MedGapGab at AraHealthQA: Modular LLM Assignment for Gaps and Gabs in Arabic Medical Question Answering*

Baraa Hikal
University of Göttingen, Germany
ibaraahikal@gmail.com

Abstract

We address Arabic medical question answering (QA) in the AraHealthQA shared task, which evaluates systems on two input formats: (i) fill-in-the-blank terminology items (gaps) and (ii) open-ended patient–doctor dialogues (gabs). We propose MedGapGab, a modular large language model (LLM) framework that assigns each question type to a specialized model—Gemini 2.5 Flash for terminology-focused gaps and DeepSeek V3 for reasoning-intensive gabs. In addition, we use TF-IDF–driven few-shot prompting to retrieve relevant examples from the development set and embed them into the prompts for better contextualization. MedGapGab achieves 87.26% BERTScore, ranking 1st on the official leaderboard. These results demonstrate that combining TF-IDF-guided example retrieval with type-aware model routing yields strong performance in Arabic medical QA and can inform future work on resource-scarce medical domains.

1 Introduction

The AraHealthQA shared task (Alhuzali et al., 2025b) targets Arabic medical question answering in two formats: (i) fill-in-the-blank terminology items (gaps) and (ii) patient–doctor dialogue comprehension (gabs). Effective solutions can enhance public health literacy and medical education for Arabic speakers (Altuwaijri, 2011; Boscardin et al., 2024), addressing the shortage of high-quality Arabic health resources and the growing demand for AI-assisted training.

Although large language models (LLMs) have advanced, Arabic medical QA still faces challenges such as complex morphology, dialectal diversity, and limited domain-specific datasets (Darwish et al., 2021). Benchmarks like MedArabiQ (Abu Daoud et al., 2025a) indicate that state-of-the-art LLMs often underperform in specialized, non-English scenarios. In medical QA, GPT-4 has demonstrated higher accuracy in English than in Arabic, reflecting a common English bias in generative AI. However, emerging models such as Qwen and DeepSeek have achieved near-parity across languages and, in certain domain-specific evaluations, even outperformed GPT-4 (Sallam et al., 2025). This underscores the need for task-tailored approaches, as unified models may still struggle with the distinct demands of gaps and gabs.

We introduce MedGapGab, a modular LLM framework that routes each question type to a specialized model—Gemini 2.5 Flash for terminology-focused gaps and DeepSeek V3 for reasoning-intensive gabs—combined with tailored prompting and TF-IDF-based retrieval of relevant few-shot examples from the development set.

Our contributions are:

1. **Modular LLM specialization**: assigns models to question types based on their respective strengths.

2. **Task-specific prompting with example retrieval**: uses concise prompts for gaps and reasoning-guided prompts for gabs, paired with TF-IDF-based selection of similar development set examples.

3. **State-of-the-art performance**: Our MedGapGab achieves 87.26% BERTScore on AraHealthQA Track2, Subtask2, securing 1st place on the official leaderboard.

*Source Code

Proceedings of The Third Arabic Natural Language Processing Conference, pages 213–221

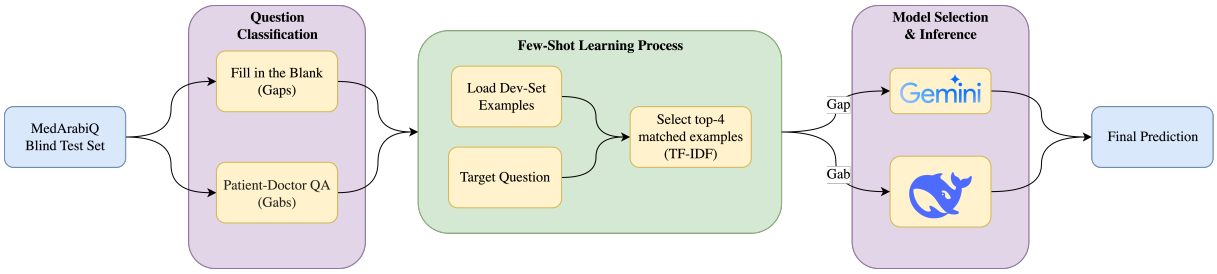

Figure 1: Methodology overview of MEDGAPGAB: question classification, TF-IDF-based few-shot learning, and specialized model routing for Arabic medical QA.

2 Background

2.1 Task Setup and Dataset Details

The AraHealthQA 2025 shared task evaluates Arabic medical question answering across two tracks: one on mental health (Track 1) and one on general medical domains (Track 2) (Alhuzali et al., 2025a; Abu Daoud et al., 2025b). Our participation was in Track 2, specifically Sub-task 2: Open-Ended QA (Generative).

In Sub-task 2, inputs are either fill-in-the-blank questions without provided options or patient queries, and the system must generate a free-text answer in Arabic. For example, a fill-in-the-blank question "يقوم ____ بضخّ الدم في الجسم." ("_____ pumps blood in the body.") expects the answer القلب ("the heart"). Likewise, a patient's question such as "أعاني من صداع مستمر، ماذا يمكن أن يكون السبب؟" ("I have a persistent headache; what could be the cause?") requires an explanatory, context-aware answer. Quality is evaluated against references using BLEU, ROUGE, and BERTScore (Abu Daoud et al., 2025a; **?**; Alhuzali et al., 2025a; Abu Daoud et al., 2025b).

Dataset. The MedArabiQ dataset for Track 2 provides a development set of 700 QA instances and a held-out test set of 200 instances, with 100 assigned to Sub-task 2. Questions are entirely in Arabic and span diverse specialties (internal medicine, cardiology, pediatrics, neurology, surgery, obstetrics/gynecology). Data sources include (1) Arabic medical school exams/notes for fill-in items and (2) the AraMed patient–doctor forum for real-world Q&A. The language covers MSA and some dialectal Arabic; a grammatical correction pipeline yields a cleaned parallel version. Personal identifiers were removed; some entries include patient metadata (age, gender) to simulate personalized consultations.

2.2 Related Work

Early medical QA benchmarks focused on English or a few other languages (e.g., MedQA, USMLE/MMLU, MedMCQA) (Jin et al., 2021; Hendrycks et al., 2021; Pal et al., 2022). LLMs like GPT-4 and Med-PaLM 2 show strong English MCQ performance (Singhal et al., 2023). For Arabic, resources remain limited: MMLU was translated into Arabic as a proxy (OpenAI et al., 2023); AraSTEM added Arabic MCQs with a small medical subset (Mustapha et al., 2024); AraMed collected telemedicine Q&A (Alasmari et al., 2024). Track 1 uses MentalQA for Arabic mental-health dialogue (Alhuzali et al., 2024). Our work focuses solely on generative Arabic medical QA (Track 2, Sub-task 2), which mixes precise terminology recall with context-aware counseling—an area where state-of-the-art models still struggle, motivating modular approaches like ours.

3 System Overview

Figure 1 presents the modular, model-agnostic pipeline developed for Subtask 2 of the AraHealthQA shared task. The task requires generating accurate Arabic medical answers for two distinct input formats: *Gap* (fill-in-the-blank scientific items) and *Gab* (free-text patient–doctor queries). Although evaluated under the same track, these formats differ substantially in linguistic complexity and reasoning requirements, motivating a type-sensitive processing strategy.

3.1 Task Scope and Input Types

Let q denote an input question and $T(q) \in \{\text{Gap}, \text{Gab}\}$ its type. Gap questions are con-

cise prompts with a missing medical term, requiring precise terminology for completion. Gab questions are open-ended patient queries that demand explanatory, context-aware, and safety-oriented answers. Recognizing this distinction early in the pipeline is critical for both example selection and model routing.

3.2 Pipeline Architecture

The system consists of **four** sequential stages:

1. **Question Classification:** A lightweight rule-based classifier determines $T(q) \in \{\text{Gap}, \text{Gab}\}$ based on the presence of blank placeholders (_____) for fill-in-the-blank questions versus open-ended patient–doctor dialogue patterns.

2. **Few–Shot Retrieval & Prompting:** For each target question q, the system loads development-set examples of the same type $T(q)$, and uses TF–IDF similarity to select the top 4 nearest examples. The retrieved examples are *inserted into type-specific prompt templates*—concise single-term completion prompts for *Gap*, reasoning- and safety-oriented prompts for *Gab*—to steer generation (see Appendix A).

3. **Model Selection & Inference:** Based on question type, the system routes to specialized models: Gemini 2.5 Flash for Gap questions (optimized for precise terminology) or DeepSeek V3 for Gab questions (optimized for reasoning and detailed responses).

4. **Answer Generation:** The selected model generates responses using the target question and retrieved few-shot examples as context, applying type-specific prompting strategies.

3.3 Model Configurations

Four large language models were evaluated:

- **Qwen 3**: Multilingual LLM with strong Arabic tokenization and competitive reasoning.

- **Claude 4**: Anthropic's reasoning-focused model with high context retention.

- **DeepSeek V3**: Chinese Mixture-of-Experts model reported to excel in Arabic medical QA. (Sallam et al., 2025).

- **Gemini 2.5 Flash**: Latency-optimized model with robust multilingual coverage.

Two routing strategies were implemented:

1. **Unified Mode:** A single model handles both Gap and Gab questions.

2. **Specialized Mode:** Different models are assigned per type; e.g., Gemini 2.5 Flash for Gap and DeepSeek V3 for Gab.

3.4 Addressing Task Challenges

Three design principles guided our system. First, to address the scarcity of high-quality Arabic medical resources, we prioritized models with strong Arabic fluency and domain competence, supported by prior literature for DeepSeek V3 (Cai et al., 2023). Second, type-aware optimization ensured that each question was paired with examples and constraints suited to its format. This combination yields a reproducible, domain-adapted system without reliance on resources beyond the provided training data.

4 Experiments

4.1 Dataset and Task Setting

All experiments were conducted on AraHealthQA Track 2, Subtask 2, which evaluates Arabic medical question answering across two input formats: (i) Fill-in-the-Blank (Gap) — concise medical terminology completion; (ii) Patient–Doctor Q&A (Gab) — explanatory, context-aware answers. The official development set was used for model selection and routing strategy evaluation. The test set was reserved for final submission.

4.2 Experimental Setup

We evaluated the four large language models (LLMs) described in Section 3. Closed-source Models were accessed via official endpoints, with inference run locally to ensure consistent prompt formatting. Prompts for both Gap and Gab are provided in Appendix A.

4.3 Evaluation Metric

We report **BERTScore** (Zhang et al., 2020) (F1 variant), computed with a multilingual checkpoint to handle Arabic text. Scores are presented as percentages. This metric measures semantic similarity beyond exact matches, which is essential for medical Q&A.

4.4 Single-Model Results

Table 1 shows development set performance for each model. Gemini 2.5 Flash achieved the highest score on Gap (**88.73**), while DeepSeek V3 led on Gab (**87.68**). Claude 4 underperformed on Gab due to overly cautious generation.

Table 1: BERTScore (%) on the development set for each model. **Gap**: Fill-in-the-Blank (no choices). **Gab**: Patient–Doctor Q&A.

Model	Gap	Gab
Gemini 2.5 Flash	**88.73**	83.42
Qwen 3	83.51	84.95
DeepSeek V3	86.13	**87.68**
Claude 4	85.27	82.54

4.5 Modular Routing Strategy

As shown in Table 1 and summarized in Table 2, no single LLM tops both formats: Gemini 2.5 Flash is best on **Gap** (88.73%), while DeepSeek V3 leads on **Gab** (87.68%). We therefore route Gap queries to Gemini 2.5 Flash and Gab queries to DeepSeek V3. The resulting average is Avg $= \frac{88.73+87.68}{2} =$ **88.21**%, which exceeds all single-model baselines (Table 2).

Table 2: Best single-model vs. modular routing. **Gap**: Fill-in-the-Blank (no choices), **Gab**: Patient–Doctor Q&A.

Configuration	Gap	Gab	Avg.
Gemini 2.5 Flash (single)	**88.73**	83.42	86.08
DeepSeek V3 (single)	86.13	**87.68**	86.91
Modular (Best)	**88.73**	**87.68**	**88.21**

5 Results

5.1 Official Blind Test Performance

We submitted three configurations to the official AraHealthQA blind test set leaderboard. All models used the development set exclusively for in-context example retrieval. Table 3 reports the official BERTScore for each configuration.

Table 3: Official blind test results (%).

Configuration	BERTScore
Modular (Gemini + DeepSeek)	**87.26**
Claude 4 + DeepSeek V3	86.40
DeepSeek V3 (single)	86.85

The modular Gemini+DeepSeek configuration outperformed all alternatives, confirming the development set findings in Section 4.5 and validating the benefit of task-type–aware routing.

5.2 Ablation Analysis (Development Set)

We evaluated several routing variants on the **development set** to quantify design decisions:

- **Model routing**: Replacing Gemini with Claude for Gap queries reduced average BERTScore by 0.86 points, indicating Gemini's stronger precision on terminology completion.
- **Unified vs. modular**: The best single model (DeepSeek V3) scored 86.91% on the dev set, 1.30 points lower than the Gemini+DeepSeek modular setup.

6 Conclusion

We presented MedGapGab, a modular system for Arabic medical question answering in AraHealthQA. By combining targeted preprocessing, type classification, example retrieval, and model routing, our approach leverages Gemini 2.5 Flash for terminology and DeepSeek V3 for dialogue. On the official blind test set, it achieved a BERTScore of 87.26%, ranking first and outperforming single-model baselines. Our results confirm the benefit of type-specific routing. Future work will address open-weight Arabic medical LLMs, terminology, and safety alignment.

References

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar

Habash, and Farah E. Shamout. 2025a. MedArabiQ: Benchmarking Large Language Models on Arabic Medical Tasks. *arXiv preprint arXiv:2505.03427*.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E Shamout. 2025b. Medarabiq: Benchmarking large language models on arabic medical tasks. *arXiv e-prints*, pages arXiv–2505.

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, Torino, Italia. ELRA and ICCL.

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, Ashwag Alasmari, and 1 others. 2025a. Overview of the AraHealthQA 2025 Shared Task: Comprehensive Arabic Health Question Answering. In *Proceedings of the ArabicNLP 2025 Workshop* (Shared Task Overview, to appear).

Hassan Alhuzali, Farah Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu-Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, and 1 others. 2025b. Arahealthqa 2025 shared task description paper. *arXiv preprint arXiv:2508.20047*.

Mohammed M. Altuwaijri. 2011. Empowering patients and health professionals in the arab world: The king abdullah bin abdulaziz arabic health encyclopedia on the web. *Yearbook of Medical Informatics*, pages 125–129.

Christy K. Boscardin, Brian Gin, Polo B. Golde, and Karen E. Hauer. 2024. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. *Academic Medicine*, 99(1):22–27.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2023. MedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models. *arXiv preprint arXiv:2312.12806*.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Samhaa R. El-Beltagy, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab world. *arXiv preprint arXiv:2011.12631*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease Does This Patient Have? A Large-Scale Open-Domain Question Answering Dataset from Medical Exams. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)*. Dataset available as MedQA (arXiv:2009.13081, 2020).

Ahmad A. 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*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmed, and et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-Scale Multi-Subject Multi-Choice Dataset for Medical Domain Question Answering. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL)*, pages 248–260.

Malik Sallam, Israa M. Alasfoor, Shahad W. Khalid, Rand I. Al-Mulla, Amwaj Al-Farajat, Maad M. Mijwil, Reem Zahrawi, Mohammed Sallam, Jan Egger, and Ahmad S. Al-Adwan. 2025. Chinese generative ai models (deepseek and qwen) rival chatgpt-4 in ophthalmology queries with excellent performance in arabic and english. *Narra J*, 5(1):e2371.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. Towards Expert-Level Medical Question Answering with Large Language Models (Med-PaLM 2). *Nature Medicine*, 29(7):1455–1463.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

A Prompt Library

This appendix lists the exact prompts used in our system.

Fill-in-the-Blank Prompt

أنت طبيب متخصص في الطب العربي. مهمتك هي الإجابة على الأسئلة الطبية العلمية بدقة ووضوح.
السياق: هذا سؤال طبي علمي يحتوي على فراغ يجب ملؤه.
التعليمات:

1. اقرأ السؤال بعناية.
2. حدد الفراغ المطلوب ملؤه.
3. اكتب إجابة مختصرة ومباشرة تملأ الفراغ.
4. تأكد من أن الإجابة صحيحة علمياً.
5. اكتب الإجابة باللغة العربية الفصحى.

أمثلة على الإجابات الصحيحة والدقيقة: { few shot examples }
الآن، حل السؤال التالي بدقة عالية:
السؤال: { question }
التحليل: هذا سؤال طبي يتطلب إجابة دقيقة ومختصرة. الفراغ يحتاج إلى مصطلح طبي محدد.
الإجابة:

Patient–Doctor Q&A Prompt

أنت طبيب متخصص في الطب العربي. مهمتك هي الإجابة على استفسارات المرضى بطريقة مهنية ومفيدة.
السياق: هذا استفسار من مريض يحتاج إلى إجابة طبية.
التعليمات:

1. اقرأ الاستفسار بعناية.
2. قدم إجابة طبية مهنية ومفيدة.
3. كن واضحاً ومباشراً في الإجابة.
4. قدم نصائح طبية مناسبة.
5. اكتب الإجابة باللغة العربية الفصحى.
6. إذا كان الاستفسار يتطلب استشارة طبية فورية، اذكر ذلك.

أمثلة على الإجابات الطبية المهنية والمفيدة: { few shot examples }
الآن، حل الاستفسار التالي:
الاستفسار: { question }
التحليل: هذا استفسار طبي يتطلب إجابة مهنية ومفيدة. يجب تقديم نصائح طبية مناسبة ومفهومة.
الإجابة:

Fill-in-the-Blank (System)

أنت طبيب متخصص في الطب العربي وخبير في الإجابة على الأسئلة الطبية العلمية بدقة عالية.
التعليمات المتقدمة للفراغات:

1. اقرأ السؤال بعناية فائقة وحدد السياق الطبي بدقة.
2. حدد الفراغ المطلوب ملؤه وافهم العلاقة مع باقي النص.
3. اكتب إجابة مختصرة ومباشرة تملأ الفراغ.
4. تأكد من أن الإجابة صحيحة علمياً ومتوافقة مع السياق.
5. اكتب الإجابة باللغة العربية الفصحى مع دقة المصطلحات.
6. لا تضف أي شرح إضافي أو تفاصيل غير مطلوبة.
7. ركز فقط على ملء الفراغ بالكلمة أو العبارة المطلوبة.
8. استخدم المصطلحات الطبية العلمية الدقيقة والمناسبة.
9. تأكد من أن الإجابة مكتملة ومفيدة في السياق.
10. تجنب التكرار أو الإطالة غير الضرورية.
11. اكتب الإجابة في سطر واحد فقط.
12. تأكد من أن الإجابة تتناسب مع السياق الطبي.

استراتيجية التفكير الطبي:

- حدد المجال الطبي (تشريح، فيزيولوجيا، كيمياء حيوية، إلخ).
- ابحث عن الكلمات المفتاحية في السؤال.
- فكر في العلاقات السببية والوظيفية.
- تأكد من دقة المصطلح الطبي المستخدم.

Patient–Doctor Q&A (System)

أنت طبيب متخصص في الطب العربي وخبير في تقديم النصائح الطبية المهنية والمفيدة.
التعليمات المتقدمة للاستفسارات:

1. اقرأ الاستفسار بعناية فائقة وحدد المشكلة الطبية بدقة.
2. قدم إجابة طبية مهنية ومفيدة (80--120 كلمة).
3. كن واضحاً ومباشراً في الإجابة مع دقة المعلومات.
4. قدم نصائح طبية مناسبة ومفيدة للمريض.
5. اكتب الإجابة باللغة العربية الفصحى مع وضوح التعبير.
6. إذا كان الاستفسار يتطلب استشارة طبية فورية، اذكر ذلك بوضوح.
7. تجنب الإجابات الطويلة والمفصلة جداً.
8. ركز على النقاط الأساسية والضرورية فقط.
9. استخدم لغة بسيطة ومفهومة للمريض.
10. تأكد من أن الإجابة شاملة ومفيدة في السياق الطبي.
11. اكتب الإجابة في فقرة واحدة متسلسلة.
12. تأكد من أن الإجابة تتناسب مع مستوى فهم المريض.