ArabicMedicalBERT-QA-82 at AbjadMed: Fighting Class Imbalance in Arabic Medical Text Classification

Gleb Shanshin
ITMO University
Saint Petersburg, Russia
gleb.shanshin@niuitmo.ru

Abstract

We present a supervised system for Arabic medical question-answer classification developed for the AbjadMed shared task. The task involves assigning one of 82 highly imbalanced medical categories and is evaluated using macro-averaged F1. Our approach builds on an AraBERT model further pretrained on a related Arabic medical classification dataset. Under a unified fine-tuning setup, this domain-adapted model consistently outperforms general-purpose Arabic backbones, with the best results obtained using a low backbone learning rate, indicating that only limited adaptation is required. The final system achieves a macro F1 score of 0.51 on the private test split. For comparison, we evaluate several cost-efficient large language models under constrained prompting and observe substantially lower performance.

1 Introduction

Medical natural language processing for Arabic remains challenging due to domain-specific terminology, limited annotated data, and severe class imbalance. These challenges are particularly pronounced in medical text classification, where fine-grained category distinctions and skewed label distributions degrade the performance of standard models. The AbjadMed Shared Task at EACL 2026 (Gupta et al., 2026) targets this problem by benchmarking multi-class classification of Arabic medical question-answer pairs. In this work, we investigate supervised transformer-based approaches with a focus on domain-aligned pretraining, and complement them with an analysis of zero-shot and few-shot classification using cost-efficient large language models under the same evaluation setting.

Arabic (Original)	English (LLM Translation)
السؤال المنجلي الدم بفقر مصاب انا عليكم السلام 72 السكلسل نسبة بأن علماً (السكلسل) تأتي الالام فأن 7 الدم نسبة تصبح فعندما الحل وما الدم نسبة لزيادة الحل فما بكثره لعلاج…	**Question** Hello, I have sickle cell anemia (sickle cell disease). My sickle-cell level is 72, and when my hemoglobin drops to 7, the pain attacks become frequent. What is the solution to increase my blood level, and what is the treatment…?
الجواب وتقوية النفسية الرضرض عن بالابتعاد الحل بالحديد غنية متوازنة غذائية حمية وتناول المناعة حاد ونقص سببها الام نوبات حدوث وعند الدم تعويض الا لايوجد الدموي بالخضاب الدم• الناقص…بنقل	**Answer** The solution is to avoid psychological stress, strengthen the immune system, and follow a balanced diet rich in iron. When painful crises occur due to a severe drop in hemoglobin, the only option is to compensate for the missing blood by transfusion.

Table 1: Example of an Arabic medical question-answer pair for category “Hematological diseases” and its English translation.

2 Dataset

The shared task training set contains 27,951 Arabic medical question-answer pairs. Test set contains 18,634 pairs split equally on public and private leaderboards without labels being revealed. An example instance and its automatic English translation are shown in Table 1.

The dataset exhibits extreme class imbalance: frequent categories contain up to 600 instances, whereas rare classes such as *In vitro fertilization (IVF)* are represented by as few as seven examples. In addition, the label space includes partially overlapping and semantically adjacent categories, such as *Dentistry*, *Dental diseases*, *Dental health*, and *Jaw and dental surgery*, as well as *Pediatric diseases*, *Child health*, and *Pediatric surgery* and others, which further increases the difficulty of accurate classification.

3 Finetuning

Since the texts are in Arabic, we focus on transformer backbones pretrained on Arabic corpora. We evaluate the following publicly

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 115–119

Backbone	Backbone LR	Val	Public	Private
Shared Task Baseline (3 epochs)				
CAMeL-Lab/bert-base-arabic-camelbert-da	2e-5	-	**0.2939**	**0.2896**
Model with remapped classes (no fine-tuning)				
AymanElbery/arabic-medical-classifier-arabertv2	–	**0.4605**	**0.4684**	**0.4848**
Fine-tuned models (90% of training data)				
CAMeL-Lab/bert-base-arabic-camelbert-da	1e-6	0.3209	0.3316	0.3232
CAMeL-Lab/bert-base-arabic-camelbert-da	2e-5	0.3558	0.3774	0.3529
aubmindlab/bert-base-arabertv2	1e-6	0.3580	0.3432	0.3372
aubmindlab/bert-base-arabertv2	2e-5	0.3871	0.3844	0.3953
AymanElbery/arabic-medical-classifier-arabertv2	1e-6	**0.5294**	**0.4986**	**0.5099**
AymanElbery/arabic-medical-classifier-arabertv2	2e-5	0.5041	0.4612	0.4714
10-fold majority voting (final submitted model)				
AymanElbery/arabic-medical-classifier-arabertv2	1e-6	–	**0.5071**	**0.5139**
Fine-tuned model (100% of training data)				
AymanElbery/arabic-medical-classifier-arabertv2	1e-6	–	**0.5165**	**0.5153**

Table 2: Validation and leaderboard results for fine-tuned models (macro F1).

Agreement	Count	Share
1-2/10	0	0%
3/10	6	0%
4/10	126	0.7%
5/10	510	2.7%
6/10	706	3.8%
7/10	801	4.3%
8/10	1,070	5.7%
9/10	1,520	8.2%
10/10	13,895	74.6%

Table 3: Distribution of cross-validation prediction agreement. k denotes the number of folds (out of 10) that predict the final majority-vote label for an instance.

available models from HuggingFace:

- CAMeL-Lab/bert-base-arabic-camelbert-da (Inoue et al., 2021)
- aubmindlab/bert-base-arabertv2 (Antoun et al., 2020)
- AymanElbery/arabic-medical-classifier-arabertv2 (Elbery, 2025)

All models are compared under a unified training setup with a batch size of 16, maximum sequence length of 512 tokens, 10% stratified validation split, class-weighted cross-entropy using inverse class-frequency weights. We used AdamW optimizer with a weight decay of 0.01 and a cosine learning rate scheduler with 5% warmup, during the training we select the best checkpoint over 10 training epochs. Also we use discriminative learning rates: 5e−4 for the classification head and 1e−6, 2e−5 for the transformer backbone; other settings are shared across models

The AymanElbery/arabic-medical-classifier-arabertv2 model was pretrained on a closely related medical classification task with 89 categories. We partially remap its label space to the shared task categories. Although the mapping is not exact, this model substantially outperforms the other backbones even without task-specific fine-tuning (Table 2). As a result, fine-tuning this model requires minimal deviation from the pretrained weights, and lower learning rates yield better performance.

Each input instance is constructed from the Arabic question-answer pair in the form `<question> [SEP] <answer>`. The explicit markers السؤال (“Question”) and الجواب (“Answer”) are removed to maintain compatibility with the pretrained models.

We additionally experimented with imbalance-mitigation techniques such as focal loss (Lin et al., 2017), label smoothing (Szegedy et al., 2016), and data augmentation via token masking and text cropping. While these methods showed minor improvements in preliminary experiments, they were not included in the final submission.

Since training is performed on 90% of the dataset, we further apply 10-fold cross-validation and aggregate predictions via majority voting. To assess ensemble stability, we analyze prediction

Model	Failed	AAPT	Public	Private
Zero-shot				
gpt-4.1-nano	21	1.36	0.2722	0.2589
deepseek-v3.2	**1**	**1.06**	**0.2964**	**0.2977**
Few-shot				
gpt-4.1-nano	25	1.43	0.2770	0.2624
deepseek-v3.2	**7**	**1.08**	**0.3196**	**0.3157**

Table 4: Performance of cost-efficient LLMs. **Failed**: no valid label within 100 attempts; **AAPT**: average attempts per text. Macro F1 is reported for public/private leaderboards.

Hallucinated label	Count	Most frequent valid label after refinement
gpt-4.1-nano		
Proctology	61	Gastrointestinal diseases
Oncology	51	Benign and malignant tumors
Herbal medicine	37	Alternative medicine
Digestive diseases	35	Gastrointestinal diseases
Gynecologic diseases	35	Women's health
deepseek-v3.2		
Urological diseases	107	Urogenital diseases
Breast diseases	31	Gynecological diseases
Proctology	24	General surgery
Oncology	19	Benign and malignant tumors
Liver diseases	12	Gastrointestinal diseases

Table 5: Most frequent out-of-list labels generated under zero-shot inference and the corresponding valid labels selected after constrained re-prompting.

agreement across folds at the instance level. The majority of instances exhibit high agreement, with over 88% receiving agreement from at least 9 folds, indicating that disagreements are confined to a relatively small subset of inputs. The resulting ensemble performance is reported in Table 2 and agreement decomposition in Table 3.

Despite the fact that fold aggregation usually gives a more stable result and this solution was used during the Shared Task, training on the entire training dataset for 10 epochs gets slightly higher metrics. This model was open-sourced and available with HuggingFace at gleb-shnshn/arabic-medical-bert-qa-82.

4 LLM

We evaluate zero-shot and few-shot classification using cost-efficient large language models, specifically deepseek-v3.2 (DeepSeek-AI and many contributors, 2025) and gpt-4.1-nano (OpenAI, 2025). These models are assessed under the same label space and macro F1 evaluation protocol as the supervised systems.

The prompt structures for zero-shot and few-shot inference are shown in Table 6. For each input instance, invalid model outputs are appended to the prompt as negative examples in subsequent attempts. If no valid category is produced within 100 attempts, the prediction is assigned to the fallback class *General Medicine*.

The most frequently produced *out-of-list* labels under zero-shot inference are summarized in Table 5, together with the most common valid label selected after refinement across attempts. These results indicate a systematic tendency to generate plausible but non-enumerated category names rather than selecting from the provided label set.

In the few-shot setting, a single labeled example per category is included in the prompt. The shortest available instance is selected for each class to minimize prompt length. Table 4 reports failure rates (no valid prediction within 100 attempts), average attempts per text (AAPT), and leaderboard macro F1 scores. Although few-shot prompting slightly increases the number of attempts and failures, it consistently improves classification performance across both models.

Part	Zero-shot	Few-shot
Role	You are a medical text classifier for Arabic medical dialogues. Your task is to classify the text into exactly one of the available categories.	
Examples	-	Example 1: Text: <Arabic text> Category: Addiction ⋮ Example 82: Text: <Arabic text> Category: Women's health
Rules	Available categories (choose ONLY from this list): ["Addiction", "Allergy", "Alternative medicine", "Anatomy" … "Vitamins and minerals", "Women's health"] Rules: 1. Output ONLY the category name, nothing else 2. Choose the most specific and relevant category 3. If unsure, use the closest match from the list 4. You MUST choose from the available categories above	
Invalid Attempts	Previous invalid attempts (these are NOT valid options, do NOT use them): 1. "Urogenital diseases" is not a valid option	
Task	Classify this Arabic medical text: Text: <Arabic text> Category:	

Table 6: Prompt structure comparison between zero-shot and few-shot; shared parts are merged across both columns.

5 Conclusion

We show that, under extreme label imbalance, transferring supervision from a related medical classification task is an effective strategy for Arabic domain adaptation. The results suggest that most performance gains come from preserving pretrained representations rather than aggressive fine-tuning, as evidenced by the effectiveness of low backbone learning rates. At the same time, in our closed-set classification setup, cost-efficient general-purpose LLMs showed substantially lower performance and frequent out-of-list label generation, indicating limited reliability without additional constraints or post-processing. Future work includes exploring large and Arabic-native models and more principled approaches to handling rare classes.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

DeepSeek-AI and many contributors. 2025. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *arXiv preprint arXiv:2512.02556*.

Ayman Elbery. 2025. arabic-medical-classifier-arabertv2: Arabic Medical Classification Model (AraBERTv2). https://huggingface.co/AymanElbery/arabic-medical-classifier-arabertv2. Hugging Face model. Fine-tuned AraBERTv2 for Arabic medical text classification (89 classes). License: Apache-2.0. Accessed: 2026-01-10.

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.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the*

IEEE International Conference on Computer Vision (ICCV), pages 2980–2988.

OpenAI. 2025. Introducing the GPT-4.1 Model Series Including GPT-4.1-nano. https://openai.com/index/gpt-4-1/.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.