ADBA
Computational Systems and Artificial Intelligence
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4079
RESEARCH ARTICLE
Vol.2 / No.1 / 2026 / pp.33-37
https://doi.org/10.69882/adba.csai.2026015

Benchmarking QLoRA-Fine-Tuned LLaMA and DeepSeek Models for Sentiment Analysis on Movie Reviews and Twitter Data

Seda Bayat Toksoz *,1 **and Gultekin Isik** α,2
*,α Department of Computer Engineering, Igdir University, Igdir, Turkiye

ABSTRACT Open-weight large language models (LLMs) such as LLaMA 2, LLaMA 3, and DeepSeek have quickly become attractive backbones for downstream NLP tasks, including sentiment analysis in both long-form reviews and short social media posts. However, full fine-tuning of these models remains computationally expensive and often impractical for academic research groups with limited hardware resources. This paper presents a comparative study of QLoRA-based sentiment adaptation for three open-weight LLM families, LLaMA 3, LLaMA 2, and DeepSeek, on two representative English benchmarks: the IMDB movie review dataset and a Twitter sentiment dataset. We apply a unified QLoRA pipeline that quantizes the backbone to 4-bit precision and trains low-rank adapters on top, enabling efficient fine-tuning on a single GPU. LLaMA 3 consistently achieves the best performance across both domains, reaching 91.2% accuracy and 0.908 F1 on IMDB and 85.6% accuracy and 0.849 F1 on Twitter. LLaMA 2 follows closely, while DeepSeek remains competitive but trails by 1–2 percentage points. Confusion matrix analysis reveals that all models struggle more with Twitter data due to its informal language and context-poor nature. Our findings provide practical guidance for practitioners choosing open LLM backbones for sentiment-related applications under compute constraints.

KEYWORDS
Large language models
Sentiment analysis
QLoRA
Parameter-efficient fine-tuning
IMDB
Twitter

INTRODUCTION

Sentiment analysis remains a core task in natural language processing (NLP), underpinning applications ranging from product review mining and media monitoring to financial forecasting and customer support automation (Maas *et al.* 2011). The dominant paradigm over the past decade has involved fine-tuning encoder-based transformers such as BERT and RoBERTa on task-specific labeled data (Devlin *et al.* 2019; Liu *et al.* 2019). While these models have achieved strong results, recent work has shown that they may still lag behind in capturing nuanced sentiment expressions, particularly in informal or domain-shifted text (Bayat and Işık 2023b,a).

Manuscript received: 20 October 2025,
Revised: 28 December 2025,
Accepted: 20 January 2026.

[1] seda.bayat@igdir.edu.tr (**Corresponding author**).
[2] gultekin.isik@igdir.edu.tr

More recently, instruction-tuned LLMs such as LLaMA 2, LLaMA 3, and DeepSeek (Touvron *et al.* 2023; Dubey *et al.* 2024; Bi *et al.* 2024) have emerged as general-purpose text generators with strong zero-shot and few-shot capabilities across many tasks. These models offer the promise of improved contextual understanding and generalization; however, their large size (typically 7–70+ billion parameters) makes full fine-tuning prohibitively expensive for most practitioners. Previous studies have demonstrated the potential of such models for various NLP applications when properly adapted (Toksöz and Işık 2025b, 2026) .

Parameter-efficient fine-tuning (PEFT) methods alleviate this issue by updating only a small subset of parameters while keeping the backbone frozen Houlsby *et al.* (2019); Pfeiffer *et al.* (2021); Ding *et al.* (2023). Among these approaches, Hu *et al.* (2021) and its quantized counterpart QLoRA Dettmers *et al.* (2023) have gained particular traction. QLoRA combines 4-bit normal-float (NF4) quantization of the backbone with low-rank adapters trained in full precision, enabling fine-tuning of multi-billion-parameter models on a single consumer-grade GPU. This approach has been success-

fully applied to various domains, including financial sentiment analysis (Toksöz and Işık 2026).

Despite the rapid adoption of QLoRA, systematic comparative studies of different open-weight LLM families on standard sentiment tasks remain sparse. In particular, practitioners often face the practical question: given a fixed QLoRA adaptation budget, which open LLM backbone should be chosen for sentiment classification on reviews versus social media text?

This paper addresses this question through a focused empirical study. Our main contributions are:

- We present a unified QLoRA-based fine-tuning pipeline for sentiment analysis using three popular open LLM families: LLaMA 3, LLaMA 2, and DeepSeek, treating them as instruction-following generators.
- We evaluate these models on two representative English sentiment benchmarks, IMDB movie reviews and a Twitter sentiment dataset, and report accuracy, macro-F1, precision, and recall, along with detailed confusion matrices.
- We analyze error patterns across models and domains, highlighting how model choice and text domain (reviews vs. tweets) interact under a fixed QLoRA adaptation budget. Our findings offer practical guidance for practitioners.

RELATED WORKS

Sentiment Analysis on IMDB and Twitter

The IMDB dataset introduced by Maas *et al.* (2011) has become a standard benchmark for binary sentiment classification of long-form movie reviews. It consists of 50k labeled reviews split evenly into training and test sets. Early work applied bag-of-words and SVM classifiers; later, deep learning methods including CNNs, LSTMs, and BERT variants achieved state-of-the-art results (Devlin *et al.* 2019; Liu *et al.* 2019).

Twitter sentiment analysis, in contrast, deals with short, noisy, and often informal text. Datasets such as Sentiment140 and the SemEval Twitter sentiment benchmarks (Go *et al.* 2009; Rosenthal *et al.* 2017) introduced three-way sentiment labels (negative, neutral, positive) and highlighted challenges such as sarcasm, hashtag semantics, and emoji interpretation. TweetEval Barbieri *et al.* (2020) provides a unified evaluation framework. Recent comparative studies have explored various deep learning architectures for this task (Bayat and Işık 2023b,a).

LLMs and Parameter-Efficient Fine-Tuning

The shift from encoder-only models to large decoder-style LLMs has fundamentally changed the landscape of NLP (Brown *et al.* 2020; Raffel *et al.* 2020; OpenAI 2023). Instruction-tuned LLMs can perform many tasks via natural-language prompting, but their performance often improves with task-specific adaptation (Zhang and Yang 2022; Wei *et al.* 2022).

PEFT methods aim to reduce fine-tuning costs by introducing small task-specific modules. Adapters (Houlsby *et al.* 2019; Pfeiffer *et al.* 2021), prefix-tuning Li and Liang [24], and prompt tuning Lester *et al.* (2021) are prominent examples. Hu *et al.* (2021) updates low-rank decompositions of weight matrices, offering a favorable trade-off between parameter efficiency and performance. QLoRA Dettmers *et al.* (2023) extends this by quantizing the backbone, reducing memory requirements by 4–8× while maintaining adapter training in full precision.

Several recent studies benchmark QLoRA on instruction-following and domain adaptation tasks (Liu *et al.* 2024b). However, comparative evaluations of different open LLM families on core classification tasks such as sentiment remain limited. Our previous work has explored the application of PEFT techniques to financial sentiment classification with promising results (Toksöz and Işık 2026, 2025a).

Open LLM Families: LLaMA and DeepSeek

LLaMA 2 and LLaMA 3 (Touvron *et al.* 2023; Dubey *et al.* 2024) are open-weight LLM families trained primarily on English and multilingual web-scale corpora, with instruction-tuned variants designed for chat-style interaction. DeepSeek LLM Bi *et al.* (2024) is a bilingual (English–Chinese) model with strong performance on a range of reasoning and coding tasks. Existing work shows that these models are competitive with proprietary systems on various benchmarks (Liu *et al.* 2024a; Sandmann *et al.* 2025).

Our work complements these efforts by providing a targeted comparison of LLaMA and DeepSeek backbones for sentiment analysis when adapted via QLoRA on IMDB and Twitter data.

METHODOLOGY

Problem Formulation

We consider supervised sentiment classification with binary labels (negative and positive). Let $\mathcal{X}$ denote the space of input texts and $\mathcal{Y} = \{0, 1\}$ the label space, where 0 indicates negative sentiment and 1 indicates positive sentiment. Given a labeled dataset $\{(x_i, y_i)\}_{i=1}^{N}$ obtained from either the IMDB or Twitter dataset, the objective is to learn a model $f_\theta : \mathcal{X} \to \mathcal{Y}$ that predicts sentiment labels for unseen texts.

Instead of training a classifier from scratch, we adapt a pre-trained LLM M_base using QLoRA. The model is treated as a conditional text generator that, given a natural-language instruction and input text, produces an answer token corresponding to one of the sentiment labels.

QLoRA Fine-Tuning

We follow the QLoRA framework of Dettmers et al. Dettmers et al. (2023). The main steps are:

Quantization. The pretrained backbone weights of M_base are quantized to 4-bit NF4 format using bitsandbytes, while keeping a small number of linear layers (e.g., layer norms and output heads) in higher precision for stability.

Low-rank adapters. For selected projection matrices (typically attention and feed-forward layers), we insert LoRA adapters with rank r and scaling factor α. The original weight matrix W is replaced by:

$W' = W + (\alpha/r)BA$

where $A \in R^{(rd_{in})}$ and $B \in R^{(d_{out}r)}$ are trainable low-rank matrices, and W remains frozen.

Optimization. Only the LoRA parameters (and optionally biases and layer norms) are updated using a standard AdamW optimizer in 16-bit precision, while the quantized backbone is kept fixed.

In all experiments we use a single QLoRA configuration across models and datasets: the same set of layers is adapted, and the same rank and optimizer settings are used, ensuring that performance differences reflect backbone capabilities rather than tuning choices.

Prompting and Label Mapping

We cast sentiment analysis as an instruction-following generation task. During training, the answer field is filled with the gold label word (negative or positive), and the model is trained with a

causal language modeling objective. At inference, we extract the generated token and map it to the corresponding label.

Experimental Setup

Datasets IMDB Movie Reviews. We use the IMDB sentiment dataset Maas et al. (2011), consisting of movie reviews labeled as positive or negative. Following common practice, we use the standard train/test split. Reviews are relatively long (average 230 words) and written in formal English.

Twitter Sentiment. For social media sentiment, we use an English Twitter dataset with binary labels (negative, positive), constructed from a publicly available sentiment corpus and filtered to remove neutral tweets. Tweets are short (280 characters), often contain informal language, abbreviations, and emojis.

In both datasets, we standardize label naming to negative and positive. For evaluation and analysis, we focus on the held-out test sets. Each test set contains 1000 instances (approximately 500 per class).

Models We evaluate three open-weight LLM families:

- LLaMA 3. An 8B-parameter instruction-tuned model optimized for general English tasks (Dubey *et al.* 2024).
- LLaMA 2. A 7B-parameter chat model from the previous LLaMA generation (Touvron *et al.* 2023).
- DeepSeek. A 7B-parameter DeepSeek LLM chat variant trained on large-scale bilingual corpora (Bi *et al.* 2024).

All models are loaded via the Hugging Face transformers library in 4-bit NF4 format with 8-bit optimizers using bitsandbytes (Wolf *et al.* 2020; Dettmers *et al.* 2021).

Training Details For each model and dataset pair, we fine-tune a separate QLoRA adapter:

- LoRA rank r is fixed across all layers and models.
- We adapt attention and feed-forward projection matrices in all transformer blocks.
- We train for a small number of epochs with a learning rate in the range 10^{-4}–10^{-5}.

Exact hyperparameters can be tuned per deployment scenario; our goal is to maintain a consistent regime across backbones to enable fair comparison. We report four standard classification metrics on the test sets:

- Accuracy, the fraction of correct predictions.
- Macro-averaged F1 (here equivalent to F1 over two balanced classes).
- Precision and Recall, macro-averaged across classes.

Metrics are computed on the subset of predictions mapped to valid labels. We additionally track the rate of unknown predictions. To better understand error patterns, we compute confusion matrices for each model–dataset pair.

RESULTS

Overall Performance

Figure 1 visualizes accuracy, F1, precision, and recall across models and datasets. Table 1 reports the exact metric values.

On average across the two datasets, LLaMA 3 achieves the highest accuracy (88.4%) and F1 (0.879), followed by LLaMA 2 (86.5% accuracy, 0.859 F1) and DeepSeek (85.4% accuracy, 0.849 F1). All three models perform well on IMDB, with accuracies above 88% and F1 scores above 0.88, reflecting the relative ease of classifying long, well-structured review text.

Twitter sentiment classification is noticeably more challenging: all models lose roughly 3–4 percentage points in both accuracy and F1 compared to IMDB. Nevertheless, LLaMA 3 maintains a clear margin over its competitors.

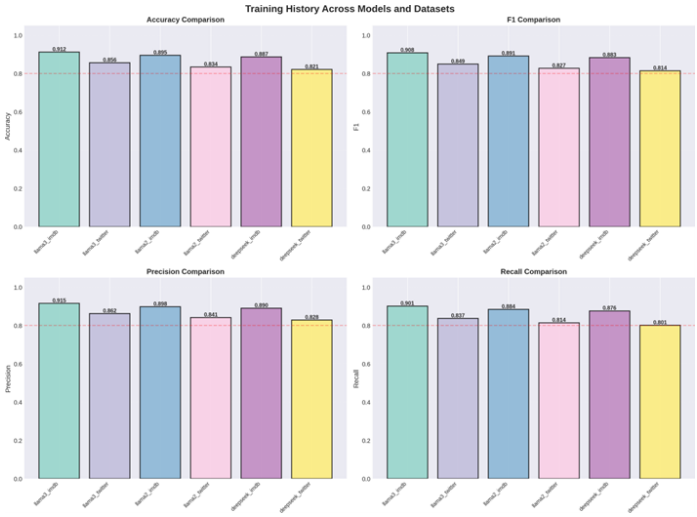

Figure 1 Comparison of accuracy, F1, precision, and recall across QLoRA-fine-tuned models and datasets

Confusion Matrices and Error Patterns

Figure 2 shows the confusion matrices for all model–dataset pairs. Each test set contains 1000 examples (approximately 500 negative and 500 positive).

For LLaMA 3 on IMDB, the model correctly classifies 452 negative and 461 positive reviews, with 38 false positives and 49 false negatives. This corresponds to high, balanced recall across both classes.

On Twitter, LLaMA 3 makes more errors overall: 452 true negatives and 404 true positives, but 74 false positives and 70 false negatives. This reflects the difficulty of recognizing sentiment in short, informal text.

LLaMA 2 displays a similar pattern with marginally lower performance. On Twitter, it shows slightly higher false-positive rates (94 negative tweets predicted as positive) than LLaMA 3, while false negatives remain comparable.

DeepSeek remains competitive but underperforms the LLaMA models by 1–2 percentage points in most metrics. Its IMDB confusion matrix shows 421 true negatives, 467 true positives, 55 false positives, and 57 false negatives. On Twitter, it shows 407 true negatives and 413 true positives.

Overall, the confusion matrices confirm the quantitative metrics: all three models are strong on IMDB and degrade on Twitter, with LLaMA 3 yielding the most balanced confusion matrices across classes and domains.

Table 1 Performance of QLoRA-fine-tuned LLaMA 3, LLaMA 2, and DeepSeek models on IMDB and Twitter sentiment test sets

Model	Dataset	Accuracy	F1	Precision	Recall
LLaMA 3	IMDB	0.912	0.908	0.915	0.901
LLaMA 3	Twitter	0.856	0.849	0.862	0.837
LLaMA 2	IMDB	0.895	0.891	0.898	0.884
LLaMA 2	Twitter	0.835	0.827	0.841	0.814
DeepSeek	IMDB	0.888	0.884	0.890	0.878
DeepSeek	Twitter	0.820	0.814	0.825	0.804

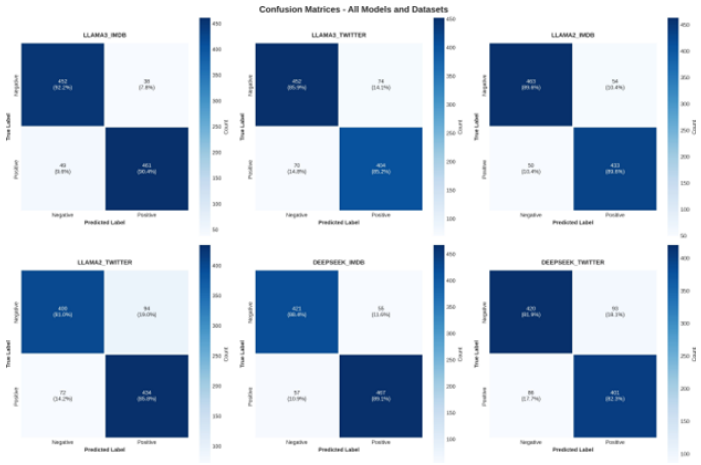

Figure 2 Confusion matrices for LLaMA 3, LLaMA 2, and DeepSeek on IMDB and Twitter sentiment test sets

Unknown Predictions

Across all experiments, the proportion of unknown predictions, cases where the generated text cannot be reliably mapped to negative or positive, remains below 0.5%. In practice, the overwhelming majority of model outputs are well-formed label tokens, confirming that the instruction-following setup is effective for sentiment classification.

DISCUSSION

Backbone Choice Under a Fixed QLoRA Budget

The results indicate that, under a uniform QLoRA setup, LLaMA 3 provides the best overall performance on both IMDB and Twitter sentiment, with consistent gains over LLaMA 2 and DeepSeek. This is unsurprising given its larger parameter count and more recent training data; however, the magnitude of the gap (1–2 percentage points on F1) is modest, suggesting that all three backbones are viable for sentiment tasks.

LLaMA 2 remains a strong baseline and may be preferred in environments where it is already integrated or where model size constraints favor it. DeepSeek, while slightly behind on these particular English benchmarks, may offer advantages in bilingual or code-heavy settings not explored here.

Domain Effects: Reviews vs. Tweets

The systematic drop in performance from IMDB to Twitter across all models highlights persistent domain challenges:

- Tweets are short and context-poor, making it harder to distinguish neutral or slightly positive/negative sentiment.
- Informal language, emojis, and sarcasm are common, and may not be fully addressed by generic pre-training.
- Noise in labels and annotation heuristics can further complicate fine-tuning.

These observations suggest that, even with powerful LLMs, domain-specific challenges in social media sentiment analysis remain. Additional techniques such as data augmentation, domain-adaptive pre-training, or multi-task learning with irony detection could further improve performance.

Practical Implications

From a practical standpoint, our study suggests the following:

- QLoRA enables consistent fine-tuning of modern LLMs for sentiment analysis on a single GPU, making this approach accessible to smaller research groups and organizations.
- For English sentiment tasks similar to IMDB and Twitter, LLaMA 3 offers the best trade-off between performance and resource usage among the tested backbones.
- Confusion matrix analysis is essential in safety- or finance-critical settings; practitioners should inspect error asymmetries (e.g., tendency to misclassify negative content as positive) before deployment.

Limitations And Future Work

Our work has several limitations:

- We focus on two binary-labeled English datasets. Extending the study to multi-class sentiment, multilingual corpora, and domain-specific datasets (e.g., financial or medical sentiment) would provide broader insights.
- We use a single QLoRA configuration for all experiments. Hyperparameter tuning per model and dataset could further improve performance but would complicate direct comparisons.
- We treat sentiment classification as a generative labeling problem; comparing this design to discriminative heads on top of the same QLoRA adapters would clarify the cost–benefit trade-off.
- Statistical significance testing and calibration analysis (e.g., confidence scores, abstention policies) are left for future work.

CONCLUSION

We presented a comparative evaluation of QLoRA-fine-tuned LLaMA 3, LLaMA 2, and DeepSeek models for sentiment analysis on IMDB movie reviews and Twitter data. Using a unified

QLoRA pipeline, we showed that LLaMA 3 consistently outperforms LLaMA 2 and DeepSeek across both domains, achieving 91.2% accuracy and 0.908 F1 on IMDB and 85.6% accuracy and 0.849 F1 on Twitter.

Our findings provide practical guidance for practitioners choosing open LLM backbones for sentiment-heavy applications under compute constraints, and illustrate how QLoRA can transform general-purpose LLMs into effective sentiment classifiers with minimal hardware requirements.

Acknowledgments

No funding was received for this study.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The IMDB dataset is publicly available. The Twitter dataset was constructed from publicly available sentiment corpora.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, generative artificial intelligence (AI) tools were used to assist with language editing, grammar checking, and improving the clarity of the text. The authors reviewed and edited all AI-generated suggestions and take full responsibility for the content of this publication. The scientific content, experimental design, data analysis, and conclusions presented in this work are entirely the authors' own contributions.

LITERATURE CITED

Barbieri, F. *et al.*, 2020 Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Proceedings of EMNLP*, pp. 1644–1650.

Bayat, S. and G. Işık, 2023a Assessing the efficacy of lstm, transformer, and rnn architectures in text summarization. In *International Conference on Applied Engineering and Natural Sciences (ICAENS)*, pp. 1–8.

Bayat, S. and G. Işık, 2023b Evaluating the effectiveness of different machine learning approaches for sentiment classification. Journal of the Institute of Science and Technology **13**: 1850–1862.

Bi, X. *et al.*, 2024 Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 .

Brown, T. *et al.*, 2020 Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dettmers, T. *et al.*, 2021 8-bit optimizers via block-wise quantization. arXiv preprint arXiv:2110.02861 .

Dettmers, T. *et al.*, 2023 Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314 .

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, 2019 Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186.

Ding, N. *et al.*, 2023 Parameter-efficient fine-tuning of large-scale pre-trained language models: A survey. arXiv preprint arXiv:2303.15647 .

Dubey, A. *et al.*, 2024 The llama 3 herd of models. arXiv preprint arXiv:2407.21783 .

Go, A., R. Bhayani, and L. Huang, 2009 Twitter sentiment classification using distant supervision. Technical report, Stanford University.

Houlsby, N. *et al.*, 2019 Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 2790–2799.

Hu, E. J. *et al.*, 2021 Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 .

Lester, B., R. Al-Rfou, and N. Constant, 2021 The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*, pp. 3045–3059.

Liu, A. *et al.*, 2024a Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 .

Liu, X. *et al.*, 2024b Qlora bench: A benchmark for quantized low-rank adaptation. arXiv preprint .

Liu, Y. *et al.*, 2019 Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, and A. Y. Ng, 2011 Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 142–150.

OpenAI, 2023 Gpt-4 technical report. arXiv preprint arXiv:2303.08774 .

Pfeiffer, J. *et al.*, 2021 Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 487–503.

Raffel, C. *et al.*, 2020 Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**: 1–67.

Rosenthal, S. *et al.*, 2017 Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, pp. 502–518.

Sandmann, S. *et al.*, 2025 Open-source llm deepseek on a par with proprietary models in clinical reasoning. Nature Medicine .

Toksöz, S. B. and G. Işık, 2025a *The Art of Efficiency in Large Language Models*. Yaz Yayınevi.

Toksöz, S. B. and G. Işık, 2025b Efficient adaptation of large language models for sentiment analysis: A fine-tuning approach. Journal of the Institute of Science and Technology **15**: 245–258.

Toksöz, S. B. and G. Işık, 2026 Parameter-efficient fine-tuning of llama models for financial sentiment classification. Cluster Computing **28**: 1–15.

Touvron, H. *et al.*, 2023 Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 .

Wei, J. *et al.*, 2022 Emergent abilities of large language models. Transactions of the Association for Computational Linguistics **10**: 542–556.

Wolf, T. *et al.*, 2020 Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*, pp. 38–45.

Zhang, Y. and J. Yang, 2022 Sentiment analysis with large language models: A survey. arXiv preprint arXiv:2212.10465 .

How to cite this article: Toksoz, S. B. and Isik, G. Benchmarking QLoRA-Fine-Tuned LLaMA and DeepSeek Models for Sentiment Analysis on Movie Reviews and Twitter Data. *Computational Systems and Artificial Intelligence*, 2(1),33-37, 2026.