NYUAD at AraHealthQA Shared Task: Benchmarking the Medical Understanding and Reasoning of Large Language Models in Arabic Healthcare Tasks

Nouar AlDahoul
Computer Science Department
New York University
Abu Dhabi, UAE
nouar.aldahoul@nyu.edu

Yasir Zaki
Computer Science Department
New York University
Abu Dhabi, UAE
yasir.zaki@nyu.edu

Abstract

Recent progress in large language models (LLMs) has showcased impressive proficiency in numerous Arabic natural language processing (NLP) applications. Nevertheless, their effectiveness in Arabic medical NLP domains has received limited investigation. This research examines the degree to which state-of-the-art LLMs demonstrate and articulate healthcare knowledge in Arabic, assessing their capabilities across a varied array of Arabic medical tasks. We benchmark several LLMs using a medical dataset proposed in the Arabic NLP AraHealthQA challenge in MedArabiQ2025 track. Various base LLMs were assessed on their ability to accurately provide correct answers from existing choices in multiple-choice questions (MCQs) and fill-in-the-blank scenarios. Additionally, we evaluated the capacity of LLMs in answering open-ended questions aligned with expert answers. Our results reveal significant variations in correct answer prediction accuracy and low variations in semantic alignment of generated answers, highlighting both the potential and limitations of current LLMs in Arabic clinical contexts. Our analysis shows that for MCQs task, the proposed majority voting solution, leveraging three base models (Gemini Flash 2.5, Gemini Pro 2.5, and GPT o3), outperforms others, achieving up to 77% accuracy and securing first place overall in the challenge[1] (Alhuzali et al., 2025). Moreover, for the open-ended questions task, several LLMs were able to demonstrate excellent performance in terms of semantic alignment and achieve a maximum BERTScore of 86.44%.

1 Introduction

Medicine relies heavily on complex reasoning, spanning tasks from diagnostic decision-making to treatment planning, especially when patient outcomes depend on understanding multi-factorial conditions (Qiu et al., 2024; Huang et al., 2025). Differential diagnosis involves generating and narrowing down possible diagnoses using clinical evidence, requiring both extensive medical knowledge and logical reasoning to evaluate multiple hypotheses.

LLMs have demonstrated superior performance across various domains and applications, such as article debiasing (Kuo et al., 2025), content moderation (AlDahoul et al., 2024b), and political leaning detection (AlDahoul et al., 2024a). In the healthcare domain, LLMs are reshaping the landscape of healthcare by transforming the way consultations, diagnoses, and treatment plans are delivered (Yang et al., 2023). They offer new avenues for improving patient education through dynamic, conversational interactions, thereby enhancing both accessibility and patient autonomy. Beyond direct patient care, LLMs also show promise in supporting medical training and streamlining administrative responsibilities, including the generation of clinical notes, referral letters, and discharge summaries (Yang et al., 2023).

Most existing benchmarks focus on English, leaving a gap in evaluating Arabic LLMs for healthcare due to the lack of high-quality clinical datasets, Arabic's linguistic diversity, and the limited performance of multilingual models in domain-specific tasks (Daoud et al., 2025). To fill these gaps, there is an increasing demand for frameworks that evaluate LLM performance in clinical tasks for Arabic-speaking communities. Our analyses and experiments center around the following research questions: **RQ1**: Do state-of-the-art proprietary base LLMs perform well in Arabic medical tasks? **RQ2**: To what extent do state-of-the-art proprietary base LLMs with reasoning capacity excel in Arabic medical tasks? **RQ3**: Do open-source-based Arabic LLMs perform well in Arabic medical tasks? and **RQ4**: How does majority voting among several LLMs enhance performance in Arabic medical

[1] https://www.codabench.org/competitions/8967/#/results-tab

Proceedings of The Third Arabic Natural Language Processing Conference, pages 119–125
November 8-9, 2025

tasks?

We address **RQ1** by running the APIs of several LLMs, such as Claude Opus, Grok 3, Deepseek v3, Llama 4 Maverick, GPT-4o-mini, and GPT-4o. To answer **RQ2**, we utilized APIs of state-of-the-art LLMs with reasoning capabilities such as GPT-o3, Gemini Flash 2.5, and Gemini Pro 2.5. Moreover, to address **RQ3**, we ran Falcon 3, Fanar, and Allam. Additionally, to answer **RQ4**, we calculated the majority vote among the predictions of three LLMs.

2 Related Work

BioBERT (Lee et al., 2020), SCIBERT (Beltagy et al., 2019), and PubMedBERT (Gu et al., 2021) improved biomedical NLP by training on domain-specific corpora, thereby outperforming the general BERT model (Yang et al., 2023). Building on this, ClinicalBERT (Alsentzer et al., 2019) enhanced performance on medical tasks by fine-tuning BERT and BioBERT using the MIMIC-III clinical dataset. Expanding further, GatorTrona (Yang et al., 2022). significantly larger model trained from scratch on extensive clinical and biomedical text—demonstrated strong results across a wide range of clinical NLP tasks (Yang et al., 2023).

Various benchmarks have been developed to evaluate LLMs' proficiency in medical reasoning and knowledge (Huang et al., 2025; Zuo et al., 2025). However, significant challenges persist, ranging from ethical and safety concerns to the risk of biased outputs and inconsistent performance across different languages and cultural settings (Yang et al., 2023; Nazi and Peng, 2024; Daoud et al., 2025).

To advance medical LLMs, researchers have increasingly focused on creating multilingual medical datasets (Qiu et al., 2024). They introduced MMedC, a 25.5-billion-token multilingual medical corpus, and MMedBench, a multilingual QA benchmark with rationales. By fine-tuning Llama 3 (8B), they found it outperformed all other open-source models and approached GPT-4 performance. However, Arabic was not one of the languages included (Qiu et al., 2024).

Arabic medical benchmarks are limited and mostly focused on question-answering tasks. While resources like MMLU (Hendrycks et al., 2020), AraSTEM (Mustapha et al., 2024), and AraMed (Alasmari et al., 2024) offer valuable contributions, they do not fully cover the breadth of Arabic medical tasks, highlighting the need for more comprehensive benchmarking efforts. The previous issue was addressed by the MedArabiQ benchmark (Daoud et al., 2025).

3 Materials and Methods

3.1 Dataset Overview

The medical data used in this work is the main dataset utilized in the AraHealthQA shared task in the MedArabiQ2025 track (Alhuzali et al., 2025) under one of the Arabic NLP challenges. It focuses on modern standard Arabic (MSA) and consists of 700 diverse clinical samples, covering both structured medical knowledge assessments and real-world patient-doctor interactions (Daoud et al., 2025; Alhuzali et al., 2025). The dataset has multiple-choice and open-ended questions that are distributed as follows:

- a random set of 100 multiple-choice questions to evaluate the models' medical understanding.
- a set of 100 multiple-choice questions with bias injected to evaluate how LLMs handle ethical or culturally sensitive scenarios.
- a set of 100 fill-in-the-blank questions with choices to evaluate the model's ability to recognize correct answers, reducing the reliance on generative capabilities.
- a set of 100 fill-in-the-blank questions without choices to assess LLMs' reasoning and generation capabilities.
- a set of 100 patient-doctor Q&As selected from AraMed (Alasmari et al., 2024) to evaluate LLMs with online real-world scenarios from medical discussion forums.
- a 100 Q&As with grammatical error correction to handle inflectional patterns and prepare the dataset for grammatical correction.
- a 100 Q&As with LLM Modifications to mitigate potential model memorization and to assess the model's reasoning and adaptability.

The previous 700 examples were used for evaluation of LLMs. Later, another set of 200 examples (100 MCQs and 100 open-ended questions) was released for testing the LLMs' reasoning and understanding.

3.2 Methods

We have evaluated state-of-the-art base LLMs to identify the best in terms of correct answer match accuracy in MCQs task and alignment score of generated answers in open-ended questions task. This LLM can understand the questions, identify the correct answers utilizing its embedded knowledge and reasoning capability, and generate the answers that align with those of experts.

We started assessing several proprietary base LLMs for the MCQs task to evaluate the accuracy of the match between real and predicted answers. We used LLMs' APIs in the inference mode utilizing two different zero-shot prompts specialized for the MCQs task (Prompt 1 and Prompt 2) shown in the Appendix. The evaluated LLMs are: Gemini Flash 2.5, Gemini Pro 2.5[2] (Team et al., 2023), GPT-4o-mini[3], GPT-4o (Hurst et al., 2024), GPT o3[4], Grok 3[5], Claude 3 Opus[6], Deepseek v3 (Liu et al., 2024), and Llama 4 Maverick[7].

Later, we selected the two LLMs that have shown high performance in the MCQs task: Gemini Flash 2.5 and Gemini Pro 2.5 and utilized them in the open-ended question task. We also demonstrated the performance of small-sized LLMs such as GPT-4o-mini in this task. We utilized three different prompts specialized for open-ended tasks (Prompt 1, Prompt 2, and Prompt 3) which are also shown in the Appendix.

Additionally, open-source-based Arabic LLMs such as Falcon3 (Almazrouei et al., 2023) ("tiiuae/Falcon3-7B-Instruct")[8,9], Fanar (Team et al., 2025) ("QCRI/Fanar-1-9B-Instruct")[10], and Allam (Bari et al., 2024)("ALLaM-AI/ALLaM-7B-Instruct-preview")[11] were assessed for both tasks.

We applied zero-shot prompting across all models and tasks, setting the temperature to 0 and top_p to 1 for all tasks to ensure deterministic responses. For the open-ended question task, BERTScore was used as an evaluation metric to measure alignment between generated and expert answers. For this purpose, we used the "XLM-RoBERTa-Large model" (Daoud et al., 2025), which was trained on multiple languages, including Arabic.

We also evaluated Arabic Falcon[12]. Since there is no API available for Arabic Falcon, we used the web interface to manually input questions into the chat version. We retained the history of previous questions to avoid clearing the context before each new query.

3.3 Results and Discussion

The results of the MCQs task using the proprietary LLMs are shown in Table 1. The dataset has MCQs related to understanding and reasoning. While understanding involves factual knowledge, reasoning mimics how doctors make decisions.

The medical reasoning capacity of GPT-o3, Gemini Flash 2.5, and Gemini Pro 2.5 makes them have superior performance compared to other LLMs. These simulate diagnostic thinking by combining multiple facts and using step-by-step reasoning to eliminate plausible but incorrect distractors in medical MCQs, which answers **RQ2**.

Model	Prompt	Accuracy%
GPT-4o-mini	1	49
GPT-4o	1	57
GPT-O3	1	72
Gemini Flash 2.5	1	73
Gemini Pro 2.5	1	75
GPT-O3	2	74
Gemini Flash 2.5	2	74
Gemini Pro 2.5	2	76
Majority voting	2	77
Grok 3	2	60
Claude 3 Opus	2	49
Falcon Arabic	2	38
Deepseek v3	2	56
Llama 4 Maverick	2	63

Table 1: Accuracy of different proprietary base LLMs using different prompts.

Even though Claude 3, Deepseek 3, Grok 3, and

[2] https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking

[3] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[4] https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf

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

[6] https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

[7] https://ai.meta.com/blog/llama-4-multimodal-intelligence/

[8] https://huggingface.co/blog/falcon3

[9] https://huggingface.co/tiiuae/Falcon3-7B-Instruct

[10] https://huggingface.co/QCRI/Fanar-1-9B-Instruct

[11] https://huggingface.co/ALLaM-AI/ALLaM-7B-Instruct-preview

[12] https://falcon-lm.github.io/blog/falcon-arabic/

Llama 4 Maverick possess strong reasoning capabilities, they exhibit modest performance on this task, likely due to limited medical knowledge or insufficient proficiency in Arabic, which addresses **RQ1** and **RQ2**. However, Llama 4 Maverick was the best among them in terms of accuracy (63%).

For sensitivity of prompt construction, we found that Prompt 2, which includes step-by-step or chain-of-thought reasoning, is generally better than simple Prompt 1 when it comes to answering medical MCQs.

The significant finding in this work is that current state-of-the-art proprietary LLMs exhibit limitations in their embedded medical knowledge of various Arabic medical tasks (maximum accuracy is 76% in Gemini Pro 2.5). The source of errors in the MCQ task may stem from misunderstanding of questions, lack of medical knowledge, or lack of medical reasoning capabilities.

To benefit from the capacity of each of three LLMs (GPT-O3, Gemini Flash 2.5, and Gemini Pro 2.5) in MCQs task, we applied a majority voting technique using the predictions from these LLMs, resulting in a final accuracy of 77%, which secured first place overall in the challenge, which answers **RQ4**.

The results of the open-ended questions task using proprietary LLMs are shown in Table 2. The dataset has questions labeled with answers. The LLMs should generate answers that are semantically aligned with reference answers.

Our finding indicates that reasoning LLMs such as Gemini Flash 2.5 and Gemini Pro 2.5 have structured answers that reduce hallucination and overconfidence, as the models are less likely to guess and more likely to justify their answers. As a result, their responses often align more closely with reference answers and perform better on semantic evaluation metrics like BERTScore, which answers **RQ2**. Furthermore, GPT-4o-mini shows good performance in terms of BERTScore.

Additionally, the three LLMs showed high sensitivity to prompts with variances in BERTScores. The maximum BERTScores were achieved by Prompt 3 that asked the LLMs to have modern standard Arabic in response, emphasized medically correct answers, and asked for concise answers that are not diluted with explanations, which usually tend to align more closely with reference answers.

Table 3 shows the accuracy and BERTScore of several open-source base Arabic LLMs. Among the models, Allam demonstrates relatively better performance (39%) in MCQs task, while Falcon 3 gave the best BERTScore (0.8493). This experiment indicates a lack of medical knowledge and/or medical reasoning in the base open-source Arabic LLMs compared to proprietary ones, which addresses **RQ3**.

Model	Prompt	BERTScore
Gemini Pro 2.5	1	0.8105
Gemini Flash 2.5	2	0.8364
GPT-4o-mini	2	0.8386
GPT-4o-mini	3	0.8581
Gemini Flash 2.5	3	0.8633
Gemini Pro 2.5	3	0.8644

Table 2: BERTScore of proprietary base LLMs using different prompts.

Model	Task	Accuracy %
Falcon 3	Task 1	36
Fanar	Task 1	31
Allam	Task 1	39
Model	**Task**	**BERTScore**
Falcon 3	Task 2	0.8493
Fanar	Task 2	0.8403
Allam	Task 2	0.8431

Table 3: Accuracy and BERTScore of different base Arabic LLMs.

Limitations

The first limitation is that multiple-choice and fill-in-the-blank with choice questions in the MedArabiQ2025 dataset are limited to only a few hundred examples. There is a clear need for larger, high-quality Arabic medical datasets to fine-tune LLMs and enhance their performance. Alternatively, storing extensive medical data in a vector database and employing retrieval-augmented generation (RAG) techniques could help retrieve more accurate and contextually relevant answers.

A second limitation of this work is the absence of bias detection and mitigation techniques during the preprocessing of questions before inputting them to LLMs. Incorporating such techniques could play a significant role in improving model performance and ensuring more reliable outputs.

The third limitation is that for open-ended and fill-in-the-blank questions without choices, we lack a robust metric for capturing semantic similarity.

In this work, we utilized BERTScore, which often yields similar values across different responses and fails to reflect subtle nuances in semantic alignment with the correct answers.

References

Ashwag Alasmari, Sarah Alhumoud, and Waad Alshammari. 2024. Aramed: Arabic medical question answering using pretrained transformer language models. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 50–56.

Nouar AlDahoul, Talal Rahwan, and Yasir Zaki. 2024a. Polytc: a novel bert-based classifier to detect political leaning of youtube videos based on their titles. *Journal of Big Data*, 11(1):80.

Nouar AlDahoul, Myles Joshua Toledo Tan, Harishwar Reddy Kasireddy, and Yasir Zaki. 2024b. Advancing content moderation: Evaluating large language models for detecting sensitive content across text, images, and videos. *arXiv preprint arXiv:2411.17123*.

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

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

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

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

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. *arXiv preprint arXiv:2505.03427*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. O1 replication journey–part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Chen Wei Kuo, Kevin Chu, Nouar AlDahoul, Hazem Ibrahim, Talal Rahwan, and Yasir Zaki. 2025. Neutralizing the narrative: Ai-powered debiasing of online news articles. *arXiv preprint arXiv:2504.03520*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ahmad Mustapha, Hadi Al-Khansa, Hadi Al-Mubasher, Aya Mourad, Ranam Hamoud, Hasan El-Husseini, Marwah Al-Sakkaf, and Mariette Awad. 2024. Arastem: A native arabic multiple choice question benchmark for evaluating llms knowledge in stem subjects. *arXiv preprint arXiv:2501.00559*.

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

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, et al. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

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

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Xi Yang, A Chen, N PourNejatian, HC Shin, KE Smith, C Parisien, C Compas, C Martin, AB Costa, MG Flores, et al. 2022. A large language model for electronic health records. npj digital medicine, 5 (1), 1–9. *URL http://dx. doi. org/10.1038/s41746-022-00742-2*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Appendix

A.1 Prompts used

This work employed different prompts for the two tasks: the orange-colored prompts were used for Task 1 (MCQs), while the red, green, and blue prompts were used for Task 2 (open-ended questions).

MCQs Prompt 1

Answer this multiple choice question in Arabic using a single letter: أ, ب, ج, د, ه

MCQs Prompt 2

السؤال: [ضع هنا السؤال الطبي متعدد الخيارات مع الخيارات أ, ب, ج, د, ه]

الخطوات: فكر خطوة بخطوة لتحديد الخيار الصحيح.

اشرح السياق الطبي والمصطلحات إن لزم، وحدد أي الخيارات خاطئة أو غير منطقية. استند إلى المعلومات الطبية المعروفة

الإجابة النهائية: أعطني فقط الحرف الصحيح للإجابة (أ, ب, ج, د, أو ه) في السطر الأخير

Open-ended questions' Prompt 1

You are a knowledgeable and concise medical expert. Provide a high-quality answer to the following open-ended medical question.
Your response should:
Begin with a direct, evidence-based answer.
Elaborate on the mechanisms, relevant anatomy or physiology, and clinical significance.
Use clear, professional medical language.
Question:
[Insert your medical question here]

Open-ended questions' Prompt 2

You are a knowledgeable and concise medical expert. Provide a high-quality answer to the following open-ended medical question.

Open-ended questions' Prompt 3

You are a knowledgeable and concise medical expert.
Your task is to generate a concise, accurate, and medically correct answer in Modern Standard Arabic.
Do not include explanations—just provide the best possible answer based on your knowledge.