Sakinah-AI at MentalQA: A Comparative Study of Few-Shot, Optimized, and Ensemble Methods for Arabic Mental Health Question Classification

Fatimah Emad Eldin, Mumina Abukar
Cairo University, The University of South Wales
{12422024441586@pg.cu.edu.eg, 74108361@students.southwales.ac.uk}

Abstract

This paper details the system developed by team Sakinah-AI for the MentalQA 2025 shared task, focusing on Arabic mental health question classification. We compare few-shot learning with Large Language Models against fine-tuning of BERT-based models (CAMeL-BERT and AraBERTv2). Few-shot learning with Palmyra-Med-70B achieved the highest weighted F1-score of 0.605, followed by hyperparameter-optimized CAMeL-BERT at 0.597. Notably, 5-fold ensemble methods proved detrimental to performance. Our results demonstrate that for low-resource specialized domains, both few-shot learning and optimized fine-tuning of appropriate base models outperform ensemble strategies. To ensure reproducibility all experimental code and final fine-tuned models are made publicly available.

1 Introduction

Arabic mental health NLP faces unique challenges due to limited annotated data and the linguistic complexity of user-generated content on mental health platforms. To address these challenges, we participated in the MentalQA 2025 shared task (Alhuzali et al., 2024), conducting a systematic comparison of three paradigms for Arabic mental health question classification: few-shot learning with large language models, optimized fine-tuning, and ensemble methods.

Our comparative study reveals critical insights for low-resource specialized domains. Few-shot learning with Palmyra-Med-70B (Kamble and Alshikh, 2023) achieved optimal performance (0.605 weighted F1-score), closely followed by hyperparameter-optimized CAMeL-BERT (0.597). Notably, CAMeL-BERT significantly outperformed AraBERTv2 (0.543), while k-fold ensemble methods proved detrimental to both models' performance. These findings challenge conventional wisdom that ensemble methods universally improve classification accuracy.

The results demonstrate that for small, specialized datasets, strategic model selection and optimization outweigh complex ensembling strategies. Domain-specific pre-training (Palmyra-Med) and careful hyperparameter tuning emerge as more effective approaches than aggregating multiple weak learners. To ensure reproducibility and facilitate future research, we provide open access to all experimental code and fine-tuned models via GitHub[1] and Hugging Face[2].

2 Background and Related Work

2.1 Task Overview and Dataset

The MentalQA 2025 shared task (Alhuzali et al., 2025) focuses on multi-label classification of Arabic mental health questions into seven categories: Diagnosis, Treatment, Anatomy/Physiology, Epidemiology, Healthy Lifestyle, Provider Choices, and Other. We participated in Track 1, Sub-Task 1, using a dataset of 500 annotated question-answer pairs (300 training, 50 development, 150 test) from Arabic mental health platforms characterized by informal, dialect-rich language.

2.2 Arabic Mental Health NLP Evolution

Early foundational work by Alghamdi et al. (2020) created the Arabic psychological forum corpus "Nafsany" and compared lexicon-based approaches against traditional machine learning models. Alasmari (2025) revealed a clear paradigm shift: pre-2022 studies relied on traditional machine learning and lexicon-based methods, while post-2022 research shifted towards transformer-based models like AraBERT (Antoun et al., 2020) and MARBERT, which consistently outperform traditional approaches.

[1] https://github.com/astral-fate/MentalQA2025/
[2] https://huggingface.co/collections/FatimahEmadEldin/sakinah-ai-at-mentalqa-689b2d707791cea458e97aaf

Proceedings of The Third Arabic Natural Language Processing Conference, pages 140–148
November 8-9, 2025

Alhuzali and Alasmari (2025) conducted comprehensive evaluation of Arabic PLMs on the MentalQA dataset, demonstrating that fine-tuned MARBERT achieved superior performance with Jaccard scores of 0.80 for question classification and 0.86 for answer classification, while few-shot learning with GPT-3.5 showed significant improvements over zero-shot approaches. Recent LLM evaluations by Zahran et al. (2025) across eight models on diverse Arabic mental health datasets found that prompt design is critical and few-shot techniques consistently improve performance. Practical applications include the "MindWave" app by Bensalah et al. (2024), which leverages AI for bilingual mental health support.

2.3 Research Gaps and Contribution

Despite progress, gaps remain: limited comparative studies between fine-tuning and few-shot approaches in Arabic mental health domains, insufficient evaluation of ensemble methods versus optimized single models in low-resource settings, and lack of systematic analysis comparing domain-specific versus general-purpose LLMs. Our work addresses these gaps by providing direct comparative evaluation between fine-tuning BERT-based models (CAMeL-BERT and AraBERTv2) and few-shot learning with large language models, systematically evaluating ensemble strategies against optimized single models in the low-resource MentalQA 2025 shared task setting.

3 Methodology

3.1 System Overview

Our system comprises two parallel pipelines for multi-label Arabic mental health question classification: Fine-Tuning and Few-Shot Learning (Figure 1). This design enables direct comparison between traditional supervised learning and contemporary in-context learning paradigms.

3.2 Fine-Tuning Pipeline

3.2.1 Base Model Selection

We selected two Arabic BERT variants with complementary strengths:

CAMeL-BERT-DA-Sentiment (Inoue et al., 2021): A specialized variant fine-tuned for sentiment analysis on Arabic dialectal text. We hypothesized its exposure to user-generated content would benefit processing informal mental health questions.

AraBERTv2 (Antoun et al., 2020): A widely-adopted baseline model for Arabic NLP tasks, providing robust comparison benchmarks.

3.2.2 Training Strategies

Optimized Single Models: We employed Optuna framework for automated hyperparameter optimization, systematically exploring learning rates (le-5 to 5e-5), batch sizes (8, 16), and epochs (10-20) to identify optimal configurations. The final hyperparameters used for the CAMEL-BERT model are detailed in Appendix B (Table 5).

K-Fold Ensembles: We trained five models using stratified cross-validation and averaged their predictions. This approach tests whether model diversity improves performance in low-resource settings.

3.2.3 Model Selection Rationale

We selected models to test three factors: domain specialization, architecture, and scale. Palmyra-Med-70B (Kamble and Alshikh, 2023) provides medical domain expertise. Mixtral-8x22B uses mixture-of-experts architecture, while Qwen3-235B represents dense transformers. Gpt-Oss-20B tests the lower performance boundary (20B parameters), and Colosseum-355B tests the upper boundary (355B parameters). This design isolates whether domain knowledge, architectural differences, or parameter scaling most impacts Arabic mental health classification. All models support Arabic and are accessible via NVIDIA NIM API.

3.3 Few-Shot Learning Pipeline

3.3.1 Model Selection Rationale

We selected models testing domain specialization (Palmyra-Med-70B), architecture (Mixtral-8x22B mixture-of-experts vs. Qwen3-235B dense transformer), and scale boundaries (Gpt-Oss-20B at 20B, Colosseum-355B at 355B parameters). All models support Arabic and are accessible via NVIDIA NIM API.

3.3.2 Prompt Engineering

We constructed structured prompts with: (1) explicit multi-label task instructions, (2) Arabic category definitions and examples, and (3) 3-5 diverse training exemplars. Models were explicitly instructed to "select ALL applicable categories" with multi-label demonstrations.

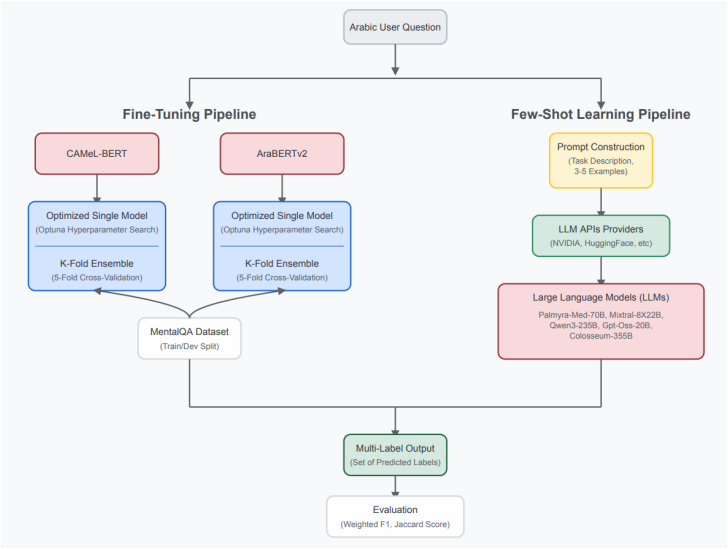

Figure 1: The Sakinah-AI System Architecture, illustrating two parallel processing pipelines.

3.4 Experimental Design

Our study follows a controlled comparison framework. For fine-tuning, we used 300 training samples with 50-sample development sets for hyperparameter optimization. For ensembles, we combined training and development sets (350 samples) for 5-fold cross-validation. Few-shot experiments used 3-5 training examples as in-context demonstrations. This design enables fair comparison across paradigms while addressing the low-resource constraints typical of specialized Arabic NLP domains.

4 Experimental Setup

4.1 Comparative Analysis Framework

We conduct a systematic comparison of three paradigms for Arabic mental health question classification: optimized fine-tuning, few-shot learning, and ensemble methods. This controlled evaluation addresses a critical research question: which approach performs best in low-resource specialized domains where traditional assumptions about ensemble superiority may not hold.

4.2 Data Configuration

The 500-sample dataset was partitioned into 300 training, 50 development, and 150 test samples. While this small size presents overfitting risks typical of specialized domains, we implement several mitigation strategies:

Fine-Tuning Protocol: Training set for model optimization, development set for hyperparameter selection, with early stopping based on development performance.

Ensemble Strategy: Combined training/development sets (350 samples) for stratified 5-fold cross-validation to maximize training data while maintaining validation integrity.

Few-Shot Design: Minimal training exposure (3-5 examples) inherently reduces overfitting risk while testing generalization from limited demonstrations. All final evaluations use the held-out test set to ensure unbiased performance estimates.

4.3 Evaluation Metrics

The primary evaluation metric is the weighted F1-score, which accounts for label imbalance (Sokolova and Lapalme, 2009). We additionally consider the Jaccard Score for multi-label evaluation (Manning et al., 2008).

Weighted F1-Score For a set of labels L, the weighted F1-score is calculated as:

$$\text{Weighted F1} = \sum_{l \in L} w_l \cdot F1_l \tag{1}$$

where w_l represents the proportion of instances of label l in the dataset, and $F1_l$ denotes the F1-score for that label, calculated as:

$$F1_l = 2 \cdot \frac{\text{Precision}_l \cdot \text{Recall}_l}{\text{Precision}_l + \text{Recall}_l} \tag{2}$$

Jaccard Score For individual predictions, where Y_{true} represents the set of true labels and Y_{pred} represents the set of predicted labels, the Jaccard score

is:

$$J(Y_{\text{true}}, Y_{\text{pred}}) = \frac{|Y_{\text{true}} \cap Y_{\text{pred}}|}{|Y_{\text{true}} \cup Y_{\text{pred}}|} \quad (3)$$

The overall score represents the average Jaccard score across all samples.

5 Results

Our evaluation, conducted on the blind test set, reveals a distinct performance hierarchy among the different modeling paradigms. As shown in Table 1, the few-shot approach with a domain-specific LLM (Palmyra-Med-70B) achieved the highest weighted F1-score of 0.605. Closely following was the single, hyperparameter-optimized fine-tuned model, CAMeL-BERT (Opt.), with a score of 0.597. These top performers significantly outpaced all other models, particularly the ensemble variants, which consistently underperformed their single-model counterparts.

5.1 Error Analysis and Performance Patterns

To better understand these results, we conducted a detailed error analysis for both fine-tuned and few-shot models. A comprehensive quantitative and qualitative breakdown of model performance is available in Appendix C.

5.1.1 Fine-Tuned Model Analysis

As detailed in Table 2, the optimized CAMeL-BERT model maintains the lowest error counts across most categories, confirming its robustness. In contrast, the AraBERTv2 ensemble suffered a catastrophic performance collapse, with error counts surging in categories like **Anatomy and Physiology** (140 errors), **Other** (147 errors), and **Provider Choices** (122 errors). This pattern suggests that for smaller, specialized datasets, ensembling can amplify systematic model biases rather than mitigate variance, leading to degraded performance.

5.1.2 LLM Performance and Multi-Label Challenges

The error analysis for LLMs (Table 3) shows that Palmyra-Med-70B maintained a more balanced error profile compared to other models, which struggled significantly in high-support categories like **Diagnosis** and **Treatment**. A critical qualitative finding was the LLMs' systematic failure to adhere to multi-label instructions. Our prompt engineering (detailed in Appendix A Table 4) was specifically designed to prevent this by including: (1) explicit instructions to "perform precise multi-label classification" and "select ALL applicable categories," (2) clear examples of multi-label outputs (e.g., "Final Answer: A,D"), and (3) a structured format. Despite these safeguards, all tested LLMs frequently defaulted to predicting only a single label, even for questions where multiple categories were clearly relevant. This suggests a fundamental limitation in current instruction-following capabilities for complex classification tasks, possibly stemming from strong priors developed during pre-training on predominantly single-output tasks. This limitation likely suppressed the overall performance of all LLMs in our study.

5.2 Key Insights from Comparative Analysis

Domain Expertise vs. General Capability. The superior performance of Palmyra-Med-70B (0.605) over the much larger, general-purpose Qwen3-235B (0.325) highlights the profound value of domain-specific pre-training. Palmyra-Med's focused medical knowledge provided a decisive advantage in correctly interpreting the nuanced language of mental health questions, demonstrating that for specialized tasks, domain expertise can be more critical than model scale alone.

The Failure of Ensemble Methods. The consistent underperformance of k-fold ensembles challenges the conventional wisdom that they universally improve model robustness. For CAMeL-BERT, the ensemble F1-score (0.537) was notably lower than the optimized single model (0.597). The degradation was even more severe for AraBERTv2 (0.328 vs. 0.543). This outcome suggests that in low-resource settings, where individual models are trained on limited and potentially noisy data, they may develop high bias. In such cases, ensembling methods like averaging predictions can amplify these shared systematic errors rather than reducing variance, ultimately harming overall performance.

6 Discussion

Our results yield several key insights for specialized, low-resource domains. The superior performance of Palmyra-Med-70B (0.605) and optimized CAMEL-BERT (0.597) demonstrates that domain-specific pre-training and strategic single-model optimization are more effective than ensembling for Arabic mental health question classification. The consistent failure of our k-fold ensembles challenges the conventional wisdom that they univer-

Fine-Tuning	
Model Name	**Weighted F1-Score**
CAMeL-BERT (Optimized)	**0.597**
AraBERTv2 (Optimized)	0.543
CAMeL-BERT (K-Fold Ensemble)	0.537
AraBERTv2 (K-Fold Ensemble)	0.328

(a) Fine-Tuning Models

Few-Shot Learning	
Model Name	**Weighted F1-Score**
Palmyra-Med-70B	**0.605**
Mixtral-8X22B	0.563
Qwen3-235B	0.325
Gpt-Oss-20B	0.147
Colosseum-355B	0.014

(b) Few-Shot Learning Models

Table 1: Final results on the test set, comparing fine-tuned models against few-shot learning with LLMs. Optimized single models and domain-specific LLMs demonstrate superior performance.

Category	**CAMeL-BERT**		**AraBERTv2**	
	Opt.	**Ens.**	**Opt.**	**Ens.**
Anatomy	31	**18**	11	140
Diagnosis	**55**	71	53	65
Epidemiology	96	**85**	39	55
Lifestyle	**57**	102	44	38
Other	**3**	52	3	147
Provider	**31**	76	6	122
Treatment	**66**	**66**	63	85

Table 2: Error counts per category for all fine-tuned models. Lower values indicate better performance. Errors are calculated as Support $\times$ (1 - Recall).

Category	Palmyra	Mixtral	Qwen3	GPT-Oss	Colosseum
Anatomy	20	17	**10**	12	**10**
Diagnosis	**49**	52	67	74	84
Epidemiology	49	42	**37**	40	35
Lifestyle	**36**	38	39	39	37
Other	**3**	5	5	**3**	**3**
Provider	11	**9**	6	7	6
Treatment	50	**49**	66	81	85

Table 3: Error counts per category for few-shot LLMs. Lower values indicate better performance.

sally reduce errors. From a bias-variance perspective, ensembles are most effective at reducing variance by averaging the uncorrelated errors of diverse base learners. However, in low-resource settings with a small and specialized dataset, this core assumption is violated. The models trained on different folds of the data are not sufficiently diverse; instead, they learn similar systematic biases from the limited data. Consequently, the ensemble averages and reinforces these shared biases rather than canceling out random errors, leading to a notable degradation in performance, as seen with both CAMEL-BERT and AraBERTv2. While this study operated within the constraints of the provided dataset, future work could address these data limitations through several mitigation strategies. Data augmentation techniques, such as back-translation or contextual synonym replacement tailored to Arabic dialects, could create novel training instances. Furthermore, semi-supervised learning approaches could be employed to leverage vast amounts of unlabeled, in-domain text. By training a model on the existing labeled data and using it to generate pseudo-labels for unlabeled data, the training set could be effectively and cheaply expanded. A final significant finding was the LLMs' systematic failure to adhere to multi-label instructions despite explicit prompting, highlighting fundamental limitations in current instruction-following capabilities.

7 Conclusion

This paper presented the Sakinah-Al system for the MentalQA 2025 shared task, comparing few-shot learning, optimized fine-tuning, and ensemble methods for Arabic mental health question classification. Our results show that a domain-specific LLM, Palmyra-Med-70B, achieved the highest weighted F1-score (0.605), closely followed by an optimized CAMEL-BERT model (0.597). Notably, ensemble methods were detrimental to performance in this low-resource setting. The primary limitations of our study include the LLMs' difficulties with multi-label adherence and the small size of the training dataset. Furthermore. Future assessments must incorporate crucial dimensions such as clinical relevance and safety considerations to prevent harmful or inaccurate outputs. Moreover, focusing on model interpretability will be essential to build trust and utility for clinicians and end-users. Future work should explore advanced prompt engineering and data augmentation techniques while embedding these human-centered principles into the evaluation process.

Acknowledgments

We thank the organizers of the MentalQA 2025 shared task for their support and assistance in providing the dataset, evaluation framework, and coordination that made this research possible.

References

Ashwag Alasmari. 2025. A scoping review of arabic natural language processing for mental health. *Healthcare*, 13(9):963.

Norah S. Alghamdi, Hanan A. H. Mahmoud, Ajith Abraham, Samar A. Alanazi, and Laura García-Hernández. 2020. Predicting depression symptoms in an arabic psychological forum. *IEEE Access*, 8:57317–57334.

Hassan Alhuzali, Aseel Alasmari, and Hajar Alsaleh. 2024. Mentalqa: An annotated arabic corpus for questions and answers of mental healthcare. *IEEE Access*.

Hassan Alhuzali and Ashwag Alasmari. 2025. Pretrained language models for mental health: An empirical study on arabic q&a classification. *Healthcare*, 13(9):985.

Hassan Alhuzali, Farah Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu-Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, and 1 others. 2025. AraHealthQA 2025 shared task description paper. *arXiv preprint arXiv:2508.20047*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15, Marseille, France. European Language Resources Association.

Nouhaila Bensalah, Habib Ayad, Abdellah Adib, and Abdelhamid Ibn El Farouk. 2024. Mindwave app: Leveraging ai for mental health support in english and arabic. In *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)*, Marrakech, Morocco. IEEE.

Go Inoue, Bashar Al-Khafaji, 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 15–28.

Kiran Kamble and Waseem Alshikh. 2023. Palmyra-med: Instruction-based fine-tuning of llms enhancing medical domain performance. Preprint available on ResearchGate. Under review.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Noureldin Zahran, Aya E. Fouda, Radwa J. Hanafy, and Mohammed E. Fouda. 2025. A comprehensive evaluation of large language models on mental illnesses in arabic context. arXiv preprint. *Preprint*, arXiv:2501.06859.

A Few-Shots examples

The prompt used for all Large Language Model (LLM) evaluations was engineered to facilitate precise multi-label classification for Arabic mental health questions. As detailed in Table 4, the prompt architecture consists of four key components: a system prompt establishing an expert persona, a complete list of all seven categories with definitions, two diverse few-shot examples demonstrating the reasoning process and required multi-label output format (e.g., "Final Answer: A,D"), and a final task instruction for the target question. This structure was explicitly designed to guide the models in selecting all applicable categories and to counteract the observed tendency of LLMs to default to single-label outputs.

B Fine-Tuning Hyperparameters

The fine-tuning of the `CAMeL's bert-base-arabic-camelbert-mix-sentiment` model was conducted using the hyperparameters detailed in Table 5. These settings were configured using the Hugging Face Transformers library.

C Detailed Performance Analysis

This appendix provides a detailed quantitative and qualitative analysis of the top-performing models, based on the output from the error analysis script.

C.1 Quantitative Performance Summary

The table below summarizes the key performance metrics for the selected models. Palmyra-Med-70B demonstrates the best overall performance, closely followed by the optimized single model, CAMEL-BERT. The AraBERTv2-Ensemble model shows a significant degradation in performance across all metrics.

C.2 Per-Category F1-Score Matrix

To understand model performance on a more granular level, the following matrix presents the F1-score for each of the seven classification categories.

Component	Content
System Prompt	You are an expert in classifying Arabic patient questions into mental health categories. Perform precise multi-label classification.
Category List	**(A) Diagnosis:** Interpreting symptoms. **(B) Treatment:** Seeking therapies or medications. **(C) Anatomy and Physiology:** Basic medical knowledge. **(D) Epidemiology:** Course, prognosis, causes of diseases. **(E) Healthy Lifestyle:** Diet, exercise, mood control. **(F) Provider Choices:** Recommendations for doctors. **(Z) Other:** Does not fit other categories.
Example 1	**Question:** هل يعتبر الخوف من عدم الإنجاب مستقبلاً حالة عادية خاصةً لما أكون متعلقة بأطفال كثيراً وأنا على وشك الزواج أنا خائفة جداً **Reasoning:** The user is asking if their fear (a symptom) is normal and is concerned about its future course (prognosis). This fits 'Diagnosis' (interpreting a symptom) and 'Epidemiology' (prognosis). **Final Answer:** A,D
Example 2	**Question:** من سنة تقريباً وأنا أؤذي نفسي بأكثر من طريقة وما أعرف كيف أتخلص من هذه العادة، وبدأت تجيني أفكار بإنهاء حياتي وحاولت أنتحر بأكثر من مرة وأكثر من طريقة **Reasoning:** The user describes self-harm and suicidal thoughts and is asking how to get rid of this habit. This is a clear call for 'Treatment' (seeking therapy/help) and relates to 'Healthy Lifestyle' (self-help, mood control). **Final Answer:** B,E
Task	Classify the following question. Provide your reasoning and then the final answer. **Question:** {Target Question} **Reasoning:** **Final Answer:**

Table 4: Structure and content of the few-shot prompt used for LLM inference.

Hyperparameter	Value
Model & Tokenizer	
Base Model	CAMeL-BERT (mix-sentiment)
Max Sequence Length	256
Training Arguments	
Epochs	15
Batch Size	8
Gradient Accum. Steps	2
Learning Rate	2e-5
Warmup Steps	100
Weight Decay	0.01
Optimizer	AdamW
FP16 Precision	True
Loss Function	
Loss Type	Focal Loss
Alpha (α)	1.0
Gamma (γ)	2.0

Table 5: Hyperparameters for the optimized fine-tuning of CAMeL-BERT.

Metric	Palmyra-Med-70B	CAMEL-BERT Opt	AraBERTv2 Ens.
Exact Match Ratio	12.67%	11.33%	0.00%
Macro Jaccard Score	0.2623	0.2445	0.1115
Weighted F1-Score	0.60	0.59	0.26

Table 6: Overall performance metrics on the blind test set.

Both Palmyra-Med and CAMEL-BERT perform strongly on high-support categories like Diagnosis (A) and Treatment (B), while the Ensemble model fails completely on Treatment and Healthy Lifestyle questions.

C.3 Error Analysis Matrix

The following examples from the test set illustrate common failure modes for different models, highlighting the challenges of multi-label classification and the pitfalls of ensembling in low-resource settings.

Category	Palmyra-Med-70B	CAMEL-BERT Opt	AraBERTv2 Ens.
(A) Diagnosis	0.75	0.76	0.71
(B) Treatment	**0.74**	0.70	0.00
(C) Anatomy/Phys.	0.09	**0.15**	0.12
(D) Epidemiology	**0.44**	0.37	0.18
(E) Healthy Lifestyle	0.38	**0.41**	0.00
(F) Provider Choices	**0.15**	0.00	0.09
(Z) Other	0.00	0.00	**0.04**

Table 7: Per-category F1-scores for each model. Higher is better.

Error Type	Question & Analysis	Labels
Multi-Label Failure (Palmyra-Med-70B)	**Question:** بكاء مفاجى ،حزن ، فقدان الوزن بدون سبب، الاكتئاب، عدم الثقة بالنفس، القلق ، التوتر ، الانطوائية، **Analysis:** The user lists numerous symptoms ('A'), is implicitly asking for a solution ('B'), and is concerned about the course of the illness ('D'). The model correctly identifies 'Diagnosis' but fails to capture the other required labels.	**True:** A, B, D **Predicted:** A
Ensemble Hallucination (AraBERTv2 Ensemble)	**Question:** ماهو افضل دواء منوم وذاا تأثير سريع وقوي لاني اعاني من ارق ولا استطيع النوم ابداا الرجاء الاجابه؟؟ **Analysis:** A direct question about medication ('B'). The ensemble model not only misses the correct label entirely but also hallucinates four incorrect and irrelevant labels, demonstrating a catastrophic failure.	**True:** B **Predicted:** A, C, F, Z
Domain Specialization (CAMEL-BERT Opt)	**Question:** رجاء أريد معلومات عن دواء جديد اسمه ايشل پلنڤرن شكرا **Analysis:** This is a clear request for information about a specific treatment ('B'). The optimized CAMEL-BERT model, attuned to user-generated dialectal content, correctly classifies this. The log shows that the baseline AraBERT model failed to produce any prediction for this item, highlighting the robustness of the optimized model.	**True:** B **Predicted:** B

Table 8: Illustrative examples of misclassification cases from the test set.