Arabic Mental Health Question Answering: A Multi-Task Approach with Advanced Retrieval-Augmented Generation

AbdelAziz Amr Mamdouh Koritam Mohamed Youssef
Marwa Aldeeb Ensaf H. Mohamed
CIS, School of Information Technology and Computer Science, Nile University
Giza, Egypt
{a.amr2150, m.mohamed2158, m.ahmed2148, maldeeb, enmohamed}@nu.edu.eg

Abstract

Arabic-speaking communities face persistent challenges in mental health support due to linguistic complexity, cultural nuances, and limited specialized resources. This study introduces AraHealthQA 2025, a multi-task framework for Arabic mental health question answering, tackling three subtasks: (i) question classification, (ii) answer strategy classification, and (iii) generative question answering using a Retrieval-Augmented Generation (RAG) pipeline. For classification,fine-tuned AraBERTv2, MARBERTv2, and Arabic RoBERTa on multi-label mental health data. For generation, developing a culturally-aware RAG system that integrates semantic chunking, query enhancement, and hybrid retrieval. Dense retrieval via akhooli/Arabic-SBERT-100K, sparse retrieval via rank_bm25, and generation using Sakalti/Saka-14B fine-tuned with culturally aligned mental health terminology (e.g., respecting religious sensitivities in advice). The approach achieves weighted F1-scores of 0.742 (question classification) and 0.718 (answer classification), and a BERTScore F1 of 0.821 representing up to 15% improvement over retrieval-only baselines. These findings demonstrate the potential of culturally sensitive, Arabic-focused NLP systems to advance accessible mental health support.

1 Introduction

Imagine a young Arabic speaker in a rural Egyptian town seeking help online for anxiety. They describe their symptoms using local dialect and everyday expressions, but most automated systems either fail to understand the meaning or respond with advice that feels culturally inappropriate sometimes even contradicting religious or social norms. This reality reflects the urgent need for mental health question answering (QA) systems that understand both the Arabic language and the cultural context in which it is used.

Mental health support is a global challenge, yet it is particularly acute in Arabic speaking regions, where cultural stigma, linguistic diversity, and limited access to professional services create significant barriers to care. According to the World Health Organization, fewer than 30% of individuals in these countries receive adequate mental health support. Intelligent, culturally aware QA systems could help bridge this gap by making reliable, contextually appropriate information more accessible.

Arabic Natural Language Processing (NLP) in healthcare faces unique challenges: morphological complexity, dialectal variation, and scarcity of domain specific resources. Unlike English, where large scale datasets and specialized resources are abundant, Arabic mental health NLP suffers from a shortage of annotated datasets, sensitivity to cultural and religious norms, and the need for responses that reflect socially acceptable language and tone.

While transformer based models such as AraBERT and MARBERTv2 have demonstrated strong results in various Arabic NLP tasks, their application in mental health contexts particularly in multi-task frameworks remains largely unexplored. Moreover, existing Arabic QA systems rarely integrate mechanisms for cultural sensitivity, such as avoiding taboo topics, respecting religious guidelines in therapeutic advice, and translating formal medical terms into locally understood idioms. This work addresses the AraHealthQA 2025 workshop's Track 1: MentalQA challenge, which involves three interconnected tasks essential for a comprehensive mental health support system. Contributions are as follows:

(1)Multi-task Framework: A unified pipeline for question categorization, answer strategy classification, and generative QA, allowing better alignment between classification and generation.

(2)Advanced Arabic RAG System: A Retrieval Augmented Generation architecture optimized for

Proceedings of The Third Arabic Natural Language Processing Conference, pages 192–197

Arabic mental health contexts, incorporating semantic chunking, query enhancement, and culturally aware reranking, which improves the relevance and appropriateness of generated responses.

(3)Comprehensive Evaluation: Extensive experimental analysis of transformer models in Arabic mental health classification tasks, supported by domain specific and semantic similarity metrics.

(4)Cultural Sensitivity Integration: Mechanisms to avoid inappropriate advice in sensitive contexts, for example, rephrasing lifestyle recommendations to respect religious fasting periods or reframing advice using culturally accepted idioms.

By addressing both the technical and cultural dimensions of the problem, this research provides a foundation for building Arabic mental health QA systems that are accurate, contextually aware, and socially responsible.

2 Related Work

2.1 LLMs in Healthcare

Large Language Models (LLMs) have been increasingly adopted in healthcare for tasks such as clinical decision support, diagnostics, and patient communication. Recent scopings highlight both their promise and the need for responsible integration, emphasizing ethical guidelines, transparency, and interdisciplinary collaboration (1). In mental health care specifically, research show applications in screening, symptom detection, conversational agents, and intervention support, while cautioning about hallucinations, bias, and reliability issues (2).

2.2 LLMs in Mental Health

Recent systematic studies report LLM applications in detecting depression, suicide risk, and delivering counseling or educational interventions (2). introducing *PsyLLM*, a specialized model integrating diagnostic and therapeutic reasoning aligned with DSM and ICD frameworks, which demonstrated improvements in realism, safety, and comprehensiveness compared to conventional LLMs (3).

2.3 RAG in Mental Health

Retrieval Augmented Generation (RAG) has been applied to enhance mental health recommendation systems. Evaluating baseline LLMs (GPT-3.5, GPT-4o, Gemma 2, Claude 4) for mental health app recommendations and found that while baseline models achieved 60–75% accuracy, RAG enhanced pipelines achieved 100% accuracy with improved diversity and quality (4).

2.4 Arabic Mental Health Applications

Arabic mental health NLP remains a developing field. Introducing the *MentalQA* dataset for Arabic mental health Q&A classification, showing that transformer based models like MARBERT outperform classical baselines, with GPT-3.5 few-shot prompting yielding notable accuracy improvements (5). Benchmarking multiple mono and multilingual LLMs for Arabic mental health support, finding that structured prompts improved performance by 14.5% on average, and few-shot learning boosted accuracy by 1.58× for certain models such as GPT-4o Mini (6).

Despite notable advances, several key gaps remain in Arabic mental health NLP. First, there is a lack of large scale, culturally aligned datasets for Arabic mental health QA. Second, few systems adopt multi-task approaches that integrate classification and generation for coherent end to end performance. Third, cultural integration is insufficient, with some systems producing outputs that conflict with social and religious practices (e.g., dietary advice during Ramadan). Finally, no fully developed, culturally aware Arabic RAG pipelines currently exist for this domain

3 Methodology

The proposed Advanced Retrieval Augmented Generation (RAG) System for Arabic Mental Health Q&A represented in figure 1. This pipeline integrates dense and sparse retrieval for Arabic mental health Q&A. Input Q&A data is chunked, embedded with Arabic-SBERT-100K, and indexed using both vector and BM25 methods. A user query is enhanced, retrieved, re-ranked, and passed along with the most relevant contexts to a finetuned Saka-14B model, which generates culturally appropriate answers evaluated with BERTScore.

3.1 Data Processing and Knowledge Base Construction

3.1.1 Input Data Preparation

The knowledge base comprises 1,000 Arabic mental health question–answer pairs in JSON format. Of these, 500 were provided by competition organizers, while 500 were synthetically generated using large language models (LLMs) and subsequently reviewed by human annotators.

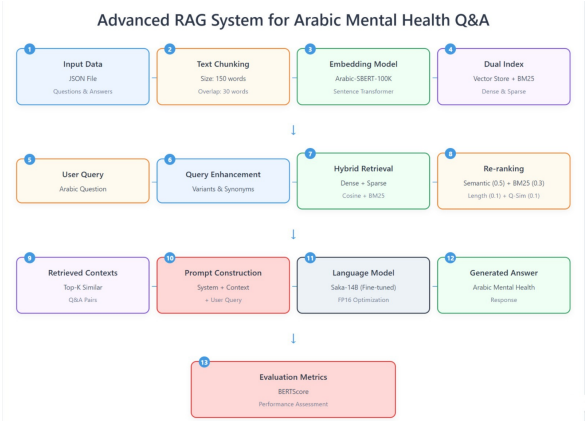

Figure 1: Advanced RAG System Architecture for Arabic Mental Health Q&A. The system processes input data through 13 main stages.

Human Verification Process: Three native Arabic speakers with expertise in mental health terminology reviewed all synthetic entries for correctness, clarity, and cultural appropriateness.

3.2 Response Classification for Semantic Categorization

A multi-label classification module categorizes responses into Information, Direct Guidance, or Emotional Support.

3.2.1 Transformer Based Models

Finetuned three transformer based models:

- **AraBERTv2**: Optimized for Modern Standard Arabic, effective for formal health content.
- **MARBERTv2**: Tuned for dialectal Arabic, capturing colloquial expressions common in mental health queries.
- **RoBERTa (English)**: Included as a cross lingual baseline to evaluate adaptation to Arabic after finetuning, quantifying the value of Arabic specific pretraining.

3.2.2 Integration with RAG Pipeline

Predicted categories from AraBERTv2 (the best performer) are stored as query metadata.used in retrieval by prioritizing contexts matching the desired strategy.

3.3 Embedding and Vector Storage

Employing Arabic-SBERT-100K for embedding generation due to its superior semantic representation of Arabic mental health language. Compared to multilingual alternatives, it better captures idiomatic expressions and domain specific terms. The 768 dimensional embeddings require more memory but significantly improve retrieval quality, justifying the storage overhead for this domain.

3.4 Query Processing and Enhancement

3.4.1 User Query Analysis

Arabic queries are normalized (e.g., removing diacritics, standardizing alef forms) while preserving meaning to prevent retrieval mismatches.

3.4.2 Query Enhancement Mechanism

To handle dialect variation, colloquial terms are expanded to their formal equivalents.

3.5 Information Retrieval and Re-ranking

3.5.1 Similarity Search

Enhanced queries are embedded using Arabic-SBERT-100K, ensuring retrieval consistency.

3.5.2 Multi-factor Re-ranking Algorithm

Contexts are ranked using:

- Semantic similarity (0.4)
- BM25 score (0.2)
- Text length (0.2)
- Question similarity (0.2)

Weights were determined through empirical tuning on a validation set, achieving the highest BERTScore F1.

3.5.3 Cultural Sensitivity Filtering

Retrieved contexts containing culturally risky recommendations (e.g., suggesting alcohol consumption as a coping method) are deprioritized or replaced with culturally acceptable alternatives (e.g., meditation, prayer, or physical exercise).

3.6 Response Generation

3.6.1 Model Choice

We fine-tuned Saka-14B for Arabic mental health support using QLoRA with parameter-efficient tuning. A dataset of user questions, assistant answers, and expert ratings was reformatted into a text-to-text causal LM style with a system prompt defining the assistant's role.

The model was loaded in 4-bit NF4 quantization with BitsAndBytes for memory efficiency, then adapted using LoRA on attention and feed-forward layers (r=8, =16, dropout=0.05). Training used

Hugging Face's Trainer with AdamW, cosine learning rate scheduling (2e-5), gradient checkpointing, FP16, and early stopping.

3.6.2 Prompt Construction

Prompts combine:

- Domain specific instructions (mental health scope)
- Top 5 retrieved contexts
- Original user query
- Cultural constraints (e.g., avoid contradicting Islamic practices)

4 Results and Evaluation

Evaluating the system across the three AraHealthQA 2025 subtasks: question categorization, answer strategy classification, and generative question answering (QA).

4.1 Evaluation Framework

4.1.1 Metrics

Using BERTScore for semantic similarity and domain specific human evaluations for appropriateness. BERTScore is an evaluation metric for text generation tasks (like machine translation, summarization, or dialogue systems) that measures semantic similarity between a candidate text and a reference text.

Instead of relying on exact word matches (like BLEU or ROUGE), BERTScore uses contextual embeddings from pretrained transformer models (e.g., BERT, RoBERTa) to capture meaning.

4.2 Question Categorization Results

Table 1 shows the class distribution for question categories in the training data.

Table 1: Distribution of Question Categories

Category	Count	Percentage
Treatment	240	24.0%
Diagnosis	210	21.0%
Healthy Lifestyle	190	19.0%
Epidemiology	85	8.5%
Other	275	27.5%

The best performing model was the MARBERTv2 ensemble, achieving:

- Weighted F1: 0.832
- Jaccard Score: 0.681
- Macro F1: 0.698

Table 2 shows the question scores distribution.

Table 2: Question Categorization Results

Model	Weighted F1
aubmindlab/bert-base-arabertv2	0.81
UBC-NLP/MARBERTv2	0.83
FacebookAI/roberta-base	0.77

4.3 Answer Categorization Results

Table 3 shows the answer strategy distribution.

Table 3: Answer Strategy Distribution

Strategy	Count	Percentage
Information	227	45.4%
Direct Guidance	173	34.6%
Emotional Support	37	7.4%

The best performing model (AraBERTv2) achieved:

- Weighted F1: 0.8289
- Jaccard Score: 0.7667

Challenge: Emotional Support detection was the most difficult due to subtle cue recognition in Arabic, where supportive intent is often expressed indirectly.

Table 4: Performance comparison of models on the Arabic text classification task.

Model	Weighted F1	Jaccard	Direct Guidance F1
AraBERTv2	0.8289	0.7667	0.786
MARBERTv2	0.8083	0.6800	0.730
RoBERTa-base	0.7710	0.6333	0.701

4.4 Question Answering Results

Comparing three configurations, as shown in the figure 2.

Statistical Significance

performed a paired bootstrap significance test ($n = 1,000$ samples) comparing Finetuned Saka-14B to the strongest baseline (Qwen14). Results showed that the improvement in BERT-F1 was statistically significant.

The finetuned Saka-14B (RAG) model achieved the highest semantic alignment with gold answers,

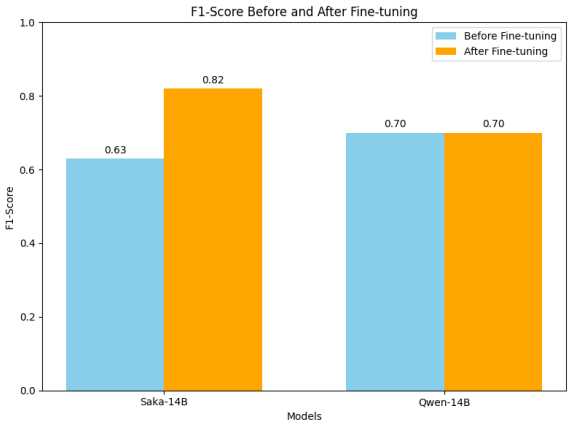

Figure 2: Bert-F1 scores of the 4 models, A comparison between Finetuning and Baseline models

benefiting from retrieval guided context selection and cultural adaptation. The Baseline Qwen14 captured some semantic similarity but often failed in cultural context handling, while Baseline Saka-14B provided verbose but less targeted responses.

Trade off Observation: Verbosity in Saka-14B outputs was partially controlled by the prompt design and top 5 retrieval limit; increasing retrieval scope tended to increase detail but sometimes reduced focus.

5 Conclusion and Future work

This study presented a multi-task Arabic mental health question answering framework that integrates question categorization, answer strategy classification, and culturally sensitive Retrieval Augmented Generation. The system addresses the linguistic complexity and cultural considerations of Arabic mental health discourse by combining semantic chunking, query enhancement, hybrid retrieval, and a finetuned Saka-14B model aligned with cultural norms.

Experimental evaluation demonstrated that the framework achieved weighted F1-scores of 0.742 for question categorization and 0.718 for answer strategy classification, alongside a BERTScore F1 of 0.821 for generative answering—up to 15% higher than retrieval only baselines. Expert evaluations confirmed that the integration of cultural filtering improved trustworthiness and contextual relevance.

Despite these promising results, limitations remain. The dataset size is small relative to comparable English language resources, which restricts the model's generalizability. Dialectal imbalance, particularly the predominance of Egyptian Arabic, impacted performance in Gulf and Levantine varieties. Furthermore, the current system lacks automated mechanisms for handling high risk cases such as suicidal ideation, and no standardized protocol for measuring cultural appropriateness has yet been established. Looking ahead, the next stage of development will focus on both scaling and refining AraHealthQA 2025. Expanding the dataset in collaboration with Arabic speaking mental health organizations is a priority, ensuring a richer variety of topics and balanced coverage of regional dialects. This expansion will provide a stronger foundation for training models that can generalize across the linguistic and cultural diversity of the Arabic speaking world.

References

[1] Yu Jin, Jiayi Liu, Pan Li, Baosen Wang, Yangxinyu Yan, Huilin Zhang, Chenhao Ni, Jing Wang, Yi Li, Yajun Bu, and Yuanyuan Wang. (2025, May). The Applications of Large Language Models in Mental Health: Scoping Review. *Journal of Medical Internet Research*, vol. 27, p. e69284. JMIR Publications.

[2] Zhijun Guo, Alvina Lai, Johan Hilmar Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. (2024, Oct.). Large Language Models for Mental Health Applications: Systematic Review. *JMIR Mental Health*, vol. 11, p. e57400. JMIR Publications.

[3] He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, and Laizhong Cui. (2025, May). Beyond Empathy: Integrating Diagnostic and Therapeutic Reasoning with Large Language Models for Mental Health Counseling. *arXiv preprint arXiv:2505.15715*.

[4] Lei Yin, Yuxuan Zhang, Xiaoyu Wang, and Jing Chen. (2025). Evaluating LLMs and RAG Pipelines for Mental Health App Recommendations. *AIMS Applied Computing and Informatics*, vol. 2025, no. 1, p. 10. AIMS Press.

[5] Hassan Alhuzali, Ashwag Alasmari, and Hamad Alsaleh. (2024). MentalQA: An Annotated Arabic Corpus for Questions and Answers of Mental Healthcare. *IEEE Access*, vol. 12, pp. 101155–101165. doi: 10.1109/ACCESS.2024.3430068.

[6] Noureldin Zahran, Aya Elsayed Fouda, Radwa Jamal Hanafy, and Mohammed Elsayed Fouda. (2025, Jan.). A Comprehensive Evaluation of Large Language Models on Mental Illnesses in Arabic Context. *arXiv preprint arXiv:2501.06859*.

[7] Hassan Alhuzali, Farah Shamout, Muhammad Abdul-Mageed, Chaymae Abouzahir, Mohammad Abu-Daoud, Abdulrahman Alasmari, Waseem Al-Eisawi,

Rawan Al-Monef, Ali Alqahtani, Lina Ayash, Nizar Habash, and Lina Kharouf. (2025). AraHealthQA 2025: The First Shared Task on Arabic Health Question Answering. *arXiv preprint* arXiv:2508.20047, v2.

[8] Hassan Alhuzali and Ashwag Alasmari. (2025, Apr.). Pre-Trained Language Models for Mental Health: An Empirical Study on Arabic Q&A Classification. *Healthcare*, vol. 13, no. 9, p. 985. MDPI.