MindLLM at AraHealthQA 2025 Track 1: Leveraging Large Language Models for Mental Health Question Answering

Nejood Abdulaziz Bin Eshaq
Department of Computer Science, King Khalid University, Abha 62521, Saudi Arabia
446818545@kku.edu.sa

Abstract

This paper presents our submission to the AraHealthQA 2025 shared task (Alhuzali et al., 2025), Sub-task 3: Arabic Mental Health Question Answering. We evaluated four large language models—GPT-4o, Gemini, Allam, and Qwen—using various prompting strategies. A simple 3-shot prompt, instructing the model to respond in Arabic, consistently outperformed zero-shot, 5-shot, and more complex methods. GPT-4o achieved the best results, with a **BERTScore F1 of 0.670** on the official hidden test set, ranking **2nd overall**. The system required no fine-tuning or external data, relying solely on prompt design and consistent evaluation.

1 Introduction

Mental health disorders, such as obsessive-compulsive disorder (OCD), depression, and suicidal ideation, affect millions worldwide, significantly impairing well-being and daily functioning (World Health Organization, 2022). Early intervention can enhance recovery, prevent severe outcomes like self-harm, and reduce the broader societal and economic burden (Patel et al., 2018). Moreover, prioritizing mental health care helps break stigma and encourages individuals to seek the support they need. The AraHealthQA 2025 shared task (Alhuzali et al., 2025) addresses the growing demand for accessible and culturally appropriate mental health resources for Arabic-speaking populations. It highlights both the social importance of providing trustworthy support and the technical challenges posed by modeling Arabic psychological discourse. The shared task comprises three subtasks; our work focuses on Subtask 3: Question Answering, which requires generating accurate, informative, and empathetic answers to mental health questions written in Arabic.

We experimented with four large language models (LLMs)—GPT-4o (OpenAI, 2024), Gemini (Team et al., 2023), Allam (Bari et al., 2024), and Qwen (Qwen Team, 2024)—and explored multiple prompting strategies, including zero-shot, few-shot, chain-of-thought (Wei et al., 2022), and self-consistency (Wang et al., 2022). Prompt selection was conducted using Meta's *LLaMA-3-70B-Instruct model (8192-token context)* as an LLM-as-a-judge, evaluated via BERTScore F1 (Zhang et al., 2019). After iterative testing, we adopted a 3-shot prompting approach and selected GPT-4o as our final submission model, based on its alignment with expert-written answers.

Our system achieved 2nd place in the official leaderboard with a BERTScore F1 of 0.670. Key challenges during development included ensuring clean and well-structured input data, enforcing consistent and controlled answer formats, and handling ambiguous or emotionally sensitive queries that require careful phrasing to avoid misinterpretation, especially in a mental health context where psychological state and cultural background may influence understanding.

The full code and evaluation scripts are available at: https://github.com/njoudae/AraHealthQA_2025_subtasck_3/tree/main.

2 Related Work

Recent years have seen significant progress in Arabic NLP for mental health, although challenges like limited data and cultural complexities still hinder its development. (Alasmari, 2025) offers a scoping review that outlines the current state of Arabic NLP in mental health, covering methods from classical machine learning models like SVM and Random Forest to more advanced transformer models such as AraBERT and MARBERT. The review notes a strong focus on detecting depression and suicidal tendencies, often leveraging social media data, and sheds light on both the strengths and drawbacks of existing techniques. While transformer models

Proceedings of The Third Arabic Natural Language Processing Conference, pages 149–154

have delivered impressive results, the study emphasizes the lack of dataset variety and the urgent need for culturally aware tools that accommodate dialectal differences and address societal stigma in Arabic-speaking regions.

Expanding on this groundwork, (Alhuzali and Alasmari, 2025) carried out a practical assessment of pre-trained language models (PLMs) for classifying Arabic mental health Q&A using the MentalQA dataset. They compared traditional machine learning techniques, Arabic PLMs like MARBERT and CAMeLBERT, and prompt-based approaches using GPT-3.5/4. Their findings revealed that PLMs significantly outperformed older feature-based models, with MARBERT delivering the best results. Interestingly, GPT-3.5 prompt-based methods excelled in few-shot learning situations, showing promise for applications in low-resource languages. However, the study also highlighted a critical limitation: the small size of the MentalQA dataset (only 500 samples), which impacts how broadly the findings can be applied.

Shifting the focus to real-world applications, (Bensalah et al., 2024) introduced MindWave, a bilingual Arabic–English mental health support app. The system uses NLP and sentiment analysis on both text and voice inputs to identify signs of burnout and depression. To tackle the shortage of Arabic sentiment datasets, the researchers built a large parallel English–Arabic medical corpus containing 945,000 sentences. They then fine-tuned machine translation models to develop classifiers tailored to Arabic. Additionally, the study compared various Arabic tokenization techniques, offering useful insights into best practices. Unlike previous efforts that focused mainly on classification or Q&A tasks, MindWave showcases how NLP tools can be seamlessly integrated into interactive support platforms and communities.

Lastly, (Zahran et al., 2025) performed a wide-ranging evaluation of large language models (LLMs) in the context of Arabic mental health. This study stands out as one of the first to deeply assess how well LLMs function in this domain. The authors pointed out both the benefits and risks of LLMs: while these models can generate meaningful and relevant responses, concerns about empathy, cultural appropriateness, and safety persist. Compared to more specialized PLMs, general-purpose LLMs showed inconsistent reliability, reinforcing the need for domain-specific adaptation and human monitoring. Collectively, these studies highlight the importance of building richer datasets, adopting multifaceted evaluation methods (beyond basic accuracy scores like BERTScore), and developing culturally sensitive NLP tools. Our research builds on these findings by focusing on prompt-based evaluation within the AraHealthQA framework, tackling both performance and ethical dimensions in this underexplored area.

3 Task and Dataset Description

The AraHealthQA 2025 shared task (Alhuzali et al., 2025) provides a benchmark dataset for evaluating Arabic mental health question answering systems. The dataset, MentalQA, was recently accepted in IEEE ACCESS and consists of 500 annotated samples of real user-submitted psychological questions and expert-written answers in Arabic (Alhuzali et al., 2024).

	question	answer	final_QT	final_AS
340	تعدحي ضروف ماكدر اروح دكتور أعصابتي ولا اخرج حالتي كدامه صرت افكر بالشغل بسبب هاي الحاله شي مخيف أكد أنواع العذاب لترجه الموت أهون الموت أهون	السلام عليكم علاج الادمان محتاج طبيب واذوية والاهم يكون عندك اراده قويه للتخلص من هذه السموم وعلى قدر عزيمتك هيكون التوفيق لنجاح العلاج ...ساعد نفسك	['B']	['1', '3']
58	شكرا الكم على هذا الموقع الرائع. كم هي المده حتي يبين فيها مفعول دواء لعلاج حالات الاكتئاب deanxit والذهان ؟	بسخرق حوالي أسبوعين لإظهار بدايات للتحسن. المراحل الأولية غير مريحة، قد يعاني Deanxit التضخم من الأرق وغيرها من الآثار الجانبية. يستخدم العلاج أساسا لمكتئاب. تأثيره في حالات الذهان ضعيف جدا. لا تتوقف عن تناول الدواء دون ابلاغ طبيبك. لا تتناول أي مكملات أو مضادات أو كحوليات جدا إلى جانب مع هذا الدواء. قد تتفاعل سلبا مع بعضها البعض. لا يستخدم هذا الدواء في الولايات المتحدة، المملكة المتحدة، الإيدان، كندا، أستراليا، الهند، أو أيرلندا. بالإضافة إلى ذلك، تم حظره أيضا في الدنمارك - بلد المنشأ، ويرجع ذلك إلى ظهور آثار جانبية عصبية خطيرة. في أي حال، يستخدم هذا الدواء فقط لفترة قصيرة.	['B']	['1']
283	السلام عليكم اذا اتناول دواء بروزاك والضعر بتحسن كبير هل تعود اعراض الاكتئاب بعد التوقف عن الدواء؟	لا بد من استمرار العلاج سنه بعد التحسن بإشراف الطبيب	['B']	['1']

Figure 1: Data samples from MentalQA.

We participated in Sub-task 3: Question Answering, which requires generating expert-level answers to Arabic mental health questions. This task builds on the earlier classification sub-tasks and aims to develop systems capable of providing accurate and useful responses. The official evaluation metric used for Sub-task 3 is BERTScore (F1).

While recent studies have begun to explore Arabic NLP for mental health, prior work has primarily focused on resource creation, small-scale evaluations, or application-level prototypes. Building on these efforts, our contribution is to systematically evaluate multiple large language models on the AraHealthQA dataset and to analyze differences in response quality and their alignment with expert-written answers in the Arabic psychological domain.

4 System Description

Our system follows a structured prompt-based generation workflow using pre-trained large language models (LLMs) without any fine-tuning. The process which consists of four stages: (1) data prepa-

ration, (2) prompt design, (3) model setup, and (4) evaluation, was provided in Appendix Figure 2

4.1 Data Preparation

We used the AraHealthQA Subtask 3 dataset, which contains 350 samples for training and development, and 150 samples for testing. All samples were kept in Arabic to preserve cultural and linguistic nuances. The dataset was cleaned, and minor inconsistencies were corrected to ensure reliability, and example selection ensured topical diversity and cultural appropriateness.

4.2 Prompt Design & Strategies

Prompts were designed using real question–answer pairs from the dataset. We experimented with:

- Zero-shot
- Few-shot (3-shot, 5-shot)
- Chain-of-thought (CoT) (Wei et al., 2022)
- Self-consistency (Wang et al., 2022)
- Ensemble refinement

Zero-shot achieved a BERTScore F1 of 0.61, while 3-shot improved to 0.66. Self-consistency with 3-shot produced stable results, but 5-shot and CoT slightly degraded performance. Ensemble refinement did not improve scores.

4.3 Model Setup

The final configuration fixed the 3-shot prompt format across all models. No external data beyond the provided samples were used. Models included:

- GPT-4o (OpenAI, 2024)
- Gemini (Team et al., 2023)
- Allam (Bari et al., 2024)
- Qwen (Qwen Team, 2024)

Models were accessed via public APIs or Hugging Face, and all runs used fixed seeds for reproducibility.

4.4 Evaluation

For each test question, a 3-shot prompt was dynamically constructed. Model outputs were compared against expert-written answers using BERTScore F1 (Zhang et al., 2019). GPT-4o achieved the highest balance between accuracy and empathy, Gemini was empathetic but less precise, Allam favored technical terminology, and Qwen tended toward generic responses.

5 Experimental Setup

Data Split Usage

For Subtask 3, the organizers released 350 annotated samples for training and development, and 150 samples as a hidden test set (Table 1). Each entry contains: (1) the question, (2) the expert-written answer, (3) the question type, and (4) the answer strategy. Question types include *diagnosis*, *treatment*, *epidemiology*, and *healthy lifestyle*, while answer strategies are *informational*, *direct guidance*, and *emotional support*.

Table 1: MentalQA dataset distribution for Subtask 3.

	Train/Dev	Test	Total
Samples	350	150	500

From the training split, we selected 10 representative question–answer pairs covering all question types and answer strategies to construct prompting examples. These examples were fixed and reused across all prompting strategies to ensure fair comparisons. Final evaluation was conducted on the entire hidden test set.

External Tools and Libraries

All models were used in their original form without fine-tuning:

- **GPT-4o** and **Gemini**: accessed via their official APIs (accessed on 20 July 2025).
- **Allam** and **Qwen**: accessed via Hugging Face Inference API (accessed on 20 July 2025).
- **LLaMA-3-70B-Instruct (Grattafiori et al., 2024)**: accessed via Groq API (Groq, 2024) for prompt evaluation (accessed on 20 July 2025).

Table 2 summarizes the full prompting configurations used for each model.

Model	Temp.	Top-p	Max tokens
GPT-4o	0.1	0.9	1024
Gemini 2.5	0.1	0.9	1024
ALLaM-7B	0.1	0.9	1024
Qwen2.5-7B	0.1	0.9	1024
LLaMA-3 70B	0.1	0.9	1024

Table 2: Prompting parameters used across models.

Key libraries

- Hugging Face Hub version: 0.34.3
- BERTScore v0.3.11
- openai v0.28
- Google Generative AI version: 0.8.5
- Python 3.11.13

Evaluation Metric

We used BERTScore F1 (Zhang et al., 2019) with the multilingual model to compare system outputs against expert-written answers. Scores were computed using the official `bert_score` implementation (v0.3.11) with default multilingual settings for Arabic. This metric measures semantic similarity between generated answers and references, accounting for lexical and contextual matches.

Detailed results and prompt strategy that used are shown in Appendix Figure 3

6 Results

Our final system, which used GPT-4o with 3-shot prompting, achieved a BERTScore F1 of 0.67 on the official test set and was ranked 2nd overall in Sub-task 3 of the AraHealthQA 2025 shared task (Alhuzali et al., 2025).

The full results of model comparisons and prompting strategies are presented in Appendix Table 4 and Table 3

Table 3: BERTScore F1 performance of different LLMs on the official train set (3-shot prompting).

Model	BERTScore F1
GPT-4o	0.6551
Allam	0.6316
Gemini	0.6210
Qwen	0.6131

Table 4: BERTScore F1 performance across different prompting strategies, evaluated using LLaMA-3-70B-Instruct.

Prompting Strategy	BERTScore F1
Zero-shot	0.6100
3-shot	0.6600
3-shot + self-consistency	0.6600
Few-shot (5-shot)	0.6400
Chain-of-thought	0.6150
3-shot + ensemble refinement	0.6100

LLaMA-3-70B-Instruct was used only as a reference model to compare prompting strategies (Table 4) and was not included in Table 3, since our leaderboard submission relied on other models.

In the development phase, we conducted extensive ablation studies to compare various prompting strategies across multiple models. 3-shot prompting consistently outperformed zero-shot, 5-shot, and more complex techniques such as chain-of-thought reasoning, self-consistency, and ensemble refinement. While chain-of-thought prompting introduced more structured reasoning, it slightly decreased performance on BERTScore metrics. Increasing to 5-shot did not yield additional benefit and often produced redundant outputs. As a result, 3-shot prompting was selected for its superior performance and simplicity.

In the development phase, we conducted extensive ablation studies to compare various prompting strategies across multiple models. 3-shot prompting consistently outperformed zero-shot, 5-shot, and more complex techniques such as chain-of-thought reasoning, self-consistency, and ensemble refinement. While chain-of-thought prompting introduced more structured reasoning, it slightly decreased performance on BERTScore metrics. Increasing to 5-shot did not yield additional benefits and often produced redundant outputs. As a result, 3-shot prompting was selected for its superior performance and simplicity. No major hallucinations or foreign-language artifacts were observed in the generated answers. Notably, the selected model (GPT-4o) avoided making explicit diagnostic claims or recommending specific medical treatments. Instead, the system provided general guidance, informative responses, and help-seeking suggestions — a desirable behavior for mental health applications where only qualified professionals should deliver clinical diagnoses or therapeutic

interventions. This aligns well with the task's goal of producing educational and supportive content without overstepping ethical boundaries. All reported results are based on the official submission. No post-submission modifications or evaluations were performed.

All reported results are based on the official submission. No post-submission modifications or evaluations were performed.

7 Limitations

The dataset is relatively limited in size, which restricts the ability to generalize the findings. As a result, there's a need to expand the database in the future. While the BERTScore F1 serves as a useful metric for quantitative assessment, relying solely on it falls short of capturing critical elements such as empathy, safety, and cultural nuances. To address this, we plan to implement a more holistic set of evaluation standards moving forward. These will encompass emotional factors, health relevance, contextual appropriateness, harm prevention, and risk awareness. We aim to combine the LLM-as-a-Judge framework with human judgment to produce outcomes that are both more trustworthy and grounded in real-world considerations.

8 Conclusion and future work

In this work, we presented a prompt-based question answering system for Arabic mental health queries, developed as part of the AraHealthQA 2025 shared task. Our final system was built on GPT-4o using 3-shot prompting with carefully selected examples from the training data. The system demonstrated the ability to generate coherent, informative, and non-diagnostic responses that were consistent with the expert-written reference answers provided in the dataset.

For future work, we plan to explore fine-tuning Arabic LLMs on the full dataset to enhance contextual alignment, as well as investigate retrieval-augmented generation (RAG) techniques to integrate external knowledge sources and improve factual accuracy in complex queries. We also intend to involve mental health professionals in the evaluation process to assess the psychological appropriateness and safety of model-generated answers.

9 Acknowledgments

We thank the organizers of the AraHealthQA 2025 shared task for providing the dataset and evaluation platform.

References

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

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

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*.

M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani, Nouf M Alotaibi, Hisham A Alyahya, Sultan Al-Rashed, Faisal A Mirza, Shaykhah Z Alsubaie, Hassan A Alahmed, Ghadah Alabduljabbar, and 1 others. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.

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)*, pages 1–6. IEEE.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Inc. Groq. 2024. Meta Llama 3 70b (8192) served via groq api. Accessed: 2025-07-20.

OpenAI. 2024. GPT-4o: Openai's omnimodal model. Accessed: 2025-07-25.

Vikram Patel, Shekhar Saxena, Crick Lund, Graham Thornicroft, Florence Baingana, Paul Bolton, Dan Chisholm, Pamela Y. Collins, Janice L. Cooper, Julian Eaton, Helen Herrman, Mazen M. Herzallah, Yu Huang, Mark J. D. Jordans, Arthur Kleinman, María Elena Medina-Mora, Graham Morgan, Unaiza Niaz, Oye Gureje Omigbodun, and 9 others. 2018. The lancet commission on global mental health and sustainable development. *The Lancet*, 392(10157):1553–1598.

Qwen Team. 2024. Qwen2.5: A party of foundation models. Accessed: 2025-07-25.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

World Health Organization. 2022. World mental health report: Transforming mental health for all. Accessed: 2025-07-02.

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 arXiv:2501.06859*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

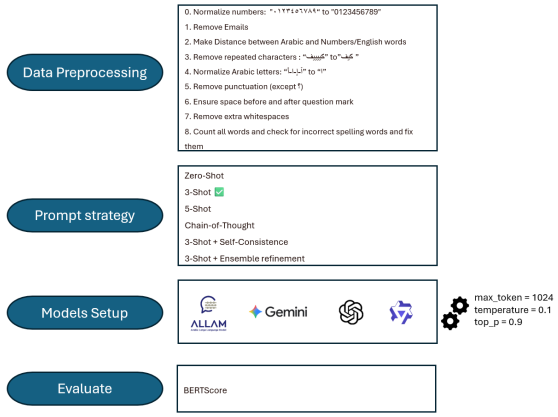

Figure 2: System pipeline

B Appendix

Figure 3: An illustrative example from the MentalQA dataset showing the question, gold reference, and generated answers using prompting strategies (3-shot).

C Appendix

All implementation details, including full prompt examples and evaluation scripts, are available in our GitHub repository: https://github.com/njoudae/AraHealthQA_2025_subtasck_3/tree/main.