Artificial Intelligence in Applied Sciences
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4060
RESEARCH ARTICLE
Vol.2 / No.1 / 2026 / pp.27-31
https://doi.org/10.69882/adba.ai.2026014

A Comparative Evaluation of QLoRA and AdaLoRA for Parameter-Efficient Fine-Tuning of Large Language Models on Medical Textbook Question Answering

Seda Bayat Toksoz ^{*,1} **and Gultekin Isik** $^{\alpha,2}$
$^{*,\alpha}$Department of Computer Engineering, Igdir University, Igdir, Turkiye.

ABSTRACT

Parameter-efficient fine-tuning methods have emerged as practical solutions for adapting large language models to specialized domains while minimizing computational overhead. This study presents a systematic comparison of two prominent approaches, QLoRA and AdaLoRA, for fine-tuning instruction-tuned language models on medical textbook question answering. We evaluated both methods using two backbone architectures, Llama-3-8B-Instruct and Qwen2-7B-Instruct, on a dataset comprising 6,500 question-answer pairs derived from 13 authoritative medical textbooks spanning diverse clinical and biomedical disciplines. Our experiments demonstrate that QLoRA consistently outperforms AdaLoRA under single-epoch training conditions, achieving validation perplexity values of 1.085 and 1.086 for Llama-3 and Qwen2, respectively, compared to AdaLoRA's 1.125 and 1.169. These results correspond to relative validation loss reductions of 30.8% for Llama-3 and 47.5% for Qwen2 when using QLoRA over AdaLoRA. Both methods maintained comparable trainable parameter counts, approximately 167 million for Llama-3 and 161 million for Qwen2, representing roughly 3.5% of total model parameters. Our findings indicate that QLoRA provides more stable convergence behavior within limited training budgets, while AdaLoRA's adaptive rank allocation mechanism may require extended training schedules to realize its theoretical advantages. These results offer practical guidance for deploying parameter-efficient fine-tuning in medical natural language processing applications where computational resources are constrained.

KEYWORDS
Parameter-efficient fine-tuning
QLoRA
AdaLoRA
Large language models
Medical question answering
Low-rank adaptation

INTRODUCTION

Large language models have demonstrated remarkable capabilities across diverse natural language processing tasks, establishing new performance benchmarks in text generation, comprehension, and reasoning (Brown *et al.* 2020; Chowdhery *et al.* 2023). The medical domain presents a particularly compelling application area, where accurate information retrieval and question answering can support clinical decision-making, medical education, and patient care (Thirunavukarasu *et al.* 2023; Singhal *et al.* 2023). However, the computational demands of training and deploying these models at scale present significant barriers, especially for healthcare institutions operating under resource constraints (Karabacak and Margetis 2023).

The standard approach of full fine-tuning, which updates all model parameters during training, becomes prohibitively expensive as model sizes grow into the billions of parameters. A model containing 7 to 8 billion parameters requires substantial GPU memory merely for inference, and fine-tuning such models with full precision necessitates specialized hardware configurations that remain inaccessible to many research groups and clinical settings. This computational barrier has motivated the development of parameter-efficient fine-tuning techniques, which seek to adapt pretrained models to downstream tasks while modifying only a

Manuscript received: 24 October 2025,
Revised: 28 December 2025,
Accepted: 20 January 2026.

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

small fraction of the total parameters (Lialin *et al.* 2023; Han *et al.* 2024). Similar approaches have been successfully applied in other domains, including financial sentiment analysis using parameter-efficient methods (Bayat Toksoz and Isik 2025).

Low-Rank Adaptation, or LoRA, introduced by Hu *et al.* (2022), represents a seminal contribution to this field. The core insight underlying LoRA is that the weight updates during fine-tuning can be effectively approximated by low-rank matrices, thereby dramatically reducing the number of trainable parameters. Rather than updating the full weight matrix W, LoRA decomposes the update into two smaller matrices: DeltaW = BA, where B and A are low-rank matrices with the rank r chosen to be much smaller than either dimension. This formulation maintains the expressiveness needed for task adaptation while reducing memory requirements by orders of magnitude.

Building upon LoRA, Dettmers *et al.* (2024) proposed QLoRA, which combines low-rank adaptation with 4-bit quantization of the base model weights. QLoRA introduces several technical innovations, including a novel 4-bit NormalFloat data type optimized for normally distributed weights, double quantization to reduce memory overhead from quantization constants, and paged optimizers to handle memory spikes during training. These contributions enable fine-tuning of models with tens of billions of parameters on single consumer-grade GPUs, democratizing access to large-scale language model adaptation.

An alternative approach, AdaLoRA, proposed by Zhang *et al.* (2023), extends the LoRA framework by introducing adaptive budget allocation. Rather than assigning fixed ranks to all weight matrices, AdaLoRA parameterizes the weight updates using singular value decomposition and dynamically adjusts the rank of each layer based on learned importance scores. The method prunes less important singular values during training, theoretically allocating more capacity to layers that contribute most significantly to task performance. This adaptive mechanism promises improved parameter efficiency by concentrating trainable parameters where they matter most.

This study addresses three primary research questions. First, we investigate how QLoRA and AdaLoRA compare in terms of validation loss and perplexity when fine-tuning instruction-tuned language models on medical textbook question answering. Second, we examine whether AdaLoRA's dynamic rank allocation mechanism provides measurable advantages over QLoRA's fixed-rank approach under practical training constraints. Third, we assess the consistency of these findings across different backbone model architectures to determine whether our conclusions generalize beyond specific model families.

Our experimental contributions are threefold. We present the first systematic comparison of QLoRA and AdaLoRA specifically targeting medical textbook question answering, using a curated dataset spanning 13 medical textbooks and covering major clinical and biomedical disciplines. We provide detailed analyses of training dynamics, convergence behavior, and parameter efficiency for both methods across two distinct backbone architectures. Finally, we offer practical recommendations for practitioners seeking to deploy parameter-efficient fine-tuning in medical natural language processing applications.

RELATED WORKS

Large Language Models in Healthcare

The application of transformer-based language models to healthcare tasks has progressed through several developmental phases. Early work focused on domain-specific pretraining, with BioBERT (Lee *et al.* 2020) demonstrating that continued pretraining on biomedical literature improves performance on named entity recognition, relation extraction, and question answering tasks within the biomedical domain. ClinicalBERT (Huang *et al.* 2019) extended this approach to clinical notes, showing that exposure to electronic health record data during pretraining enhances model understanding of clinical language patterns and medical terminology.

The emergence of instruction-tuned large language models opened new possibilities for medical applications. Singhal *et al.* (2023) introduced Med-PaLM, demonstrating that appropriately prompted large language models can approach physician-level performance on medical licensing examination questions. Subsequent work explored fine-tuning strategies specifically tailored to medical dialogue and consultation scenarios. ChatDoctor (Li *et al.* 2023) applied supervised fine-tuning to create a medical conversational agent, while MedAlpaca (Han *et al.* 2023) demonstrated effective medical adaptation using instruction-following datasets derived from clinical guidelines and medical literature.

Parameter-Efficient Fine-Tuning Methods

Parameter-efficient fine-tuning encompasses a diverse family of techniques designed to adapt pretrained models while updating only a small subset of parameters. Comprehensive surveys by Lialin *et al.* (2023) and Han *et al.* (2024) provide taxonomies of these approaches, broadly categorizing them into adapter-based methods, prompt-based methods, and reparameterization-based methods.

Adapter methods insert small trainable modules between frozen pretrained layers. The original adapter formulation by Houlsby *et al.* (2019) demonstrated that adding bottleneck layers with as few as 3% additional parameters could achieve competitive performance on diverse natural language understanding benchmarks. Prompt-based methods modify the input representation rather than the model architecture. Prefix tuning (Li and Liang 2021) prepends trainable continuous vectors to the input sequence, while prompt tuning (Lester *et al.* 2021) optimizes task-specific soft prompts.

QLoRA (Dettmers *et al.* 2024) combines low-rank adaptation with aggressive quantization, introducing the 4-bit NormalFloat format that preserves information content while reducing memory footprint. AdaLoRA (Zhang *et al.* 2023) addresses a potential limitation of standard LoRA, namely the assumption that all weight matrices benefit equally from a given rank allocation. The method parameterizes weight updates as DeltaW = PAQ, where A is a diagonal matrix of singular values that can be pruned based on importance scores computed during training.

METHODS

Dataset Description

Our experiments utilize the MedicalTextbook_QA dataset, which comprises question-answer pairs extracted from 13 medical textbooks representing core disciplines in medical education and clinical practice. Table 1 presents the complete list of source textbooks along with their respective subject domains. Each textbook contributes 500 question-answer pairs to the dataset, yielding a total of 6,500 instances that cover anatomy, biochemistry, cell biology, gynecology, histology, immunology, neurology, obstetrics, pathology, pediatrics, pharmacology, physiology, and psychiatry.

Table 1 Medical textbooks comprising the MedicalTextbook_QA dataset, organized by clinical and basic science domains

Textbook	Domain	Samples
Gray's Anatomy	Anatomy	500
Lippincott Biochemistry	Biochemistry	500
Alberts Cell Biology	Cell Biology	500
Novak's Gynecology	Gynecology	500
Ross Histology	Histology	500
Janeway's Immunology	Immunology	500
Adams Neurology	Neurology	500
Williams Obstetrics	Obstetrics	500
Robbins Pathology	Pathology	500
Nelson Pediatrics	Pediatrics	500
Katzung Pharmacology	Pharmacology	500
Levy Physiology	Physiology	500
DSM-5	Psychiatry	500
Total		6,500

The dataset was partitioned into training and validation subsets, allocating 5,000 instances for training and 500 instances for validation. This split was performed after shuffling with a fixed random seed to ensure reproducibility across experimental conditions.

Backbone Models

We selected two instruction-tuned language models as backbone architectures for our comparative evaluation: Meta-Llama-3-8B-Instruct and Qwen2-7B-Instruct. Both models represent current state-of-the-art open-weight language models optimized for instruction following and dialogue applications. Table 2 summarizes the key specifications of both models.

Table 2 Specifications of backbone models used in experimental evaluation

Specification	Llama-3-8B	Qwen2-7B
Total Parameters	8.03B	7.62B
Quantized Params	4.71B	4.51B
Pretraining Tokens	15T	7T
Attention Type	Grouped-Query	Grouped-Query
Position Encoding	RoPE	RoPE

Parameter-Efficient Fine-Tuning Configurations

QLoRA Configuration QLoRA was implemented using the standard formulation where low-rank adapters are applied to frozen quantized base model weights. The base model weights were quantized to 4-bit precision using the NormalFloat4 data type. The low-rank adaptation matrices were configured with rank r = 64 and scaling factor alpha = 16, yielding an effective learning rate scaling of alpha/r = 0.25 applied to the adapter outputs. Dropout regularization was applied to the adapter layers with probability 0.05 to prevent overfitting.

AdaLoRA Configuration AdaLoRA was configured to enable dynamic rank allocation through importance-based pruning of singular values. The initial rank was set to r_init = 64, matching the QLoRA configuration, with a target rank of r_target = 8 representing an 87.5% reduction in rank through the pruning process. Orthogonal regularization with weight 0.5 was applied to encourage orthogonality between the left and right singular vector matrices.

Training Procedure

All experiments were conducted on a single NVIDIA A100-SXM4-80GB GPU. The training procedure employed the AdamW optimizer with a learning rate of 2e-4 and weight decay of 0.01. A cosine learning rate schedule was applied with a warmup period comprising 3% of total training steps. Each model was trained for a single epoch with a per-device batch size of 4 and gradient accumulation over 4 steps, yielding an effective batch size of 16.

Evaluation Metrics

Model performance was assessed using validation loss and perplexity as primary evaluation metrics. Perplexity, defined as the exponential of the validation loss, offers an interpretable measure of model uncertainty, with lower values indicating better predictive performance. The relationship between these metrics is expressed as: Perplexity = exp(L_val), where L_val denotes the average validation loss.

RESULTS

Overall Performance Comparison

Table 3 presents the complete experimental results across all model and method combinations. QLoRA consistently achieved lower validation loss and perplexity values compared to AdaLoRA for both backbone models. On the Llama-3-8B-Instruct backbone, QLoRA attained a validation loss of 0.0814 and perplexity of 1.085, compared to AdaLoRA's validation loss of 0.1177 and perplexity of 1.125. This represents a relative reduction of 30.8% in validation loss when using QLoRA over AdaLoRA.

The performance gap was more pronounced on the Qwen2-7B-Instruct backbone. QLoRA achieved a validation loss of 0.0821 and perplexity of 1.086, while AdaLoRA produced a validation loss of 0.1563 and perplexity of 1.169. The relative improvement in validation loss for QLoRA over AdaLoRA reached 47.5% on this backbone.

Parameter Efficiency Analysis

Both methods achieved comparable trainable parameter counts. For Llama-3-8B, QLoRA utilized 167.77 million trainable parameters while AdaLoRA used 167.79 million parameters, representing approximately 3.56% of the total 4.71 billion parameters in the quantized model. For Qwen2-7B, the trainable parameter counts were 161.48 million for QLoRA and 161.49 million for AdaLoRA, corresponding to approximately 3.58% of the 4.51 billion total parameters. The minimal difference in trainable parameters between methods indicates that the performance disparities observed cannot be attributed to differences in model capacity.

Table 3 Comprehensive experimental results comparing QLoRA and AdaLoRA across backbone models. Lower values indicate better performance for all metrics except trainable parameters

Backbone	Method	Train Loss	Val. Loss	Perp.	Param (M)
Llama-3-8B	QLoRA	0.0817	0.0814	1.085	167.77
Llama-3-8B	AdaLoRA	0.1234	0.1177	1.125	167.79
Qwen2-7B	QLoRA	0.0822	0.0821	1.086	161.48
Qwen2-7B	AdaLoRA	0.1986	0.1563	1.169	161.49

DISCUSSION

Interpretation of Results

Our experimental findings reveal a consistent performance advantage for QLoRA over AdaLoRA under single-epoch training conditions on medical textbook question answering. Several factors may explain this pattern.

First, the single-epoch training budget may be insufficient for AdaLoRA's adaptive rank allocation to reach its optimal configuration. AdaLoRA's pruning schedule progressively reduces ranks from the initial value of 64 to the target value of 8 between the warmup period at 10% of training and the finalization point at 70% of training. Extended training over multiple epochs would allow the model more time to stabilize after each rank reduction, potentially enabling AdaLoRA to realize its theoretical advantages.

Second, the medical question answering task may not exhibit the layer-wise importance heterogeneity that AdaLoRA is designed to exploit. If the importance of different weight matrices is relatively uniform across the model for this particular task, adaptive rank allocation provides no advantage over fixed-rank approaches.

Implications for Medical NLP

Our results have practical implications for deploying parameter-efficient fine-tuning in medical natural language processing applications. The consistent performance advantage of QLoRA suggests that this method represents a reliable choice for healthcare institutions seeking to adapt large language models to medical domains under computational constraints. The 4-bit quantization employed by QLoRA substantially reduces memory requirements, enabling fine-tuning on consumer-grade hardware without sacrificing performance on medical question answering tasks.

Limitations

This study has several limitations. First, our evaluation relied exclusively on perplexity and loss metrics. Second, the single-epoch training schedule may not represent optimal training conditions for either method. Third, we employed default hyperparameters without extensive optimization. Fourth, our experiments were limited to two backbone models. Fifth, the MedicalTextbook_QA dataset represents a specific question answering format that may not transfer directly to other medical NLP tasks.

CONCLUSION

This study presented a systematic comparison of QLoRA and AdaLoRA for parameter-efficient fine-tuning of large language models on medical textbook question answering. Our experiments across two backbone architectures, Llama-3-8B-Instruct and Qwen2-7B-Instruct, demonstrate that QLoRA consistently outperforms AdaLoRA under single-epoch training conditions. QLoRA achieved validation perplexity values of 1.085 and 1.086 for the two backbones respectively, compared to AdaLoRA's 1.125 and 1.169, representing relative validation loss reductions of 30.8% and 47.5%.

Both methods maintained comparable trainable parameter counts at approximately 3.5% of total model parameters, indicating that the performance differences stem from how each method utilizes its parameter budget rather than from differences in model capacity. Our analysis suggests that QLoRA's fixed-rank approach provides more stable convergence behavior within limited training budgets, while AdaLoRA's adaptive rank allocation mechanism may require extended training schedules to realize its theoretical advantages. These findings offer practical guidance for medical NLP practitioners: QLoRA represents a reliable and effective choice for adapting large language models to medical domains when computational resources are constrained.

Acknowledgments

This work was supported by Igdir University.

Ethical standard

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

Availability of data and material

The MedicalTextbook_QA dataset is available through the Hugging Face Hub at https://huggingface.co/datasets/winder-hybrids/MedicalTextbook_QA.

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

The authors declare that generative artificial intelligence (AI) tools were used during the preparation of this manuscript. Specifically, AI assistance was utilized for language editing, text refinement, and formatting purposes. The authors take full responsibility for the content and have carefully reviewed and verified all AI-assisted outputs.

LITERATURE CITED

Bayat Toksoz, S. and G. Isik, 2025 Parameter-efficient fine-tuning of llama models for financial sentiment classification. Cluster Computing **29**: 41.

Brown, T. B., B. Mann, N. Ryder, *et al.*, 2020 Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901.

Chowdhery, A., S. Narang, J. Devlin, *et al.*, 2023 Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**: 1–113.

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer, 2024 Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115.

Han, T., L. C. Adams, J.-M. Papaioannou, *et al.*, 2023 Medalpaca: An open-source collection of medical conversational ai models and training data. arXiv preprint .

Han, Z., C. Gao, J. Liu, *et al.*, 2024 Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint .

Houlsby, N., A. Giurgiu, S. Jastrzebski, *et al.*, 2019 Parameter-efficient transfer learning for nlp. In *Proceedings of ICML*, volume 97, pp. 2790–2799.

Hu, E. J., Y. Shen, P. Wallis, *et al.*, 2022 Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Huang, K., J. Altosaar, and R. Ranganath, 2019 Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint .

Karabacak, M. and K. Margetis, 2023 Embracing large language models for medical applications. Cureus **15**: e39305.

Lee, J., W. Yoon, S. Kim, *et al.*, 2020 Biobert: A pre-trained biomedical language representation model. Bioinformatics **36**: 1234–1240.

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.

Li, X. L. and P. Liang, 2021 Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the ACL*, pp. 4582–4597.

Li, Y., Z. Li, K. Zhang, *et al.*, 2023 Chatdoctor: A medical chat model fine-tuned on llama. Cureus **15**: e40895.

Lialin, V., V. Deshpande, and A. Rumshisky, 2023 Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv preprint .

Singhal, K., S. Azizi, T. Tu, *et al.*, 2023 Large language models encode clinical knowledge. Nature **620**: 172–180.

Thirunavukarasu, A. J., D. S. J. Ting, *et al.*, 2023 Large language models in medicine. Nature Medicine **29**: 1930–1940.

Zhang, Q., M. Chen, A. Bukharin, *et al.*, 2023 Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*.

How to cite this article: Toksoz, S. B. and Isik, G. A Comparative Evaluation of QLoRA and AdaLoRA for Parameter-Efficient Fine-Tuning of Large Language Models on Medical Textbook Question Answering. *Artificial Intelligence in Applied Sciences*, 2(1),27-31, 2026.