baellouf at AbjadMed: Efficient Fine-tuning with All-Linear LoRA for Arabic Medical QA Classification

Abdallah Khallouf
Independent Researcher
abdallahkhallouf2@gmail.com

Abstract

We describe our system for the AbjadMed shared task on Arabic medical text classification at AbjadNLP 2026. Our approach combines efficient fine-tuning of Qwen3-8B using QLoRA with a Dice+CrossEntropy hybrid loss designed for Macro F1 optimization. Taking inspiration from recent research on optimal LoRA configurations, we apply low-rank adapters to all linear layers of the model rather than attention layers only, which we validate improves performance by 4.0 points. We also explore data augmentation through machine translation of external medical QA data, though this did not improve generalization. Our best submission achieves a Macro F1 score of 0.4441 on the test set.

1 Introduction

Arabic medical text classification presents unique challenges due to the morphological complexity of Arabic, domain-specific terminology, and the scarcity of labeled medical data in Arabic (Gupta et al., 2026). The AbjadMed shared task addresses these challenges by providing a dataset of Arabic medical question-answer pairs across 82 fine-grained categories, with significant class imbalance ranging from 7 to over 600 samples per class.

A key challenge in this task is the fine-grained nature of the classification taxonomy itself. Many categories represent closely related medical domains with subtle distinctions. For instance, dental-related content is split across 5 categories (Dental diseases, Dental health, Dentistry, Jaw and dental surgery, Oral diseases), while mental health spans 3 overlapping categories (Mental health, Psychiatric diseases, Psychology). Similar fine-grained distinctions exist for pediatric, dermatology, and women's health domains. This granularity requires models to learn nuanced semantic boundaries between related medical specialties, making the classification task substantially harder than a coarse-grained alternative.

To address these challenges, we draw inspiration from the "LoRA Without Regret" research (Schulman and Thinking Machines Lab, 2025), applying QLoRA adapters to **all linear layers** of Qwen3-8B rather than attention layers only. Combined with a Dice+CrossEntropy hybrid loss, we achieve a Macro F1 score of 0.4441 on the test set. Our LoRA adapter is available at https://huggingface.co/baellouf/qwen3-8b-medical-QA-classification-adapter.

2 Task and Data

The AbjadMed task requires classifying Arabic medical question-answer pairs into one of 82 medical categories. The training dataset contains 27,951 samples with severe class imbalance: the largest class has over 600 samples while the smallest has only 7 samples, resulting in an imbalance ratio of approximately 85:1.

Each sample consists of an Arabic text field containing a medical question and its corresponding answer, along with a category label. The categories span diverse medical domains including cardiology, dermatology, mental health, and various surgical specialties.

3 System Description

3.1 Base Model

We use Qwen3-8B (Qwen Team, 2024) as our base model. Qwen3 is a multilingual large language model with strong performance on Arabic text, making it suitable for this task. The 8B parameter variant provides a good balance between model capacity and training efficiency.

3.2 Efficient Fine-tuning with QLoRA

We employ QLoRA for parameter-efficient fine-tuning (Hu et al., 2021). Following the "LoRA Without Regret" research (Schulman and Thinking Machines Lab, 2025), we apply low-rank adapters

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 124–126
March 28, 2026. ©2026 Association for Computational Linguistics

to **all linear layers** of the model rather than only attention layers. The key insight from this work is that "attention-only LoRA significantly underperforms MLP-only LoRA, and does not further improve performance on top of LoRA-on-MLP." Crucially, this underperformance is not explained by having fewer parameters: attention-only LoRA with rank 256 underperformed MLP-only LoRA with rank 128 despite similar parameter counts.

Our LoRA configuration uses rank $r = 256$ and $\alpha = 16$, with dropout disabled. Following the finding from Schulman and Thinking Machines Lab (2025) that optimal learning rates for LoRA are approximately 10× higher than for full fine-tuning, we use a learning rate of 2×10^{-4}. The high rank ensures sufficient capacity for learning the 82-class classification task, while LoRA's compute efficiency (approximately $\frac{2}{3}$ of the FLOPs of full fine-tuning) enables faster experimentation.

3.3 Loss Function

To directly optimize for the Macro F1 evaluation metric, we employ a Dice+CrossEntropy hybrid loss (Li et al., 2020). The Dice loss component provides a differentiable approximation of the F1 score, while CrossEntropy ensures stable gradients during training.

The combined loss is defined as:

$$\mathcal{L} = 0.5 \cdot \mathcal{L}_{\text{Dice}} + 0.5 \cdot \mathcal{L}_{\text{CE}} \tag{1}$$

4 Experimental Setup

4.1 Training Configuration

We train for 3 epochs with a batch size of 4 and gradient accumulation over 8 steps, yielding an effective batch size of 32. We use the AdamW optimizer with a learning rate of 2×10^{-4}, 100 warmup steps, and weight decay of 0.01. Training is performed in bfloat16 precision.

We use a maximum sequence length of 1024 tokens and reserve 10% of the training data for validation. The model achieving the best validation Macro F1 is selected for final submission.

4.2 Infrastructure

Training was performed on a single NVIDIA H100 80GB GPU. A typical training run on the augmented dataset required approximately 7–8 hours for 6,168 gradient steps; runs on the original dataset completed in approximately 3–4 hours.

5 Results

Our best submission achieves a Macro F1 score of 0.4441 on the test set. This result was achieved using the original training data without augmentation. We note that this score comes from an unselected submission; our selected submission, which was trained on the augmented dataset, scored 0.4398, though this does not change our ranking.

5.1 LoRA Without Regret

Following the recommendations from Schulman and Thinking Machines Lab (2025), we apply LoRA to all linear layers rather than attention layers only. To confirm this choice on our task, we compared our all-linear LoRA configuration against an attention-only variant. Both configurations use identical hyperparameters (rank 256, learning rate 2×10^{-4}, Dice+CE loss) and train on the original data without augmentation.

LoRA Target Modules	**F1**
Attention only (q, k, v, o)	0.404
All linear layers	**0.444**

Table 1: Comparison of LoRA target module configurations on Qwen3-8B (Macro F1 on test set).

As shown in Table 1, applying LoRA to all linear layers outperforms attention-only LoRA by 4.0 points, confirming the findings from Schulman and Thinking Machines Lab (2025) that MLP layers are critical for LoRA performance. Using attention-only LoRA would have resulted in a significantly lower score.

5.2 Ablation Studies

We conducted ablation experiments to understand the contribution of each component. Model selection during development was based on validation Macro F1; the results in Table 2 are reported on the competition test set.

Loss Function. Switching from Focal Loss to Dice+CE improved Macro F1 by 1.6 points on the 4B model. The Dice loss component, which directly approximates F1, provides better alignment with the evaluation metric.

Model Comparison. Interestingly, MedGemma-27B, despite having 3× more parameters than Qwen3-4B and being specifically trained on medical data, only marginally outperformed it (0.423

Model	Configuration	F1
Loss Function (Qwen3-4B)		
Qwen3-4B	Focal Loss (γ=2)	0.403
Qwen3-4B	Dice+CE	0.419
Model Comparison (Dice+CE)		
Qwen3-4B	Original data	0.419
MedGemma-27B	Original data	0.423
Data Augmentation (Qwen3-8B, Dice+CE)		
Qwen3-8B	Original only	**0.444**
Qwen3-8B	+ iCliniq aug	0.4398

Table 2: Ablation results on the test set (Macro F1).

vs 0.419). While this is a single comparison, it suggests that multilingual capabilities may be at least as important as domain-specific pretraining for this task.

Data Augmentation. Contrary to our expectations, adding the translated iCliniq data actually hurt performance, dropping from 0.444 to 0.4398. We attribute this to label noise introduced by LLM-based classification of the translated data, which caused the model to overfit to noisy training patterns.

6 Conclusion

We presented our system for the AbjadMed Arabic medical text classification task. Our approach combines Qwen3-8B fine-tuning with QLoRA applied to all linear layers, validated to outperform attention-only LoRA by 4.0 points, and a Dice+CE hybrid loss for Macro F1 optimization. We achieve a Macro F1 of 0.4441 on the test set. Notably, data augmentation through LLM-translated medical QA data did not improve generalization, likely due to label noise from automated classification.

Limitations

Our approach has several limitations:

- Our LLM-based data augmentation introduced label noise that hurt test performance despite dramatically improving validation metrics, a cautionary finding for similar approaches.
- We did not perform extensive hyperparameter tuning due to computational constraints.

Ethics Statement

Medical text classification systems should be used as decision support tools rather than autonomous diagnostic systems. Misclassification of medical queries could potentially lead to inappropriate medical advice. We recommend human expert review for any clinical applications.

Acknowledgements

We thank the AbjadMed shared task organizers for providing the dataset and evaluation infrastructure.

References

Pranav Gupta, Niranjan Kumar M, Balaji Nagarajan, Imed Zitouni, and Mo El-Haj. 2026. Abjadmed: Arabic medical text classification at abjadnlp 2026. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), co-located with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*, Rabat, Morocco.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476.

Qwen Team. 2024. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

John Schulman and Thinking Machines Lab. 2025. Lora without regret. *Thinking Machines Lab: Connectionism*. https://thinkingmachines.ai/blog/lora/.