AraMinds at AraHealthQA 2025: A Retrieval-Augmented Generation System for Fine-Grained Classification and Answer Generation of Arabic Mental Health Q&A

Mohamed Zaytoon, Ahmed Salem, Ahmed Sakr, Hossam Elkordi
Department of Computer and Systems Engineering
Alexandria University, Egypt
{mohamed.zaytoon24,es-AhmedMahmod2022,
es-ahmedsakr20,es-hossam.elkordi2018}@alexu.edu.eg

Abstract

We present a mental health support system for Arabic that can classify both patient questions and doctor answers, and generate answers for new questions. The classification model organizes the input text to understand better the intent of the user and the response style, while the generation model produces accurate and empathetic responses. In evaluations, our system ranked 3rd in answer classification and 4th in answer generation, with only a small margin from the top-ranked systems. These results highlight the effectiveness of multi-label classification and RAG for improving access to mental health information and support in Arabic.

1 Introduction

Mental health and human psychology have been studied and practiced as separate fields of medicine for centuries (Grob, 1998), yet, studies show that the general population in the MENA region still refrains from seeking medical help when it comes to mental health-related problems, mainly due to social considerations (Nazmy, 2025), leading to a significant degradation in the public mental health status(Altobaishat et al., 2025).

After the rise of large language models (LLMs) and Agentic Artificial intelligence (AI) applications (Minaee et al., 2024; Plaat et al., 2025), especially for complex reasoning and questions answering (QA) tasks, many researches have explored the use of advanced language models to provide not only assistance for medical professionals (Nazi and Peng, 2024), but also as an alternative means of delivering mental health care (Guo et al., 2024), due to their ability to provide human like responses and interactions (Zaki and Hassan, 2023; Zahran et al., 2025), offering scalable, accessible, private, and stigma-free pathways for psychological support.

Besides the aforementioned cultural barriers, providing automated mental health support for Arabic speakers faces other challenges, including:

- **Higher Accuracy Standard:** the medical field -especially psychology- has a very low tolerance of error, as opposed to other AI applications, where in some cases, accuracy could be traded off for speed or power efficiency (Han et al., 2015), the cost of error in medical applications could lead to unquantifiable losses, causing -in the worst cases- human fatalities (Topol, 2019).
- **Data Scarcity:** this challenge is two-fold: 1) Arabic datasets are scarce and generally have lower quality annotations in general, 2) Datasets for mental health-related problems (Alhuzali et al., 2024) are not as abundant as other health-related datasets (**?**Alasmari, 2025).
- **Patient Confidentiality:** unlike other medical disciplines, obscuring patient identity is more challenging, as personal information, such as background and upbringing circumstances, has to be included in every case.
- **Linguistic Complexity:** Apart from data problems, Arabic is morphologically rich and includes many dialects with a high level of diversity (Habash, 2010), which leads to a wide performance gap of language models between Arabic and other languages.

Our contribution in this task could be summarized as:

- We developed a classification system to label questions and answers into multiple fine-grained categories.
- We integrate additional external knowledge from Arabic medical platforms to develop

Proceedings of The Third Arabic Natural Language Processing Conference, pages 198–203

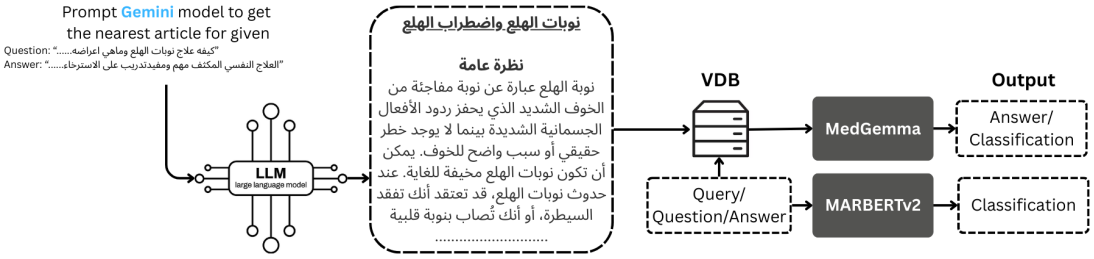

Figure 1: General overview of our system starting from the document collection phase, the RAG system for all three subtasks, and multi-label classification, and question answering.

a simple yet effective retrieval-augmented generation (RAG) system tailored to Arabic mental health Q&A, improving overall classification accuracy and reducing hallucination in answers.

- We achieved 0.56[1] f1 score on the first track, 0.76 on the second one, and 0.663 on the BERTScore (Zhang et al., 2020) metric.

2 Background

The challenge (Alhuzali et al., 2025) is divided into three main sub-challenges. The first two are **multilabel classification** for human questions and their answers, respectively, with seven different classes for questions and three for answers. The third sub-challenge is **answer generation** to provide mental health assistance to patients by generating contextually appropriate and medically accurate responses to user queries.

A lot of work has been put into data collection for mental health, from different sources such as social media platforms, for example *Reddit* (Cohan et al., 2018; Di Cara et al., 2023). This data is then manually annotated to identify different mental health-related conditions, such as depression detection (Han et al., 2022), anxiety, bipolar disorder, and suicidal intent (Ji et al., 2022).

Recent work explored the capabilities of LLMs in the mental health domain for different tasks. First, classification tasks to detect different mental health conditions (Racha et al., 2025) and their causes (Yang et al., 2023). Second, user question answering, to respond to different user queries, either informative or for advice seeking. Third, chatbots offering their users a safe space for relief and conversation (Shan et al., 2022) or chatbots that help doctors in the application and monitoring of different treatments, including cognitive behavioral therapy (CBT) (Farzan et al., 2025), and self-attachment technique (SAT) (Elahimanesh et al., 2023).

The use of external knowledge in Retrieval Augmented Generation (RAG) systems to enhance the factuality and robustness of LLMs, especially in the medical domain, has been explored, for example (Vladika and Matthes, 2024) retrieved from a large corpus that included diverse topics such as *dietary supplements, heart and lungs, reproductive health, cancer, and mental health*. This demonstrated that retrieval strategies prioritizing fewer, more recent, and highly cited sources—especially at the sentence level—significantly improved answer quality in health question answering tasks.

Recently, many researchers have taken an interest in the mental health domain in the Arabic language. starting with simpler classification tasks, such as depression detection (Maghraby and Ali, 2022; Hassib et al., 2022). Others worked on more advanced tasks such as (Zahran et al., 2025).

3 System Overview

In this section, we present our system setup in detail. First, we go through the data collection process, then we describe our RAG pipeline, used models, and embedding vectorestore. Finally, we go through how we built our multi-label classifier for the question and answer classification tracks.

3.1 Building Knowledge Base

To effectively build our knowledge base, we followed a two-step process. First, we used the Gemini API[2] to collect articles related to each question from the training set. Then, to reduce the size of the knowledge base and guarantee smoother retrieval of relevant information, we fed articles with

[1] due to the limited number of submissions, this score didn't show on the leaderboard, the metrics were computed in the post-evaluation phase.

[2] https://gemini.google.com

Task	RAG	Pretraining	F1-score
Question Classification	✓	✗	0.490
	✗	✓	**0.558***
	✗	✗	0.549*
Answer Classification	✓	✗	**0.760**
	✗	✓	0.735*
	✗	✗	0.733*

Table 1: Macro-F1 scores on the test set using the two approaches we developed, RAG and multi-label classification with and without continuous pretraining. (*) denotes experiments done during the post-evaluation phase, and were not submitted to the leaderboard due to the limited number of submissions.

high relevance to the same question to Gemini to compress them into one article. We then used the same knowledge base for both the evaluation and test phases, without any updates from the test set.

3.2 Retrieval Augmented Generation Pipeline

After the data collection phase, we embedded those related articles using `BAAI/bge-m3` (Chen et al., 2024) for its *multi-functional*, *multi-granular*, and especially *multi-lingual* capabilities. Those embeddings are then stored in a `chroma-db`[3] vectorstore. For every sample, we retrieved the top articles related to it and used them as additional context to generate answers for different classification and generation tasks using `google/medgemma-4b-it` (Sellergren et al., 2025).

Model	RAG	Validatoin	Test
Qwen2.5 7b	✗	0.608	–
Llama3.2 3b	✗	0.597	–
Phi-mini 4b	✗	0.606	–
Gemma3-4b	✗	0.619	–
MedGemma3-4b	✗	0.620	0.632
MedGemma3-4b	✓	**0.630**	**0.663**

Table 2: BertScore results on the validation and test sets for the answer generation sub-task.

3.3 Multi-label Classification for Question and Answer

For the multi-label classification tasks, we relied on `MARBERTv2` (Abdul-Mageed et al., 2021) model. To improve the model's understanding capabilities in the mental health domain, we applied masked language modeling (MLM) pretraining (Devlin et al., 2019) using different Arabic mental health books such as DSM-5 (EDITION, 1980) and articles about different psychiatric and mental health conditions from the renowned *Royal College Of Psychiatrists* [4]. The pretrained model was then fine-tuned for both the question and answer multi-label classification tasks, using the *binary cross-entropy* loss–reference the loss function–.

4 Experimental Setup

4.1 Dataset

4.1.1 Shared Task Dataset - MentalQA

The given dataset includes two different tasks: multi-label classification and answer generation. The classification part is for both questions and answers, and the generation is to reply to the given question. The QA pairs were collected from *altibbi* [5] medical platform for advisory and information, then 500 samples were manually annotated (Alhuzali et al., 2024). Questions are categorized into seven classes, while answers are classified into only three.

4.1.2 Knowledge Base Dataset

The data was collected using the Gemini API, the pre-comrpression articles were mainly from Arabic medical websites, such as `islamweb` [6], `Mayo Clinic` [7], `Mind Clinic Group` [8]. The collected articles were then curated as mentioned in section 3.1. The compressed articles were then added to a vectorstore database to facilitate retrieval

For the rag system, we use `BAAI/bge-m3` model (Chen et al., 2024), and to store the data, we used chroma-db.

4.1.3 Continue Pretraining Dataset

To better ground our base model's capacity in understanding text from the mental health domain, we utilized text scraped from multiple resources, including 32 articles from the *Royal College of Psychiatrists* and a collection of 72 books, either originally written in Arabic or manually translated into Arabic by professional human translators. The combined dataset contained approximately 4.5 million tokens.

[3] https://github.com/chroma-core/chroma

[4] https://www.rcpsych.ac.uk/mental-health/translations/arabic

[5] https://altibbi.com

[6] https://islamweb.net/ar/

[7] https://www.mayoclinic.org/ar

[8] https://mindclinicgroup.com/ar

Task	Input	Ground Truth	Prediction
Question Classification	ماهو افضل دواء منوم وذاا تأثير سريع وقوي لاني اعاني من ارق ولا استطيع النوم ابداا الرجاء الاجابه؟؟	Treatment	Diagnosis Anatomy and Physiology Epidemiology
Answer Classification	لا يجوز اخذ هذه الادوية دون استشارة الطبيب لان لها اثار جانبية كثيرة فيجب مراجعة الطبيب	Information Direct Guidance	Direct Guidance
Answer Generation	لما تصير مشكله تافهه او قويه مع اشخاص اصير افكر فيها فجأه واخلق سيناريوهات وحوار كله هواش لدرجة بالواقع علاقتي تأثرت بهالاشخاص حاولت اتجاهل الافكار بس فجأه عقلي يفكر فيها	مهم العلاج النفسي والتدريب على التركيز والاسترخاء	يبدو أنك تعاني من أفكار سلبية ومشاعر قلق بشأن علاقاتك مع الآخرين، مما يؤثر على حياتك اليومية. هذه المشاعر قد تكون ناتجة عن سيناريوهات تخيلية أو حوارات افتراضية مليئة بالجدال.**إليك بعض الاقتراحات التي قد تساعدك:*** **التحدث مع شخص تثق به:** مشاركة مشاعرك ....

Table 3: Some fauiler cases for our system on the three sub tasks.

4.2 Training Details

To continue the pretraining of the base `MARBERTV2` model, we used the AdamW optimizer, with a learning rate 2e-5, 0.01 weight decay, a cosine annealing learning rate scheduler, and a batch size of 32. We trained the model for 20 epochs. For training the Multi-Label classification models, in both sub-tasks 1,2, we used the same optimizer and learning rate, with a linear scheduler, and a batch size of 4. Both models were trained for 10 epochs. All training was done on a single NVIDIA RTX-3090 GPU.

5 Results

5.1 Classification Tasks Results

For the classification task, we tested the RAG approach to label both questions and answers separately. Then, we compared this approach with our multi-label classification. Also, we assessed the effect of the continue pretraining phase. We can see from table 1 that in both cases the pretraining phase improved the results by 0.009 and 0.002 on the question and answer classification tasks, respectively. Also, the RAG approach achieved better results in the answer classification task by 0.025, but came short by 0.068 in the question classification task.

5.2 Generation Tasks Results

To choose the best model, we run some initial tests on the answer generation task using various open-source LLMs. After picking the highest scoring model, we used it for the remaining experiments in our RAG system. The results of this experiment are shown in table 2.

5.3 Analysis

Table 3 presents failure cases from our system. In the **Question Classification**, the model misclassifies a request for a sleeping medication as a query about *Diagnosis*, *Anatomy and Physiology*, and *Epidemiology*, indicating a misunderstanding of user intent by over-focusing on the symptom keyword أرق (insomnia). For **Answer Classification**, the system correctly identifies *Direct Guidance* but misses the *Information* label, showing challenges in capturing all nuances. In **Answer Generation**, responses are verbose and generic, such as suggesting التحدث مع شخص تثق به (Talk to someone you trust), instead of specific, actionable advice, underscoring a preference for supportive text over professional recommendations.

6 Conclusion

This paper investigated our work in the first track of the AraHealthQA 2025 shared task for mental health question and answer multi-label classification and answer generation. We collected a large corpus of Arabic mental health-related books and articles and used them to continue pretraining our base encoder. Then, we fine-tuned this model on the classification tasks. We also benefited from the provided questions and collected articles from the internet using an agentic search tool. These articles are then used in a retrieval-augmented generation system for all three sub-tasks. Our system achieved 0.558 and 0.760 F1-score on question and answer multi-label classification tasks, respectively, while achieving 0.663 BertScore on the answer generation task.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Ashwag Alasmari. 2025. A scoping review of arabic natural language processing for mental health. In *Healthcare*, volume 13-9, page 963. MDPI.

Hassan Alhuzali, Ashwag Alasmari, and Hamad Alsaleh. 2024. Mentalqa: An annotated arabic corpus for questions and answers of mental healthcare. *IEEE Access*, 12:101155–101165.

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 Arabic-NLP 2025*.

Obieda Altobaishat, Mohamed Abouzid, Deemah Omari, Walid Sange, Ahmad K Al-Zoubi, Abdallah Bani-Salameh, and Yazan A Al-Ajlouni. 2025. Examining the burden of mental disorders in jordan: an ecological study over three decades. *BMC psychiatry*, 25(1):218.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Nina H Di Cara, Valerio Maggio, Oliver SP Davis, and Claire MA Haworth. 2023. Methodologies for monitoring mental health on twitter: systematic review. *Journal of Medical Internet Research*, 25:e42734.

FIFTH EDITION. 1980. Diagnostic and statistical manual of mental disorders. *American Psychiatric Association, Washington, DC*, pages 205–224.

Sina Elahimanesh, Shayan Salehi, Sara Zahedi Movahed, Lisa Alazraki, Ruoyu Hu, and Abbas Edalat. 2023. From words and exercises to wellness: Farsi chatbot for self-attachment technique. *arXiv preprint arXiv:2310.09362*.

Maryam Farzan, Hamid Ebrahimi, Maryam Pourali, and Fatemeh Sabeti. 2025. Artificial intelligence-powered cognitive behavioral therapy chatbots, a systematic review. *Iranian journal of psychiatry*, 20(1):102.

Gerald N Grob. 1998. A history of psychiatry: From the era of the asylum to the age of prozac. *Bulletin of the History of Medicine*, 72(1):153–155.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li, and 1 others. 2024. Large language models for mental health applications: systematic review. *JMIR mental health*, 11(1):e57400.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Song Han, Huizi Mao, and William J. Dally. 2015. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. *arXiv: Computer Vision and Pattern Recognition*.

Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 94–104, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mariam Hassib, Nancy Hossam, Jolie Sameh, and Marwan Torki. 2022. Aradepsu: Detecting depression and suicidal ideation in arabic tweets using transformers. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 302–311.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2022. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Comput. Appl.*, 34(13):10309–10319.

Ashwag Maghraby and Hosnia Ali. 2022. Modern standard arabic mood changing and depression dataset. *Data in Brief*, 41:107999.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

Sarah Aly Nazmy. 2025. *Cultural Factors and Mental Health Help-Seeking Behaviors Among Middle Eastern/North African Adults in the United States*. Ph.D. thesis, Alliant International University.

Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. 2025. Agentic large language models, a survey. *arXiv preprint arXiv:2503.23037*.

Suraj Racha, Prashant Joshi, Anshika Raman, Nikita Jangid, Mridul Sharma, Ganesh Ramakrishnan, and Nirmal Punjabi. 2025. Mhqa: A diverse, knowledge intensive mental health question answering challenge for language models. *arXiv preprint arXiv:2502.15418*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Yong Shan, Jinchao Zhang, Zekang Li, Yang Feng, and Jie Zhou. 2022. Mental health assessment for the chatbots. *arXiv preprint arXiv:2201.05382*.

Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56.

Juraj Vladika and Florian Matthes. 2024. Improving health question answering with reliable and time-aware evidence retrieval. *arXiv preprint arXiv:2404.08359*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*.

Mohamed Zahran and 1 others. 2025. A comprehensive evaluation of large language models on mental illnesses in arabic context. *arXiv preprint arXiv:2501.06859*.

A Zaki and R Hassan. 2023. Optimizing large language models for arabic healthcare communication: A focus on patient-centered nlp applications. *Multimodal Technologies and Interaction*, 8(11):157.

Tianyi Zhang, Kishore, Wu, Q. Weinberger, and Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.