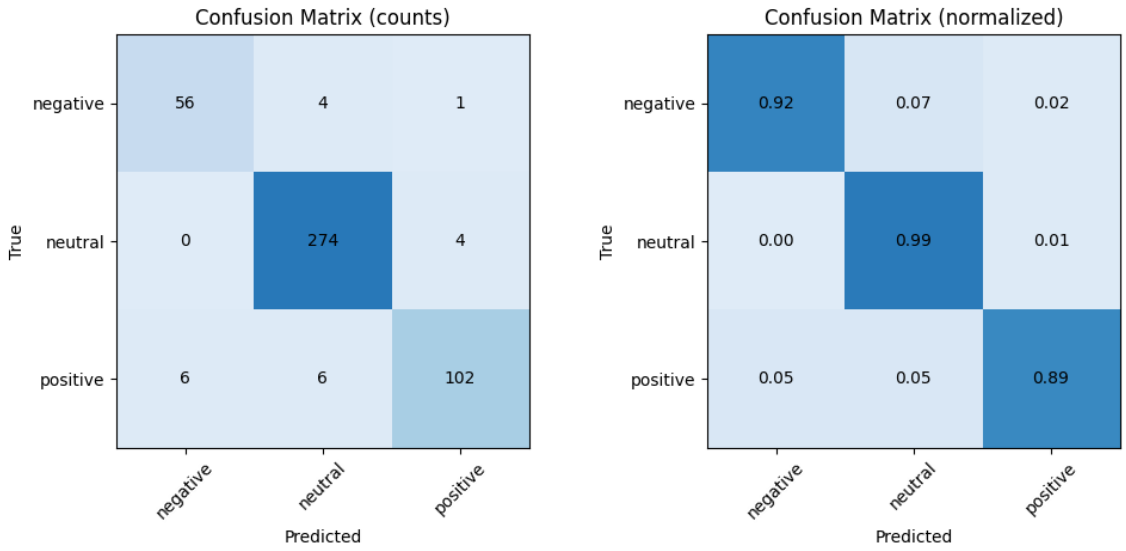

Fig. 1 Confusion matrices for the LoRA fine-tuned Gemma 3 270M: raw counts (left) and row-normalized values (right)

Prediction Confidence and Calibration

We further analyze the model's probability estimates by examining the reliability curve and the Expected Calibration Error (ECE). The reliability curve in Fig. 2 shows that the model is best calibrated at high confidence levels, where most predictions are concentrated. The resulting 10-bin ECE is 0.0178, a substantial improvement over the zero-shot base model's ECE of 0.0760.

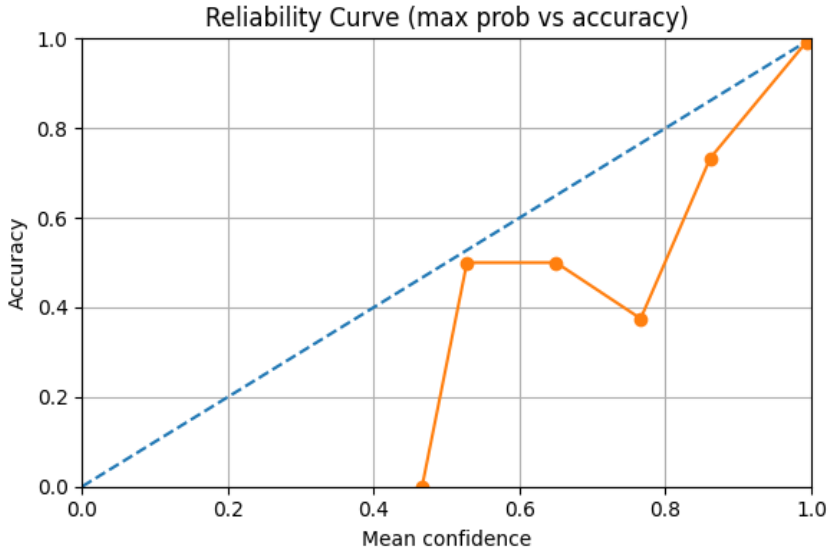

Fig. 2 Reliability curve for the LoRA-tuned Gemma 3 270M model on the FinancialPhraseBank all-agree test set

A clear trend emerges in the higher-confidence region (above approximately 0.85), where empirical accuracy rises sharply and the highest-confidence bin aligns closely with perfect calibration. This behavior is important for financial sentiment classification because high-confidence predictions are the ones most likely to influence downstream decision workflows.

The mid-confidence region displays the largest deviations, including bins where predicted confidence is higher than achieved accuracy. These bins correspond to a small subset of difficult or borderline statements rather than the dominant prediction regime. Such deviations are common in compact models and do not significantly impact the overall ECE because most correct predictions occur at higher confidence.

Overall, the curve indicates useful calibration characteristics: high-confidence predictions are reliably correct, and the remaining deviations are moderate and interpretable. The low ECE score demonstrates that the LoRA-tuned Gemma-3 model produces confidence estimates that are meaningful enough for thresholding, review queues, or other risk-aware financial NLP workflows.

Calibration is an important property in financial decision-making settings, where confidence estimates influence downstream risk assessment. The strong observed calibration highlights the advantage of compact parameter-efficient architectures in providing reliable probability estimates, ensuring that predicted confidence levels correspond closely to empirical correctness.

Computational Efficiency

Beyond predictive performance, an important consideration in financial NLP applications is the computational cost associated with model adaptation. The LoRA-based fine-tuning of the Gemma-3 270M model demonstrates a substantially lighter resource footprint than adapting multi-billion-parameter models.

In our experiments, the parameter-efficient setup trained only the LoRA adapter layers, corresponding to 5,898,240 trainable parameters out of 273,996,416 total parameters. This means that only 2.1527% of the model parameters were updated. The limited parameter update scope reduces both memory requirements and computational load while preserving the expressive capacity of the underlying model.

During training on a 16 GB Nvidia GPU, peak memory usage reached 5,144.91 MiB, approximately 5.03 GiB, well within the constraints of modest hardware typically available in academic and industry environments.

Although we do not directly reimplement the Gemma-7B baseline due to hardware constraints, adapting a seven-billion-parameter model generally involves substantially higher memory and storage requirements even when PEFT is used. This implies a higher cost barrier in terms of hardware availability and deployment complexity.

Overall, the results indicate that adapting a compact 270M-parameter model with LoRA achieves strong performance while remaining feasible on widely accessible hardware. This makes the approach particularly suitable for financial institutions or research groups that require reliable domain adaptation but operate under strict hardware or deployment constraints.

Summary of Findings

Overall, our results demonstrate that a compact 270M-parameter model, when adapted using LoRA, can outperform the reported Gemma-7B FinancialPhraseBank baseline while remaining close to strong specialized encoder systems. The reduced computational burden, strong macro ROC-AUC, and improved calibration make the LoRA-based approach highly suitable for real-world financial applications requiring rapid iteration, low latency, or deployment on constrained hardware.

Conclusion

This study examined the effectiveness of parameter-efficient fine-tuning for financial sentiment analysis by applying LoRA to the 270M-parameter Gemma-3 model and evaluating its performance on the FinancialPhraseBank dataset. Despite its compact size, the adapted model achieved accuracy, macro-F1, and per-class performance that exceed the Gemma-7B results reported in prior work by Mo

et al. [Mo24], while remaining competitive with published FinBERT-style systems on the 100%-agreement setting. These results demonstrate that strong performance on FinancialPhraseBank does not require reliance on multi-billion-parameter architectures. Instead, a smaller model, when paired with a carefully configured LoRA adapter, can provide highly competitive results while avoiding the substantial computational burden associated with large-scale fine-tuning.

Beyond predictive accuracy, the analysis highlighted several practical advantages of the LoRA-based approach. The reduced number of trainable parameters and modest memory requirements make the method feasible on widely accessible hardware, lowering the cost barrier for both experimentation and deployment. Furthermore, the model exhibited stable calibration and well-structured confidence behavior, indicating that improvements in efficiency do not come at the expense of reliability – an important consideration in financial decision-making contexts.

Overall, the findings underscore the value of revisiting established sentiment benchmarks through the lens of resource efficiency. Parameter-efficient methods such as LoRA enable compact open models to approach or exceed larger generative baselines while remaining more practical for real-world applications where hardware, energy consumption, and deployment constraints play a central role. Future work may extend this analysis to larger financial corpora, multi-task settings, or more complex downstream applications, but the present results provide strong evidence that compact PEFT-tuned models represent a compelling alternative for financial sentiment analysis when efficiency and deployability matter alongside raw benchmark accuracy.

REFERENCES

[Ara19] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," arXiv preprint arXiv:1908.10063, 2019. 10.48550/arXiv.1908.10063.

[Ats24] A. Atsiwo, "Financial sentiment analysis: Leveraging actual and synthetic data for supervised fine-tuning," arXiv preprint arXiv:2412.09859, 2024. 10.48550/arXiv.2412.09859.

[Det23] T. Dettmers, A. Pagnoni, et al., "QLoRA: Efficient fine-tuning of quantized LLMs," Advances in Neural Information Processing Systems, vol. 36, pp. 10088-10115, 2023. 10.52202/075280-0441.

[Dev19] J. Devlin, M.-W. Chang, et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171-4186, 2019. 10.48550/arXiv.1810.04805.

[Du24] K. Du, F. Xing, et al., "Financial sentiment analysis: Techniques and applications," ACM Computing Surveys, vol. 56, no. 9, 56, 9, art. 220, 42 pp., 2024. 10.1145/3649451.

[Han24] Z. Han, C. Gao, et al., "Parameter-efficient fine-tuning for large models: A comprehensive survey," arXiv preprint arXiv:2403.14608, 2024. 10.48550/arXiv.2403.14608.

[Hu22] E. J. Hu, Y. Shen, et al., "LoRA: Low-rank adaptation of large language models," ICLR, vol. 1, no. 2, p. 3, 2022. 10.48550/arXiv.2106.09685.

[Liu22] H. Liu, D. Tam, et al., "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," Advances in Neural Information Processing Systems, vol. 35, pp. 1950-1965, 2022. 10.48550/arXiv.2205.05638.

[Mal14] P. Malo, A. Sinha, et al., "Good debt or bad debt: Detecting semantic orientations in economic texts," Journal of the Association for Information Science and Technology, vol. 65, no. 4, pp. 782-796, 2014. 10.1002/asi.23062.

[Mo24] K. Mo, W. Liu, et al., "Fine-tuning Gemma-7B for enhanced sentiment analysis of financial news headlines," in 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI), pp. 130-135, IEEE, 2024. 10.1109/ICETCI61221.2024.10594605.

[Nas25] D. K. Nasiopoulos, K. I. Roumeliotis, et al., "Financial sentiment analysis and classification: A comparative study of fine-tuned deep learning models," International Journal of Financial Studies, vol. 13, no. 2, p. 75, 2025. 10.3390/ijfs13020075.

[Sun25] Y. Sun, H. Yuan, and F. Xu, "Financial sentiment analysis for pre-trained language models incorporating dictionary knowledge and neutral features," Natural Language Processing Journal, p. 100148, 2025. 10.1016/j.nlp.2025.100148.

[Tea25] G. Team, A. Kamath, J. Ferret, et al., "Gemma 3 technical report," arXiv preprint arXiv:2503.19786, 2025. 10.48550/arXiv.2503.19786.

[Tou23] H. Touvron, T. Lavril, et al., "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023. 10.48550/arXiv.2302.13971.

[Vas17] A. Vaswani, N. Shazeer, et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, 2017. 10.48550/arXiv.1706.03762.

[Wan25] L. Wang, S. Chen, et al., "Parameter-efficient fine-tuning in large language models: A survey of methodologies," Artificial Intelligence Review, vol. 58, no. 8, p. 227, 2025. 10.1007/s10462-025-11236-4. PYPSSL.

ОБ АВТОРАХ | ABOUT THE AUTHORS

ГАРДАШОВА Латафат Аббас-кызы
Азербайджанский государственный университет нефти и промышленности.
l.qardashova@asoiu.edu.az ORCID: 0000-0003-3227-2521
Проректор по науч. работе. Проф. каф. компьютерной инженерии. Дипл. прикл. математик (Азерб. гос. ун-т). Д-р техн. наук по сист. анализу, управлению и обработке информации (Азерб. гос. академия нефти). Засл. деятель науки Респ. Азербайджан. Иссл. в обл. мягких вычислений, нечеткой математики и систем, систем управления.

ГАСЫМОВ Айдын Эльдениз-оглы
Азербайджанский государственный университет нефти и промышленности.
l.qardashova@asoiu.edu.az ORCID: 0009-0000-3889-2823
Аспирант, кафедра компьютерной инженерии.

GARDASHOVA Latafat Abbas
Azerbaijan State Oil and Industry University.

l.qardashova@asoiu.edu.az ORCID: 0000-0003-3227-2521
Vice-Rector for Scientific Affairs. Prof., Dept. of Computer Engineering. Dipl. applied mathematics (Azerbaijan State Uni.). Dr. of Technical Sciences in system analysis, control and information processing (Azerbaijan State Oil Academy). Honored Scientist of the Republic of Azerbaijan Research: soft computing, fuzzy mathematics and systems, control systems.

GASIMOV Aydin Eldaniz
Azerbaijan State Oil and Industry University,

l.qardashova@asoiu.edu.az ORCID: 0009-0000-3889-2823
PhD student, Dept. of Computer Engineering.

МЕТАДАННЫЕ | METADATA

Заглавие: Оптимизация параметров Gemma-3 для анализа финансовых настроений.

Авторы: Гардашова Л. А., Гасымов А. Э.

Аннотация: Анализ финансовых настроений играет центральную роль в мониторинге рынка и оценке рисков, однако многие современные подходы основаны на полной тонкой настройке больших языковых моделей с высокими вычислительными затратами. В данной работе мы рассматриваем ресурсоэффективную альтернативу, применяя адаптацию низкого ранга (LoRA) к модели Gemma-3 с 270 миллионами параметров и оценивая ее производительность на наборе данных FinancialPhraseBank (все согласны). Несмотря на свой компактный размер, модель, настроенная с помощью LoRA, достигает точности 0,9536, макро-F1-показателя 0,9362 и взвешенного F1-показателя 0,9533, что явно превосходит базовый показатель анализа финансовых настроений Gemma-7B, представленный Mo и др., и остается конкурентоспособной по сравнению с опубликованными системами в стиле FinBERT на том же эталонном наборе данных. Дополнительный анализ показывает стабильную производительность по всем классам настроений, макро-ROC-AUC 0,997, хорошие оценки достоверности и стабильную калибровку. С практической точки зрения, параметрически эффективная конфигурация существенно сокращает использование памяти и количество обучаемых параметров, что позволяет эффективно адаптировать модель к предметной области на широко доступном оборудовании. Эти результаты показывают, что высококачественная классификация финансовых настроений не требует полной тонкой настройки моделей с миллиардами параметров, и что компактные модели в сочетании с параметрически эффективными методами обеспечивают привлекательный баланс между точностью, эффективностью и удобством внедрения.

Ключевые слова: Анализ финансовых настроений; Gemma-3; Адаптация низкого ранга; LoRA; параметрически эффективная тонкая настройка; FinancialPhraseBank..

Язык: Английский.

Статья поступила в редакцию 29 июня 2026 г.

Title: Parameter-efficient fine-tuning of Gemma-3 for financial sentiment analysis.

Authors: Gardashova L. A., Gasimov A. E.

Abstract: Financial sentiment analysis plays a central role in market monitoring and risk assessment, yet many state-of-the-art approaches rely on fully fine-tuning large language models with high computational demands. In this work, we examine a resource-efficient alternative by applying Low-Rank Adaptation (LoRA) to the 270M-parameter Gemma-3 model and evaluating its performance on the FinancialPhraseBank (all-agree) dataset. Despite its compact size, the LoRA-tuned model achieves an accuracy of 0.9536, a macro-F1 score of 0.9362, and a weighted-F1 score of 0.9533, clearly improving over the Gemma-7B financial-sentiment baseline reported by Mo et al. while remaining competitive with published FinBERT-style systems on the same benchmark. Additional analysis shows consistent performance across sentiment classes, a macro ROC-AUC of 0.997, well-behaved confidence estimates, and stable calibration. From a practical standpoint, the parameter-efficient setup substantially reduces memory usage and trainable parameter count, enabling effective domain adaptation on widely accessible hardware. These results demonstrate that high-quality financial sentiment classification does not require full fine-tuning of multi-billion-parameter models, and that compact models paired with parameter-efficient methods offer an attractive balance between accuracy, efficiency, and deployability.

Key words: Financial sentiment analysis; Gemma-3; Low-Rank Adaptation; LoRA; parameter-efficient fine-tuning; FinancialPhraseBank.

Language: English.

The editors received the article on 29 June 2026.