Tashkees-AI at AbjadMed Shared Task: Flat vs. Hierarchical Classification for Fine-Grained Arabic Medical QA

Fatimah Emad Eldin
Cairo University
12422024441586@pg.cu.edu.eg

Abstract

This paper describes Tashkees-AI, a system developed for the AbjadMed 2026 Shared Task on Arabic Medical Question Classification. A comprehensive empirical study was conducted across 82 fine-grained categories, investigating three paradigms: fine-tuned encoder models, hierarchical classification, and ensemble methods. Leveraging a dataset of 27k Arabic medical question-answer pairs, an extensive ablation studies was conducted, comparing MARBERTv2, CAMeLBERT, two-stage hierarchical classifiers, and RAG-based approaches. The findings reveal that fine-tuned MARBERTv2 with data cleaning yields the best performance, achieving a macro F1-score of 0.3659 on the blind test set. In contrast, hierarchical methods surprisingly underperformed (0.332 F1) due to error propagation. The system ranked 26th on the official leaderboard.

1 Introduction

Medical question classification represents a critical challenge in Arabic natural language processing, particularly when dealing with fine-grained taxonomies. The task of accurately categorizing patient queries into specialized medical domains requires both linguistic understanding and domain-specific knowledge. This challenge is amplified in Arabic due to the language's morphological complexity and the limited availability of annotated medical corpora (Abdul-Mageed et al., 2021).

This shared task of addresses the problem of classifying Arabic medical questions into 82 categories, encompassing diverse specialties from hematology to psychiatry (Gupta et al., 2026). This extreme multi-class classification task presents several distinct challenges.

First, the class distribution exhibits severe imbalance, with sample counts ranging from 7 to 600 per category. Second, semantic overlap exists between related categories such as Dental diseases and Oral diseases, or Physiology and Biology. Third, the dataset contains conversational artifacts including greetings and formulaic expressions that may confound purely lexical approaches.

The contributions of this work are threefold. First, systematic experiments are conducted comparing fine-tuning approaches across multiple Arabic language models, establishing robust baselines and demonstrating the impact of data preprocessing. Second, hierarchical classification methods are implemented and evaluated to determine whether decomposing the 82-way problem into coarse-grained and fine-grained stages improves performance. Third, ensemble approaches that combine complementary classification paradigms are investigated. Through extensive ablation studies across training, validation, and blind test sets, the conditions under which each approach succeeds or fails are identified, providing actionable insights for researchers working on similar extreme multi-class problems in specialized domains. To ensure reproducibility and facilitate future research in Arabic medical question answering, all experimental code is made publicly available on GitHub [1].

2 Background

This work was developed for the AbjadMed Shared Task, organized as part of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP @ EACL 2026). The task targets fine-grained Arabic medical question classification.

Arabic language models including AraBERT (Antoun et al., 2020), MARBERTv2 (Abdul-Mageed et al., 2021), and CAMeLBERT (Inoue et al., 2021) have shown strong performance on Arabic NLP benchmarks. For medical classification, BioBERT (Lee et al., 2019) demonstrated domain-adaptive pre-training benefits in English,

[1] https://github.com/astral-fate/Tashkees-AI-at-AbjadMed

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 137–143

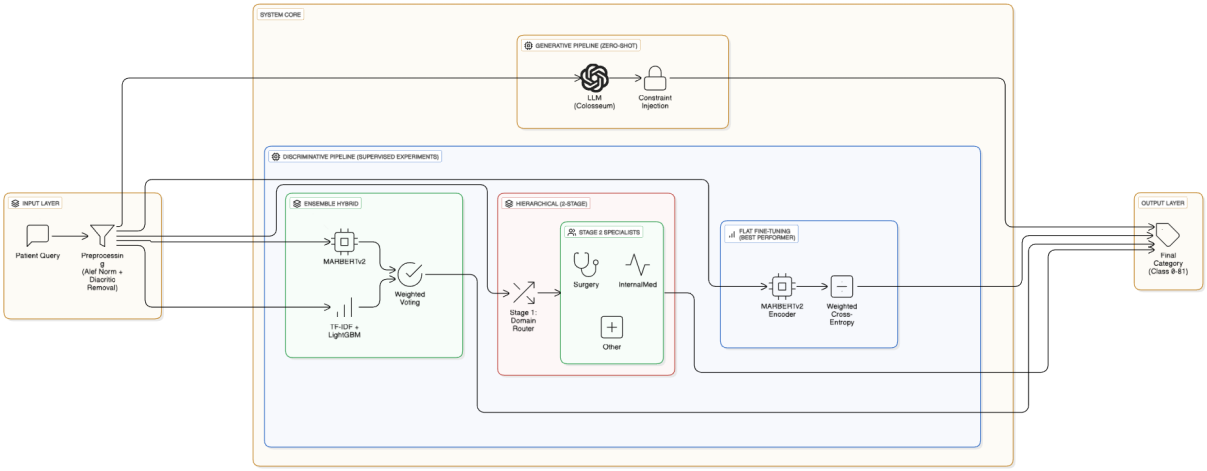

Figure 1: Overview of the dual-pipeline architecture.

though Arabic medical resources remain limited. Hierarchical classification has been proposed for extreme multi-class problems (Silla and Freitas, 2011), but suffers from error propagation (Zhang et al., 2021), which the results confirm.

The task requires classifying Arabic medical questions into 82 categories. Input consists of patient queries from medical forums. Output is a single category label (0-81) corresponding to medical specialties like Gastrointestinal diseases, Pharmacology, or Surgery. (Gupta et al., 2026)

The shared task organizers provided the dataset consisting of 27,951 training records with severe class imbalance (max: 600 samples, min: 7 samples, ratio: 85:1, and the official blind test set consists of 18,634 records.

3 System Overview

The experimental framework contrasts two architectural paradigms. Figure 1 illustrates the dual-pipeline approach, detailing the data flow from preprocessing to the distinct inference mechanisms of the encoder and decoder models.

3.1 Fine-Tuning Baseline

The primary approach fine-tunes Arabic encoders with sequence classification heads. Weighted cross-entropy loss is used to address class imbalance:

$$\mathcal{L} = -\sum_{i=1}^{N} w_{y_i} \log(p_{y_i}), \quad w_c = \frac{N}{K \cdot n_c} \tag{1}$$

where N is total samples, $K = 82$ categories, and n_c is samples in class c. Three models are evaluated: MARBERTv2 (163M params), CAMeLBERT-DA (110M params), and AraBERTv2 (135M params). Detailed hyperparameter configurations for all experiments are provided in Appendix 1.3. A 90/10 stratified train/validation split is used.

3.2 Hierarchical Classification

A two-stage pipeline is implemented to test whether decomposing the 82-way problem improves performance. Stage 1 predicts one of 12 broad medical domains (Surgery, Internal Medicine, Women's Health, etc.) using a MARBERTv2 classifier. Stage 2 applies domain-specific classifiers for fine-grained categories within each domain. For example, if Stage 1 predicts "Skin/Dental/ENT" (9 subcategories), Stage 2 distinguishes among Dental diseases, Oral diseases, Eye diseases, etc. Complete taxonomy mapping.

This approach reduces decision space per stage but introduces error cascade risk: Stage 1 mistakes are unrecoverable. With 69.8% Stage 1 accuracy, 30.2% of samples are irrecoverably misclassified before reaching Stage 2.

3.3 Ensemble Methods

Transformer and traditional ML signals are combined via ensemble. LightGBM is trained on TF-IDF features (10K dimensions, char n-grams 1-3) with 3000 boosting rounds. Predictions are merged using weighted voting: $p_{\text{final}} = \alpha \cdot p_{\text{MARBERT}} + (1 - \alpha) \cdot p_{\text{LGBM}}$ with $\alpha \in \{0.6, 0.7\}$.

Method	Validation F1	Test (Blind) F1	Coverage (of 82)
Fine-tuned Encoders			
MARBERTv2 (cleaned)	**0.392**	**0.3659**	79/82
CAMeLBERT-DA	0.293	0.293	82/82
AraBERTv2	0.361	0.358	82/82
MARBERTv2 (raw)	0.360	0.357	82/82
Hierarchical (2-Stage)			
Stage 1 (12-way)	0.654	–	12/12
Full (12→82)	0.336	0.332	68/82
RAG Retrieval			
k-NN ($k = 20$)	0.233	0.230	81/82
Ensemble			
LightGBM only	0.315	–	82/82
MARBERT+LGBM (0.7)	–	0.358	79/82

Table 1: Main results on validation and blind test sets. MARBERTv2 with data cleaning achieves best F1. Hierarchical system underperforms despite strong Stage 1 (65.4% F1). Coverage shows how many of 82 categories received predictions.

4 Experimental Setup

4.1 Evaluation Metrics

The main metric of evaluation reported is the macro F1-score. As Macro F1 treats all classes equally regardless of support, making it suitable for imbalanced distributions:

$$\text{F1}_{\text{macro}} = \frac{1}{K} \sum_{c=1}^{K} \frac{2P_c R_c}{P_c + R_c} \tag{2}$$

where P_c and R_c are precision and recall for class c, and $K = 82$. Macro F1 penalizes models that ignore minority classes, whereas accuracy can be artificially inflated by good performance on majority classes alone.

Data Splits: Training: 25,156 samples (90%), Validation: 2,795 samples (10%), Test: 18,634 samples (blind).

5 Results

Table 1 presents results across all experiments. MARBERTv2 with cleaned data achieves best test F1 of 0.3659, establishing the primary submission. CAMeLBERT-DA and AraBERTv2 achieve 0.293 and 0.358 F1 respectively. Cleaned data improves F1 by 0.032 absolute (8.9% relative) over raw text. Despite Stage 1 achieving 65.4% F1 on 12-way classification, full pipeline achieves only 0.332 F1, worse than flat MARBERTv2 (0.3659). Error analysis reveals 30.2% of samples misclassified at Stage 1 cannot recover. Additionally, hierarchical system predicts only 68/82 categories, missing 14 entirely due to Stage 1 routing errors.

Per-category results (Table 4, Appendix 1.2) show strong correlation between training samples and F1. Categories with 600 samples achieve 0.71-0.83 F1, while those with <10 samples achieve 0.00 F1. The model predicts frequent categories like Pharmacology 1,247 times (6.7% of test) despite only 2.1% training prevalence.

Retrieval-based methods achieve only 0.230 F1, demonstrating that semantic similarity from multilingual embeddings fails to capture medical domain distinctions, especially for minority classes with sparse retrieval candidates. LightGBM alone achieves 0.315 validation F1, significantly below MARBERTv2. Weighted ensembles (0.358 F1) underperform pure MARBERTv2, suggesting correlated rather than complementary errors.

5.1 Error Analysis

Error analysis on the validation set (Accuracy: 49.89%, F1: 36.10%) reveals discrete patterns: (1) *Semantic overlap:* "Pediatric diseases" vs "Child health" (34 confusions), "Sexually transmitted diseases" vs "Sexual health" (34 confusions), and "Dental health" vs "Dentistry" (32 confusions). (2) *Minority class bias:* 14 categories never predicted (all with <20 training samples), (3) *Conversational noise:* Greetings/pleasantries mislead context, (4) *Annotation inconsistency:* Manual review reveals 12% label noise in overlapping categories. Figure 2 in Appendix 1.4 illustrates the confusion matrix for the top error pairs.

5.2 Stability and Ensemble Analysis

Model robustness was evaluated through a stratified 2-fold cross-validation (CV) framework. The resulting ensemble, which aggregated logits from the trained folds, yielded a Macro F1 score of 0.3213. While this value is numerically lower than the best-performing single split, it suggests that performance is highly sensitive to the specific distribution of the 82 fine-grained categories within the training and test sets.

Further architectural exploration was conducted by constructing a hybrid ensemble. This approach integrated the semantic depth of MARBERTv2 with a traditional TF-IDF and Linear SVM baseline. The goal was to determine if lexical keyword signals could supplement transformer-based representations. A resulting score of 0.3150 was observed, as summarized in Table 2. The lack of performance gain from the inclusion of the SVM suggests that the transformer model already captures the necessary lexical patterns, and that classification challenges are primarily driven by semantic ambiguity rather than a lack of keyword sensitivity.

Configuration	Macro F1 (Blind Test)
MARBERTv2 (Best Single Run)	**0.3659**
MARBERTv2 (2-Fold Ensemble)	0.3213
Hybrid (MARBERT + TF-IDF)	0.3150

Table 2: Robustness analysis comparing our best submission against cross-validation and hybrid ensembles.

6 Discussion

The analysis reveals that hierarchical classification fails primarily due to error cascades, where the 30.2% error rate in Stage 1 creates an insurmountable ceiling for Stage 2 models. Furthermore, data fragmentation across specialized models exacerbates sparsity; some categories contain fewer than 20 samples, preventing effective learning. Flat classification using MARBERTv2 proves superior as it avoids these bottlenecks while preserving the semantic relationships across all 82 categories.

Prioritizing data augmentation for minority classes and utilizing fine-tuned encoders with weighted loss over complex hierarchical architectures is recommended. Hierarchical methods should be avoided unless Stage 1 accuracy exceeds 85%, as the loss of morphological information and training data density outweighs the benefits of a narrower decision space.

7 Conclusion

A comprehensive study of Arabic medical question classification comparing flat fine-tuning, hierarchical decomposition, and ensemble methods is presented. The best system, fine-tuned MARBERTv2 with Arabic text normalization, achieves 0.3659 macro F1 on the test set. Contrary to expectations, hierarchical classification underperforms (F1: 0.332) due to error cascade from 30% first-stage errors. Severe class imbalance (85:1 ratio) is identified as the primary obstacle, with minority classes achieving near-zero F1.

Future work should explore focal loss variants for extreme imbalance, domain-adaptive pretraining on Arabic medical corpora, and data augmentation for minority categories. The findings suggest that for extreme multi-class tasks with noisy annotations, simple discriminative fine-tuning with proper preprocessing outperforms complex hierarchical architectures.

Acknowledgments

The organizers of the AbjadMed shared task are thanked for providing the dataset and evaluation infrastructure.

References

Muhammad Abdul-Mageed and 1 others. 2021. MARBERT: Deep bidirectional transformers for arabic sentiment analysis. In *ACL-IJCNLP 2021*.

Wissam Antoun and 1 others. 2020. AraBERT: Transformer-based model for arabic language understanding. In *LREC 2020*.

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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Carlos N. Silla and Alex A. Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1):31–72.

Jesse Zhang, Haonan Yu, and Wei Xu. 2021. Hierarchical reinforcement learning by discovering intrinsic options. In *International Conference on Learning Representations*.

A Appendix

1.1 Medical Category Hierarchy

Table 3 presents the complete mapping of 82 fine-grained medical categories into 12 broad clinical domains, as defined in the hierarchical classification experiments. This taxonomy was constructed to group semantically related specialties and address the extreme label space cardinality.

1.2 Per-Category Performance

Category	Support	Precision	Recall	F1
Hematological diseases	600	0.77	0.90	0.83
Pharmacology	600	0.72	0.92	0.81
Benign and malignant tumors	510	0.80	0.78	0.79
Dental diseases	600	0.72	0.78	0.75
Women's health	600	0.69	0.72	0.71
Mid-Resource				
Urogenital diseases	50	0.56	0.76	0.64
Medicinal herbs	35	0.20	0.06	0.09
Low-Resource				
Biochemistry	< 10	0.00	0.00	0.00
IVF	< 10	0.00	0.00	0.00
Anesthesiology	< 10	0.00	0.00	0.00

Table 4: Representative per-category results showing performance disparity between high and low resource categories.

Table 4 shows representative per-category results demonstrating the strong correlation between support (number of training samples) and F1 performance. High-resource categories like Hematological diseases achieve high F1, while low-resource categories suffer significant performance drops.

1.3 Hyperparameter Configuration

Table 5 details the specific hyperparameters used for the main experimental settings.

1.4 Confusion Matrix Analysis

Figure 2 visualizes the confusion matrix for the most frequently misclassified category pairs, highlighting the semantic overlaps discussed in the error analysis.

Broad Domain	Fine-Grained Categories
Surgical Specialties	General surgery, Orthopedic surgery, Plastic surgery, Neurosurgery, Cardiothoracic surgery, Urological surgery, Pediatric surgery, Jaw and dental surgery, Vascular surgery, Gynecologic surgery
Internal Medicine & Chronic	Cardiovascular diseases, Respiratory diseases, Diabetes, Hypertension, Endocrine diseases, Gastrointestinal diseases, Rheumatic diseases, Internal medicine diseases, Hematological diseases
Women's & Reproductive	Women's health, Pregnancy and childbirth, Infertility, Gynecological diseases, Embryology, In vitro fertilization (IVF)
Children's Health	Child health, Pediatric diseases
Men's & Sexual Health	Men's health, Sexual health, Sexually transmitted diseases
Mental Health & Neurology	Psychiatric diseases, Mental health, Psychology, Neurological diseases
Skin, Dental & ENT	Dermatological diseases, Skin and beauty, Dental diseases, Oral diseases, Dental health, Dentistry, Ear nose and throat (ENT), Eye diseases, Optometry
Infectious & Immune	Infectious diseases, Allergy, Immunology, Vaccines and immunizations
Oncology	Benign and malignant tumors
Basic Sciences	Anatomy, Biology, Biochemistry, Physiology, Genetics, Microbiology, Pathology, Radiology, Laboratory, Diagnosis
Pharmacology	Pharmacology, Medicinal herbs, Alternative medicine, Vitamins and minerals, Toxicology, Chemistry, Hormones, Anesthesiology
General Health	Nutrition, Public health, Physiotherapy, Preventive medicine, Medical services, First aid, Health and sports, Geriatric health, Ramadan, Addiction, Congenital malformations, Genetic diseases, Musculoskeletal and joint diseases, Urogenital diseases, General medicine, History of medicine

Table 3: The two-level hierarchy mapping 82 fine-grained categories to 12 broad medical domains used in the hierarchical classification approach.

Model/Experiment	LR	Batch	Epochs	Max Len	Other
MARBERTv2 (Baseline)	2e-5	16	4	128	AdamW
MARBERTv2 (Optimized)	6e-5	32	15	128	WD=0.1
Hierarchical (Stage 1)	2e-5	32	10	128	WD=0.01
Hierarchical (Stage 2)	2e-5	32	15	128	WD=0.01
CAMeLBERT-DA	6e-5	32	15	128	WD=0.1
LightGBM	0.05	–	3000	–	Leaves=100

Table 5: Hyperparameter settings for the various models and stages reported in the experiments. Abbreviations: LR (Learning Rate), WD (Weight Decay).

Figure 2: Confusion matrix of the top misclassified category pairs.