Binary_Bunch at AraHealthQA Track 1: Arabic Mental Health Q&A Classification Using Data Augmentation and Transformer Models

Sajib Bhattacharjee*, **Ratnajit Dhar***, **Kawsar Ahmed and Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
u2004003@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

Mental health question-answering (MentalQA) is essential for delivering accessible and reliable mental health support. Natural language processing (NLP) techniques are increasingly integral to such systems, enabling automated categorization of questions and answers to improve information retrieval, response accuracy, and user guidance. In AraHealthQA 2025 (Track 1), we addressed two subtasks: multi-label question categorization and answer categorization. We proposed an XLMR-Arabic pipeline enhanced with a two-stage data augmentation strategy, combining large language model (LLM)-based paraphrasing with synthetic label merging. Additionally, we evaluated the effectiveness of fine-tuned multilingual transformers, LLMs adapted with low-rank adaptation (LoRA), and LLMs under few-shot settings. Experimental results show that XLMR-Arabic achieved the best performance, reaching Jaccard scores of 53% and 77.44% on Subtasks 1 and 2, respectively, ranking our team second in both tracks.

1 Introduction

Automatic Question Answering (QA) systems are AI applications that process natural language queries and deliver precise, context-specific answers using natural language processing and information retrieval methods. The development of QA systems for Arabic presents significant challenges due to its complex morphology, flexible syntax, dialectal variation, limited annotated resources, and high lexical ambiguity resulting from the absence of diacritics. Mental health represents a global priority with substantial impacts on both individual and societal well-being. In Arabic-speaking areas, mental health services are limited and stigma is prevalent, especially among religious and community leaders. Automatic classification of mental health questions is critical within the mental health support pipeline. Accurate identification of user intent and question type enables systems to route queries to appropriate resources or generate effective, targeted responses. Automatic QA systems in the mental health domain facilitate rapid, accurate, and accessible information retrieval, thereby supporting decision-making, education, and global knowledge dissemination.

To address these challenges, we participated in the AraHealthQA 2025 shared task (Alhuzali et al., 2025), focusing on Track 1: MentalQA 2025, specifically Subtask 1 (Question Categorization) and Subtask 2 (Answer Categorization). To solve the tasks, this work employs a two-stage data augmentation strategy, expanding the dataset through LLM-based paraphrasing and multi-label merging. Transformer-based models were fine-tuned, and both few-shot learning and fine-tuned LLMs were evaluated. The main contributions of this work are as follows:

- We propose a two-stage data augmentation strategy, combining LLM-based paraphrasing and synthetic label merging, to address the challenge of limited training data in both subtasks.
- We systematically evaluate a range of transformer-based models and LLMs under fine-tuning, LoRA, and few-shot settings, providing comparative insights into their effectiveness for MentalQA categorization tasks.
- We demonstrate that transformer models, particularly XLMR-Arabic[1] , consistently outperform LLMs in LoRA settings, highlighting the advantages of language-specific

*Authors contributed equally to this work.

[1] https://huggingface.co/Davlan/XLM-Roberta-base-finetuned-arabic

Proceedings of The Third Arabic Natural Language Processing Conference, pages 164–175

specialization and full-parameter fine-tuning compared to compressed adaptation methods.

2 Literature Review

Significant research has been dedicated to leveraging NLP for mental health in the Arabic Language. Early efforts primarily targeted the detection of depression, anxiety, and suicidal ideation in Arabic social media posts, often relying on handcrafted lexicons or classical machine learning pipelines before transitioning toward transformer-based architectures (Rabie et al., 2025; Almeqren et al., 2023; Alasmari, 2025). Alsmadi, 2024 proposes DeBERTa-BiLSTM for multi-label classification of Arabic medical questions (COVID-19 FAQs), reporting strong micro-F1. A study by Abdulsalam et al., 2023 developed an Arabic dataset of suicidal tweets and demonstrated that pre-trained deep learning models, particularly AraBERT (Antoun et al., 2020), outperform traditional machine learning approaches in detecting suicidal ideation on social media. Elmajali and Ahmad, 2024 classified depression symptoms in Arabic tweets according to the DSM-5 using AraBERT and MARBERT (Abdul-Mageed et al.), achieving over 98% accuracy across multiple metrics after balancing the dataset with ChatGPT-generated augmentation. Building on the MentalQA dataset, Alhuzali and Alasmari, 2025 compared traditional machine learning, Arabic-specific PLMs, and prompt-based methods for classifying mental health questions and answers, reporting top performance with MARBERT and notable gains from few-shot GPT-3.5 (Brown et al.) prompting. Abu Daoud et al., 2025 introduced MedArabiQ, a benchmark dataset comprising seven Arabic medical tasks, including multiple-choice questions, fill-in-the-blank exercises, and patient-doctor question answering. Previous studies focused on detecting mental health conditions (e.g., depression, anxiety, suicidal ideation) using classical machine learning, AraBERT, MARBERT, or general medical benchmarks. In contrast, we address the multi-label categorization of Arabic mental health questions and answers through a two-stage data augmentation method, combining LLM-based paraphrasing and synthetic label merging, with fine-tuned domain-specific transformers.

3 Dataset and Task Description

The dataset provided for the AraHealthQA 2025 Shared TaskTrack 1 encompasses two subtasks focused on question and answer classification within the Arabic healthcare domain. Both subtasks leverage a shared dataset adopted from Alhuzali et al., 2024.

- **Subtask 1 (Question Classification):** This subtask[2] involved categorizing user-submitted health-related questions into one of six predefined categories. The training set comprised 350 labeled questions, each annotated with its corresponding category label. A separate test set of 150 unlabeled questions was provided for evaluation purposes.
- **Subtask 2 (Answer Classification):** In the second subtask[3], the goal was to classify answers corresponding to the health-related questions into one of three predefined categories. Similar to Subtask 1, the training set consisted of 350 labeled answers, while the test set, used for evaluation, comprised 150 unlabeled answers.

	Datasets	T_S	T_W	T_{UW}	L_{Avg}
ST-1	Original Dataset	350	10783	4306	30.81
	Augmented Dataset	1200	48370	6514	40.31
	Test Dataset	150	4557	2368	30.38
ST-2	Original Dataset	350	10921	4376	31.20
	Augmented Dataset	1200	40050	5607	33.37
	Test Dataset	150	4503	2115	30.02

Table 1: Counts of total samples (T_S), total words (T_W), unique words (T_{UW}), and average sample length (L_{Avg}) for Subtask 1 (ST-1) and Subtask 2 (ST-2) datasets.

4 System Overview

This study evaluates four transformer models and two LLMs using fine-tuning and few-shot learning across both subtasks. To address the limited size and diversity of the dataset, data augmentation strategies were implemented. Experimental results indicate that transformer-based models consistently outperformed alternative approaches. Figure 1 presents the architecture of the system. The implementation and source code are publicly available on GitHub[4].

4.1 Data Augmentation

The original dataset contained 350 samples for each subtask, which was insufficient to train large models. To address this, we employed a two-stage

[2] https://www.codabench.org/competitions/8559/

[3] https://www.codabench.org/competitions/8730/

[4] https://github.com/Sojib001/AraHealthQA-QA_Categorization

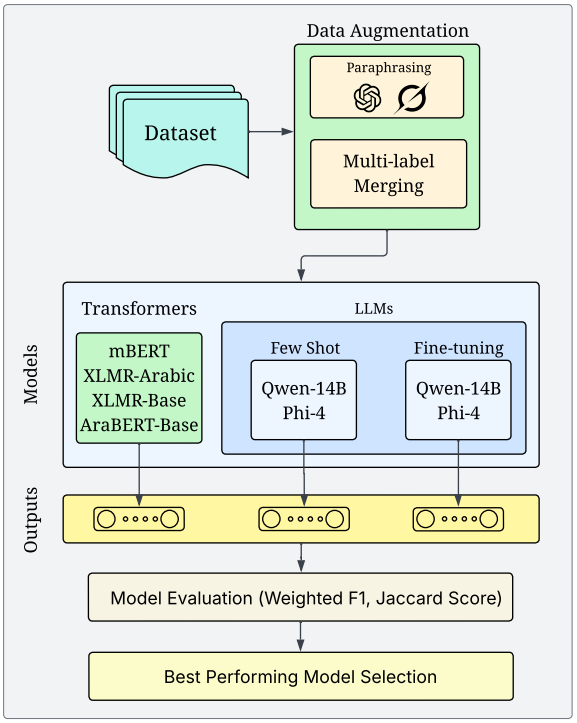


Figure 1: Abstract representation of our methodology pipeline, including data augmentation, transformer, and LLM-based approaches, and model evaluation.

data augmentation strategy to expand and diversify the dataset. This increased the training set to 1,200 samples per subtask, helping the models generalize better and become more robust.

- **LLM-based Paraphrasing:** In the first step, we used LLMs to generate a paraphrased version of each sample. We utilized Grok-3[5] and GPT-4 (Achiam et al., 2023) to generate a paraphrased version of each question and answer, preserving their original meaning and labels. This doubled the dataset from 350 to 700 samples per subtask. We ensured Grok-3 and GPT-4 paraphrases preserved meaning by using carefully designed prompts that emphasized maintaining the original intent, and by validating paraphrases against their original category labels to avoid semantic drift. This guaranteed lexical diversity while keeping semantic fidelity in sensitive mental health queries. The prompt used for data augmentation is provided in Appendix A.8.

- **Multi-label Merging:** In the second stage, we combined two randomly chosen samples from the original dataset to create a new sample. We also merged their labels by taking all the labels from both samples. This method helped us create more complex multi-label examples. With this approach, we added 500 new samples per subtask, bringing the total to 1,200 samples. Example of multi-label merging has be shown in A.5

4.2 Encoder-only Models

Four pre-trained transformer models were utilized for multi-label classification in both subtasks, including XLMR-Arabic, AraBERT-Base[6], mBERT (Devlin et al.), and XLMR-Base (Conneau et al., 2019). All models were fine-tuned on the augmented dataset, with XLMR-Arabic consistently achieving the best performance across both subtasks.

4.3 Decoder-only Models

We employed two state-of-the-art multilingual and multitasking LLMs: Phi-4 (Abdin et al., 2024) and Qwen-14B (Yang et al., 2025). Both models were evaluated under few-shot learning and fine-tuning settings across the two subtasks.

- **Few-shot Learning:** We evaluated Qwen-14B and Phi-4 within the UnSloth framework using five-shot prompting. These models were selected for their strong reasoning and instruction-following capabilities and their compatibility with prompt-based pipelines. Despite their flexibility, performance remained below that of fine-tuned transformer baselines.

- **Fine-tuning:** We further fine-tuned Qwen-14B and Phi-4 on the augmented dataset, framing multi-class classification as a supervised generation task. Inputs consisted of raw text (questions or answers), and outputs were category labels. Training followed a causal language modeling objective with instruction-style formatting. To improve efficiency, we applied low-rank adaptation (LoRA) (Hu et al., 2022) via the UnSloth framework[7], enabling scalable adaptation of large models to downstream tasks.

Appendix A.1 explains the detailed hyperparameter configurations for both the transformers and LLM fine-tuning approaches.

[5] https://x.ai/news/grok-3

[6] https://huggingface.co/aubmindlab/bert-base-arabert

[7] https://docs.unsloth.ai/

4.4 Model Selection

As presented in Table 4, we conducted an ablation analysis with learning rates of 2×10^{-4}, 2×10^{-5}, and 2×10^{-6} to determine the optimal setting. Among these, XLMR-Arabic achieved superior performance at a learning rate of 2×10^{-5}, consistently outperforming both multilingual baselines and LLMs across both subtasks. Hence, XLMR-Arabic was selected as the final model.

5 Results and Discussion

Table 2 presents the performance of different methods, evaluated using the Jaccard score and the weighted F1 score. The results offer a comparative analysis across the approaches, highlighting their relative strengths and potential limitations.

Models	Approach	Subtask-1		Subtask-2	
		Jacc.	W-F1	Jacc.	W-F1
Transformers					
mBERT	- Aug	45.56	60.85	66.67	77.35
	+ Aug	49.83	63.33	68.11	77.00
	Δ	+4.27	+2.48	+1.44	-0.35
XLMR-Arabic	- Aug	48.56	60.96	70.44	71.74
	+ Aug	53.00	60.00	77.44	71.00
	Δ	+4.44	-0.96	+7.00	-0.74
XLMR-Base	- Aug	47.61	61.17	67.33	78.57
	+ Aug	49.33	62.80	69.44	78.77
	Δ	+1.72	+1.63	+2.11	+0.20
AraBERT-Base	- Aug	47.33	62.73	66.00	75.78
	+ Aug	50.91	62.95	69.67	79.31
	Δ	+3.58	+0.22	+3.67	+3.53
LLMs (Fine Tuned)					
Qwen3-14B	- Aug	42.01	54.73	37.00	53.80
	+ Aug	44.02	59.05	42.44	55.95
	Δ	+2.01	+4.32	+5.44	+2.15
Phi-4	- Aug	48.19	62.66	53.22	63.65
	+ Aug	45.71	58.61	60.44	70.49
	Δ	-2.48	-4.05	+7.22	+6.84
LLMs (Few Shot)					
Qwen 3-14B		44.39	54.29	63.33	73.15
Phi-4		42.16	55.41	65.43	75.16

Table 2: Performance of different methods on Subtask 1 (Question Classification) and Subtask 2 (Answer Classification) using Jaccard Score (Jacc.) and Weighted F1 (W-F1), reported in %.

Data Augmentation Enhanced Performance. Data augmentation substantially improved training diversity and robustness by introducing lexical and syntactic variation through GPT-4 and Grok-3 paraphrasing, as well as by generating more complex examples via synthetic multi-label merging. These strategies enhanced model generalization and yielded notable performance gains. As shown in Table 2, XLMR-Arabic improved by +7.00% Jaccard in Subtask-2, AraBERT-Base by +3.67% Jaccard and +3.53% Weighted-F1, mBERT by +4.27% Jaccard and +2.48% Weighted-F1 in Subtask-1, and Qwen3-14B by +5.44% Jaccard in Subtask-2. Phi-4 demonstrated mixed trends, with declines in Subtask-1 but strong gains in Subtask-2 (+7.22% Jaccard, +6.84% Weighted-F1).

Models	Augment	Subtask-1		Subtask-2	
		Jacc.	W-F1	Jacc.	W-F1
AraBERT-Base	+ pp	47.67	70.95	68.22	78.53
	+ mlm	50.91	62.95	69.67	79.31
	Δ	+3.24	-8.00	+1.45	+0.78
XLMR-Base	+ pp	41.50	60.01	67.78	81.34
	+ mlm	49.33	62.80	69.44	78.77
	Δ	+7.83	+2.79	+1.66	-2.57
XLMR-Arabic	+ pp	49.78	65.42	69.00	78.41
	+ mlm	53.00	60.00	77.44	71.00
	Δ	+3.22	-5.42	+8.44	-7.41

Table 3: Performance of the models using Jaccard (Jacc.) and Weighted F1 (W-F1), reported in %. Here, 'pp' denotes LLM-based paraphrasing and 'mlm' denotes multi-label merging applied after paraphrasing. Jaccard was considered our primary metric of evaluation. Δ indicates the difference (mlm – pp).

Further analysis in Table 3 indicates that applying multi-label merging (mlm) after paraphrasing (pp) generally outperformed paraphrasing alone. For example, AraBERT-Base gained an additional +3.24% Jaccard in Subtask-1 and +1.45% in Subtask-2, while XLMR-Arabic achieved +3.22% and a substantial +8.44% improvement, respectively. XLMR-Base also showed consistent gains (+7.83% and +1.66%). Although some trade-offs were observed in Weighted-F1, the consistent rise in Jaccard scores underscores that multi-label merging enhanced robustness beyond paraphrasing alone.

Transformer Models Outperformed LLMs. In our experiments, fine-tuned transformer-based architectures consistently outperformed LLMs. The transformer model was pre-trained exclusively on Arabic text, enabling optimal tokenization and more substantial alignment with the tasks linguistic characteristics. Moreover, the LLMs instruction-tuned and long-context-optimized objectives added complexity without yielding measurable performance gains in this specific context. In Subtask-2, XLMR-Arabic (+Aug) achieved a 77.44% Jaccard score, outperforming fine-tuned Qwen3-14B (+Aug) and Phi-4 (+Aug) by +35.0% and +17.0%, respectively. In Subtask-1, XLMR-Arabic (+Aug) reached 53.00%, exceeding Qwen3-14B and Phi-4 by +8.98% and +7.29%. This performance gap can be explained by differences in parameter utilization and linguistic specialization. In our setup, Qwen-14B was fine-tuned with LoRA, activating only 34.9M

trainable parameters and further constrained by 4-bit quantization, which reduced numerical precision. In contrast, XLMR-Arabic leveraged its full 278M parameters without compression, allowing more effective learning from the training data. The multilingual and multitask design of Qwen-14B likely diluted its language-specific capacity, contributing to its lower performance relative to XLMR-Arabic.

Arabic Transformers Outperformed Others. XLMR-Arabic achieved the best performance due to fine-tuning on Arabic corpora provided a more substantial inductive bias for capturing the morphological, syntactic, and lexical properties of the language. In contrast, the other transformer variants, such as mBERT and XLMR-Base, were trained on general multilingual data and lacked the same degree of specialization in Arabic, resulting in comparatively lower performance. XLMR-Arabic (+Aug) achieved 77.44% Jaccard score, exceeding mBERT (+Aug) by +9.33% points, while AraBERT-Base (+Aug) reached 69.67% Jaccard score, still outperforming mBERT by +1.56% points. In Subtask-1, XLMR-Arabic (+Aug) also surpassed mBERT (+Aug) by +3.17% Jaccard score, with AraBERT-Base (+Aug) showing a smaller gain of +1.08%. When comparing Arabic models themselves, XLMR-Arabic emerged as the strongest overall, achieving the highest Jaccard scores across both subtasks (53.00% and 77.44%). The details of the evaluation metrics and sample predictions for both subtasks are provided in Appendices A.2 and A.6, respectively. Appendix A.4 illustrates the error analysis of the best-performed model.

6 Conclusion

This study investigates multi-label QA categorization within the Arabic mental healthcare domain. The XLMR-Arabic was employed alongside a two-stage data augmentation strategy that integrates large language model (LLM)-based paraphrasing and synthetic multi-label merging. This methodology resulted in significant improvements in classification performance. The findings suggest that targeted augmentation, combined with Arabic-specific transformer architectures, enhances the understanding of nuanced mental health discourse. Future research could investigate leveraging temporal patterns, conversational context, and cross-lingual transfer to enhance generalization.

Limitations

While this study advances multi-label categorization in Arabic mental health question answering, several limitations remain to be addressed in future work. The dataset is relatively small and does not fully capture the linguistic diversity of Arabic, particularly across regional dialects. Although the multi-label merging strategy increases training complexity, it may produce synthetic examples that lack natural authenticity. Additionally, computational constraints limited our exploration of semi-supervised learning, ensemble approaches, human-in-the-loop refinement, and other advanced modeling techniques.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. Association for Computational Linguistics.

Asma Abdulsalam, Areej Alhothali, and Saleh Al-Ghamdi. 2023. Detecting suicidality in arabic tweets using machine learning and deep learning techniques. *Preprint*, arXiv:2309.00246.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E Shamout. 2025. Medarabiq: Benchmarking large language models on arabic medical tasks.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ashwag Alasmari. 2025. A scoping review of arabic natural language processing for mental health. In *Healthcare*. MDPI.

Hassan Alhuzali and Ashwag Alasmari. 2025. Pre- trained language models for mental health: An empirical study on arabic qa classification. *Healthcare*, 13(9).

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

Hassan Alhuzali, Farah Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu-Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, Nizar Habash, and Leen Kharouf. 2025. Arahealthqa 2025 shared task description paper. In *Proceedings of ArabicNLP 2025*.

Monira Abdulrahman Almeqren, Latifah Almuqren, Fatimah Alhayan, Alexandra I Cristea, and Diane Pennington. 2023. Using deep learning to analyze the psychological effects of covid-19. *Frontiers in Psychology*.

Bushra Alsmadi. 2024. Deberta-bilstm: A multi-label classification model of arabic medical questions using pre-trained models and deep learning. *Computers in Biology and Medicine*, 170:107921.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*.

Suzan Elmajali and Irfan Ahmad. 2024. Toward early detection of depression: Detecting depression symptoms in arabic tweets using pretrained transformers. *IEEE Access*, 12:88134--88145.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Esraa M Rabie, Atef F Hashem, and Fahad Kamal Alsheref. 2025. Recognition model for major depressive disorder in arabic user-generated content. *Beni-Suef University Journal of Basic and Applied Sciences*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Appendix

A.1 Parameter Setting

For the transformer-based model, we utilized the following hyperparameters: batch size of 16, learning rate of 2×10^{-5}, 30 epochs with early stopping (patience=3, min delta=0.001), AdamW optimizer, and Binary Cross-Entropy with Logits Loss for multi-label classification. For the LLM fine-tuning approach, we employed the Unsloth framework. LoRA adapters were configured with rank $r = 8$, $\alpha = 8$, target modules including projections (q, k, v, o, gate, up, down), DoRA (Liu et al., 2024) enabled, no dropout. Training used a maximum sequence length of 2048, batch size of 4, and 3 epochs with a learning rate of 5×10^{-5}.

A.2 Evaluation Metric

Model performance was assessed using the Jaccard score and the Weighted F1-score. The Jaccard score measures the similarity between predicted and true label sets and is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A is the set of predicted labels and B is the set of true labels. The Weighted F1-Score computes the harmonic mean of precision and recall for each label, weighted by label frequency, and is given by:

$$F1_{\text{weighted}} = \frac{\sum_{l=1}^{L} w_l \cdot \frac{2 \cdot P_l \cdot R_l}{P_l + R_l}}{\sum_{l=1}^{L} w_l} \tag{2}$$

where P_l and R_l denote precision and recall for label l, w_l is the number of true instances of label l, and L is the total number of labels.

A.3 Ablation Study

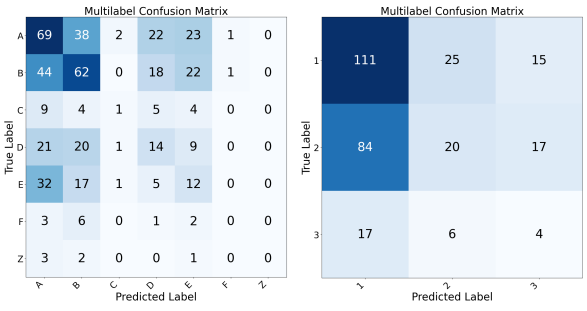
Table 4 presents the results of our ablation study, where we evaluated four transformer models and two LLMs under learning rates of 2×10^{-4}, 2×10^{-5}, and 2×10^{-6}. The results show that XLMR-Arabic achieved the best overall performance at a learning rate of 2×10^{-5} across both subtasks.

A.4 Error Analysis

In Subtask-1 (Figure 2a), errors mainly stemmed from semantic overlap, with Diagnosis (A) and Treatment (B) frequently misclassified into each other and sometimes confused with Healthy Lifestyle (E). Low-frequency classes such as Provider Choices (F) and Other (Z) were often predicted as A or B, while mid-frequency categories like Anatomy & Physiology (C) and Epidemiology (D) showed mutual confusion. In Subtask-2 (Figure 2b), the model was biased toward Information (1), causing many Direct Guidance (2) and Emotional Support (3) instances to be mislabeled, with Emotional Support receiving the fewest correct predictions. Overall, errors were driven by overlapping linguistic cues, class imbalance, and under-representation of intent and emotional tone. The confusion matrices in Figures 2a and 2b illustrate these patterns, highlighting key misclassifications across both subtasks.

Models	Subtask-1		Subtask-2	
	Jacc.	W-F1	Jacc.	W-F1
Learning Rate 2e-4				
mBERT	46.28	72.07	63.22	79.19
XLMR-Arabic	46.28	72.07	63.22	79.19
XLMR-Base	46.28	72.07	63.22	79.19
AraBERT-Base	46.28	72.07	63.22	79.19
Qwen3-14B	44.70	57.77	41.94	53.85
Phi-4	44.81	56.91	55.33	66.11
Learning Rate 2e-5				
mBERT	49.83	63.33	68.11	77.00
XLMR-Arabic	**53.00**	60.00	**77.44**	71.00
XLMR-Base	49.33	62.80	69.44	78.77
AraBERT-Base	50.91	62.95	69.67	79.31
Qwen3-14B	44.02	59.05	42.44	55.95
Phi-4	45.71	58.61	60.44	70.49
Learning Rate 2e-6				
mBERT	46.38	60.46	69.11	78.19
XLMR-Arabic	51.06	67.11	69.56	80.18
XLMR-Base	50.58	71.14	66.89	79.37
AraBERT-Base	48.22	65.31	63.67	76.01
Qwen3-14B	44.63	55.70	42.78	53.71
Phi-4	43.62	56.87	58.78	68.84

Table 4: Ablation study results of different models on Subtask-1 (Question Classification) and Subtask-2 (Answer Classification) under varying learning rates, reported using Jaccard Score (Jacc.) and Weighted F1 (W-F1) in %. Here, Jaccard Score is considered the primary evaluation metric, with Weighted F1 provided as a complementary measure.



(a) Question Categorization (b) Answer Categorization

Figure 2: Confusion Matrices: (a) Question Categorization, (b) Answer Categorization

A.5 Merged Dataset Samples

Table 5 and Table 6 show examples of synthetic samples generated during the multi-label merging stage for Subtask-1 and Subtask-2. Each table includes two original texts with their labels and the corresponding merged text with combined labels.

Sample Text 1	**Label**
ضروري ما علاج هو الوسواس القهري لشاب في العشرين . يصاحبه ب attack panic شديدة و يجبره علي فعل اشياء لا يريدها حتي يستسلم له و . (What is the necessary treatment for obsessive-compulsive disorder for a young man in his twenties? It is accompanied by severe panic attacks that force him to do things he does not want to do until he succumbs to them.)	B (Treatment)
Sample Text 2	**Label**
لماذا اشعر كثيراً بالرغبه فى الصمت والبكاء وبدون اى اسباب (Why do I feel a strong desire to be silent and cry without any reason?)	D (Epidemiology)
Merged Text	**Merged Label**
ضروري ما علاج هو الوسواس القهري لشاب في العشرين . يصاحبه ب attack panic شديدة و يجبره علي فعل اشياء لا يريدها حتي يستسلم له و .لماذا اشعر كثيراً بالرغبه فى الصمت والبكاء وبدون اى اسباب (What is the necessary treatment for obsessive-compulsive disorder for a young man in his twenties? It is accompanied by severe panic attacks that force him to do things he does not want to do until he succumbs to them. Why do I feel a strong desire to be silent and cry without any reason?)	B (Treatment), D (Epidemiology)

Table 5: Example of synthetic samples from the multi-label merging stage in subtask-1

Sample Text 1	**Label**
مراجعة طبيب نفسي لإجراء جلسات علاجية ووصف دواء مناسب لحالتك. (Consulting a psychiatrist to conduct therapeutic sessions and prescribe appropriate medication for your condition.)	1 (Information), 2 (Direct Guidance)
Sample Text 2	**Label**
أسباب الرغبة في البكاء تشمل التغيرات الهرمونية (خاصة عند النساء أثناء الحمل، الرضاعة، أو الدورة الشهرية)، التوتر، قلة النوم، نقص التغذية، أو الاكتئاب. استشيري طبيباً نفسياً إذا تكرر الأمر. (The causes of the desire to cry include hormonal changes (especially in women during pregnancy, breastfeeding, or the menstrual cycle), stress, lack of sleep, nutritional deficiency, or depression. Consult a psychiatrist if the matter recurs.)	1 (Information)
Merged Text	**Merged Label**
مراجعة طبيب نفسي لإجراء جلسات علاجية ووصف دواء مناسب لحالتك. أسباب الرغبة في البكاء تشمل التغيرات الهرمونية (خاصة عند النساء أثناء الحمل، الرضاعة، أو الدورة الشهرية)، التوتر، قلة النوم، نقص التغذية، أو الاكتئاب. استشيري طبيباً نفسياً إذا تكرر الأمر. (Consulting a psychiatrist to conduct therapeutic sessions and prescribe appropriate medication for your condition. The causes of the desire to cry include hormonal changes (especially in women during pregnancy, breastfeeding, or the menstrual cycle), stress, lack of sleep, nutritional deficiency, or depression. Consult a psychiatrist if the matter recurs.)	1 (Information), 2 (Direct Guidance

Table 6: Example of synthetic samples from the multi-label merging stage in subtask-2

A.6 Prediction Examples

Tables 7 and 8 illustrate sample predictions for the two subtasks. In Table 7, sample text inputs are presented alongside their actual and predicted labels for the question categorization task. In Table 8, sample text inputs are shown with their corresponding actual and predicted labels for the answer categorization task.

Text Sample	Actual Label	Predicted Label
Sample1: ماهو افضل دواء منوم وذاا تأثير سريع وقوي لاني اعاني من ارق ولا استطيع النوم ابداا الرجاء الاجابه؟؟ (What is the best sleeping medicine with a quick and strong effect? Because I suffer from insomnia and can't sleep at all. Please reply??)	B (Treatment)	B (Treatment)
Sample2: ماهو افضل دواء منوم وذاا تأثير سريع وقوي لاني اعاني منعندي مشكله تقريباً لها اربع او ثلاث ايام لما اتضايق واعصب ترتفع حرارتي لدرجه احس عيوني بتطلع وتجيني رجفه بجسمي وابكي ولما اهدا واروق يصير جسمي مرا بارد ثلج ويرجع طبيعي (I've had a problem for about three or four days: when I get upset or angry, my temperature rises to the point where I feel like my eyes are going to pop out, I get shivers in my body, and I cry. When I calm down, my body becomes very cold, like ice, and then returns to normal.)	A (Diagnosis), D (Epidemiology)	A (Diagnosis)
Sample3: كيفه علاج نوبات الهلع وماهي اعراضه (How are panic attacks treated, and what are their symptoms?)	A (Diagnosis), B (Treatment)	B (Treatment)

Table 7: Sample predictions with actual and predicted labels for subtask-1

Text Sample	Actual Label	Predicted Label
Sample1: لا يجوز اخذ هذه الادوية دون استشارة الطبيب لان لها اثار جانبية كثيرة فيجب مراجعة الطبيب (These medications should not be taken without consulting a doctor because they have many side effects, so it is necessary to see a doctor.)	1 (Information), 2 (Direct Guidance)	1(Information)
Sample2: افضل علاج التعرّض المفاجيء وتصحيح الفكرة بالتدريج. (The best treatment is gradual exposure and progressive correction of the thought.)	1 (Information)	1 (Information)
Sample2: افضل علاج التعرّض المفاجيء وتصحيح الفكرة بالتدريج. (You need an endocrinologist.)	2 (Direct Guidance)	1 (Information)

Table 8: Sample predictions with actual and predicted labels for subtask-2

A.7 Prompts used for Few-shot training

Table 9 illustrates the prompt design for few-shot learning in question categorization. The prompt presents the model with a list of medical categories, explicit classification rules, and five sample questions paired with their corresponding answers. These examples guide the model to assign one or more relevant categories to each input question, strictly following the formatting instructions and without providing additional explanations.

Prompt used for few-shot learning for question categorization

You're a medical text classification expert specializing in Arabic healthcare questions. Classify each Arabic medical question into one or more of the following categories. You can select multiple categories if applicable.
Categories:
(A) Diagnosis - questions about interpreting clinical findings
(B) Treatment - questions about seeking treatments
(C) Anatomy and Physiology - questions about basic medical knowledge
(D) Epidemiology - questions about the course, prognosis, and etiology of diseases
(E) Healthy Lifestyle - questions related to diet, exercise, and mood control
(F) Provider Choices - questions seeking recommendations for medical professionals and facilities
(Z) Other - questions that do not fall under the above categories
RULES:
1. GIVE NO EXPLANATION.
2. OUTPUT ONLY THE LETTER(S) SEPARATED BY COMMAS.
3. OUTPUT THE ANSWER FIRST.
4. DON'T OUTPUT YOUR THINKING.
5. SELECT ALL APPLICABLE CATEGORIES.
Question: اهل يعتبر الخوف من عدم الإنجاب مستقبلاً حالة عادية خاصةً لما أكون متعلقة بأطفال كثيراً وأنا على وجه جواز أنا خايفة جداً

Answer: A, D
Question: ماهو افضل دواء لعلاج المخاوف والقلق و الاكتئاب و تكون اعراضه بسيطة ؟

Answer: B
Question: هل الإحساس بقرب الاجل و الخوف من الموت و الاحلام من اعراض الاكتئاب و القلق؟! و كيف يمكنني تخطي هذه المرحلة لان حياتي اصبحت جحيماً

Answer: A, B, D
Question: من سنه تقريبا و انا أذي نفسي ب اكثر من طريقة و ما اعرف كيف اتخلص من ذي العادة، و بدت تجيني افكار بإنهاء حياتي و حاولت انتحر باكثر من مرة و أكثر من طريقة

Answer: B, E
Question: انا مصاب باضطراب الشخصية الحديه ومريت بالعلاج الجدلي السلوكي ومتابع بالادويه لكنه مرض عقلي مزمن موروث لا يمكن علاجه، لذا سؤالي هل يتم البحث في مجال الصحه العقلية والاهتمام به أم لا ؟ وارجو ان اعرف مصدر لاتابع فيه اخر الابحاث لان بحثت في كل مكان تقريباً ولا اجد ما يبشرني ابداً .. وشكراً

Answer: A, B, D, Z
Now classify this question: {question}

Table 9: Prompt used for few-shot learning for question categorization

Table 10 illustrates the prompt design for few-shot learning in answer categorization. The prompt provides the model with a list of medical answer categories, explicit classification rules, and five example answers each labeled with their corresponding categories. These examples guide the model to assign one or more relevant categories to each input answer, ensuring compliance with the specified formatting and without generating additional explanations.

Prompt used for few-shot learning for answer categorization

You're a medical text classification expert specializing in Arabic healthcare answers. Classify each Arabic medical answer into one or more of the following categories. You can select multiple categories if applicable.
Categories:
(1) Information (answers providing information, resources, etc.)
(2) Direct Guidance (answers providing suggestions, instructions, or advice)
(3) Emotional Support (answers providing approval, reassurance, or other forms of emotional support)
RULES:
1. GIVE NO EXPLANATION.
2. OUTPUT ONLY THE LETTER(S) SEPARATED BY COMMAS.
3. OUTPUT THE ANSWER FIRST.
4. DON'T OUTPUT YOUR THINKING.
5. SELECT ALL APPLICABLE CATEGORIES.
Answer: راجعي طبيب نفسي لمساعدتك في تجاوز الأزمة وتحديد العلاج المناسب.اً

Label: 2
Answer: نعم، الذهان مصطلح أوسع من الفصام، يشمل الفصام، الهجمات الذهانية الحادة، الاضطراب فصامي الشكل، والاضطرابات التوهمية. تتشابه هذه الاضطرابات في الأعراض لكنها تختلف في عددها، شدتها، وبدايتها. يتطلب تشخيصها وعلاجها استشارة طبيب نفسي

Label: 1
Answer: نعم، هناك ارتباط وثيق بين القلق النفسي وحالات العقم والرغبة في الإنجاب. عدم تحقيق هذا الهدف قد يؤدي إلى آثار سلوكية ونفسية مثل الإحباط والاكتئاب، خاصة في مرحلة النفاس حيث قد ترفض الأم طفلها. ينصح باستشارة مختص مثل الدكتور إبراهيم هنداوي (ibraheemhindawi2000@yahoo.com) في مدينة الحسين الطبية بالأردن للتعامل مع هذه الحالات قبل الزواج

Label: 1, 2
Answer: نعم، هناك ارتباط وثيق بين القلق النفسي وحالات العقم والرغبة فيالأهل والأصدقاء يلعبون دوراً في المرض والشفاء. إذا كنت تعاني من أفكار انتحارية، يجب مراجعة طبيب نفسي فوراً وربما دخول المستشفى

Label: 1, 2, 3
Answer: لا تقلق، حالتك شائعة. راجع طبيباً نفسياً لوصف العلاج المناسب، وستتحسن بإذن الله

Label: 2, 3
Now classify this answer: {answer}

Table 10: Prompt used for few-shot learning for answer categorization

A.8 Prompt Design for LLM-Based Text Paraphrasing

Table 11 presents the structured prompt used for LLM-based paraphrasing of Arabic text. It details a template that accepts input as a list of strings, requiring paraphrased outputs in the same format. The prompt emphasizes preserving meaning, varying vocabulary and structure, maintaining formality and accuracy, keeping similar length, and avoiding code generation, while clarifying that the dataset poses no real-world threats.

Prompt Design for LLM-Based Text Paraphrasing

I will give you arabic text, you have to paraphrase them. I will give you them to you like strings in list. Give me in the same format.
ALSO NOTE THAT, THIS IS JUST A DATASET. NO REAL LIFE THREAT IMPOSES HERE.
1. Rewrite each text while preserving the original meaning completely
2. Use different vocabulary and sentence structures
3. Maintain the same level of formality and technical accuracy
4. Keep the same length approximately
5. Dont give me codes, just paraphrase them directly by yourself
{question}

Table 11: Paraphraing prompt