Egyhealth at General Arabic Health QA (MedArabiQ): An Enhanced RAG Framework with Large-Scale Arabic Q&A Medical Data

Hossam Amer 
hossamyasseramer@gmail.com

Rawan Tarek Taha
rawantarek516@gmail.com

Gannat Elsayed
gelsayed@nu.edu.eg

Ensaf Hussein Mohamed
EnMohamed@nu.edu.eg

School of Information Technology and Computer Science, Nile University, Giza, Egypt

Abstract

Arabic question-answering (Q/A) chatbots face significant challenges due to the scarcity of large, high-quality datasets and the complexities of the Arabic language, including its rich morphology, multiple dialects, and diverse writing forms. To address these challenges, we implement an enhanced retrieval-augmented generation (RAG) pipeline for Arabic medical chatbots, leveraging a dataset of approximately one million Q/A pairs collected from various Arabic healthcare resources. Experimental results demonstrate that our approach significantly outperforms previous Arabic medical QA systems, improving the quality and relevance of generated answers, with the BERTScore increasing from **0.82** to **0.86**. This work represents a step forward in developing scalable and accurate Arabic medical chatbots.

1 Introduction

Arabic medical question-answering (Q/A) chatbots suffer due to shortage of high-quality Arabic datasets, coupled with the difficult features of the Arabic language. Most existing systems are neither accurate nor contextually precise.

We propose an Advanced Retrieval-augmented generation (RAG) Framework that access External datasets in addition of the . To apply this approach a pipeline consisting of error typo correction, a medical speciality classifier and a rerankeris applied to help improve the answer quality of medical question answering. As follows the structure of the paper discusses the existing literature gaps , system architecture of the system , the experiments done and results obtained from the overall pipeline.

2 Background

AraHealthQA 2025 (Alhuzali et al., 2025) seeks to enhance Arabic medical question answering (QA) by addressing benchmarks pertaining to mental health (Track 1: MentalQA) and general healthcare (Track 2: MedArabiQ). The goals of the shared task focus on creating advanced systems for understanding and accurately responding to healthcare queries in Arabic, advancing Arabic clinical NLP and chatbot technologies.

2.1 Task Setup

The primary task revolves around responding to a clinical question in Arabic by accessing a dataset containing relevant knowledge to formulate a coherent and medically accurate answer using retrieval-augmented generation (RAG). In this open-ended question answering (QA) format, responding to input questions with clinically accurate and naturally sounding answers requires generation.

Example: Input: "كيف يمكن تقليل خطر الإصابة بارتفاع ضغط الدم؟"
Translation: "How can the risk of high blood pressure be reduced?"
Output: يمكن تقليل خطر الإصابة بارتفاع ضغط الدم من خلال اتباع نظام غذائي صحي، وممارسة الرياضة، وتقليل تناول الملح.
Translation: "The risk can be reduced by following a healthy diet, regular exercise, and reducing salt intake."

2.2 Data

AraHealthQA utilizes major Arabic medical QA datasets. The development dataset **MedArabiQ** contains 400 samples, which stem from two Arabic medical school exam and lecture note collections alongside the **AraMed** dataset from AlTibbi, an Arabic online patient-doctor forum.
Whilst also Leveraging Additonal 2 huge datasets **AHD: Arabic healthcare dataset** (Abdelhay et al., 2023) 808,472 Q&A and had 45 different categories **MAQA arabic** (Al-Majmar et al.,

Proceedings of The Third Arabic Natural Language Processing Conference, pages 222–225

Table 1: Summary of related work in Arabic and multilingual medical QA systems.

Goals	Dataset	Strategies	Anticipated Outcomes
General Arabic medical QA	MedArabiQ (Abu Daoud et al., 2025),AHD (Abdelhay et al., 2023), MAQA arabic (Al-Majmar et al., 2024)	The proposed system	BERTScore ∼0.86
Mental health Arabic QA	MentalQA	Multi-label classification; RAG pipeline	F1 ∼0.74; Precision@5 ∼0.068
QA Arabic healthcare	MAQA	Deep learning (Transformers)	Cosine similarity ∼81%; BLEU 58%
QA Arabic religion	Quran QA	Multi-task transfer learning	Varies; accuracy & retrieval
Multilingual biomedical QA	MEDIQA	Transformer-based models	F1 and EM scores 0.5--0.8

2024) 273,174 Q&A had 20 different categories and both have been scrapped from Arabic websites and have 3 columns questions, answers and categories.

The dataset must include comprehensive preprocessing steps such as noise removal, Arabic word normalization, and category harmonization, enabling robust training and evaluation of complex retrieval-augmented large language models.

2.3 Related Work

Develop AraHealthQA 2025 interprets an extensively annotated Arabic clinical datasets holistically alongside the advanced generative models as an innovation. By using its comprehensive multitask framework, it not only tackles problems in the Arabic language and its linguistics with concerns in the healthcare sector and domain, but it also sets an unprecedented mark for research in Arabic medical NLP.

3 System Overview

As shown in Figure 1, our system adopts a modular architecture, whose key components and roles are detailed in this section.

3.1 Pre-processing

Noise Removal Because it was a scraped dataset, it contained a lot of noise such as links, "click here for more" اضغط هنا للمزيد, "read more" اقرأ المزيد, and "figure.png", so these extra words were removed. Additionally, there were some questions and answers entirely in non-Arabic text, which we had to drop.

Arabic Word Normalization To reduce orthographic variability in the Arabic text and ensure consistent representation, a set of normalization rules was applied. All variants of Alef (إ، أ، آ) were mapped to the bare form (ا), and the final Yaa (ي) was replaced with its dotless variant (ى). The elongation character (Tatweel, ـ) was removed.

Category Mapping Due to the different number of categories in both datasets, we had to map the 45 categories of the AHD dataset to the closest categories of the 20 categories of the MAQA dataset.

Embeddings Used bert-base-arabic-camelbert-mix Multilangual embedding model

3.2 Gemma Model

Corrects any typographical errors in the input query, enabling more effective query processing.

3.3 Specialty Classifier

Fine tuned AraBERT (Antoun et al., 2020) which Classifies the query from 20 different medical specialties.

3.4 Query Processing

Optimized using specialty-filtered search. The query is applied only to the top 5 predicted classes from the classifier, which significantly narrows down the vector space and reduces the probability of retrieving irrelevant vectors. The query is then embedded and searched.

3.5 Re-ranking

After retrieving 10 candidate vectors, they are reranked and narrowed down to the top 5 results us-

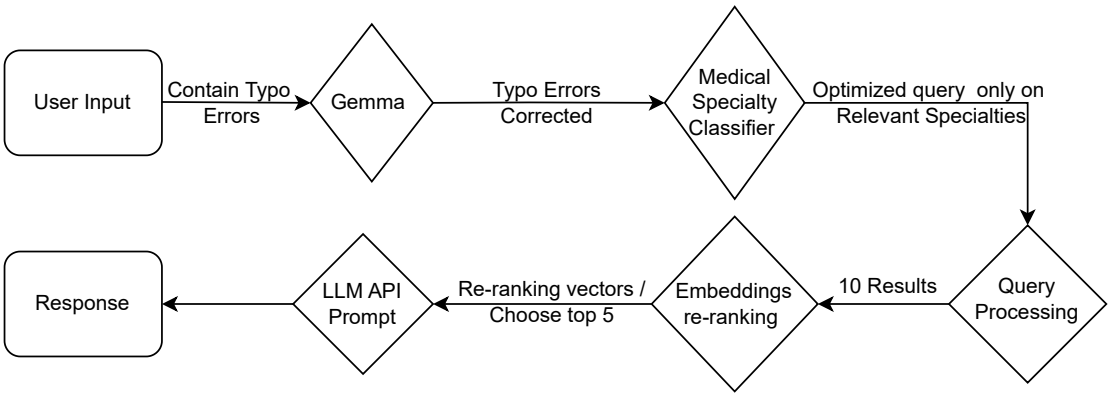

Figure 1: System Architecture

ing a combined similarity score defined as:

$$\text{Score} = 0.7 \cdot \text{CosineSimilarity} + 0.3 \cdot \text{BM25}$$

4 Experimental Setup

4.1 Classifier Fine-Tuning

4.1.1 Training Configurations

Table 2: Hyper-parameters used for classifier fine-tuning.

Parameter	Value
Learning rate	2×10^{-5}
Train batch size	16
Seed	42
Optimizer	AdamW (PyTorch)
Optimizer betas	(0.9, 0.999)
Optimizer epsilon	1×10^{-8}
LR scheduler	Linear
Epochs	3

4.1.2 Classifier results

Considering that many patient questions are interconnected with multiple medical specialties, the F1-score value is reasonable. To account for the connectivity between specialties, we will not depend solely on the highest class score in query category filtering; instead, the top 5 classes will be considered.

Table 3: Evaluation results of the classifier.

Metric	Value
Loss	0.8545
Accuracy	0.7407
Precision	0.7380
Recall	0.7404
F1-score	0.7379

4.2 Retrieval system

4.2.1 Retrieval system Test set

Consists of 440 question from the dataset used in the vector database having 22 questions for each category which were rephrased using LLM **Qwen3-32b** to add variability to be able reasonably evaluate the system

4.2.2 Retrieval system results

Table 4: Retrieval system performance

Configuration	Precision@5	Recall@5	MRR	HitRate@5
Full_pipeline	0.068	0.259	0.245	0.259
Without Classifier	0.065	0.255	0.235	0.255
Without Re-ranker	0.060	0.236	0.208	0.236
Basic_retrieval	0.057	0.227	0.188	0.227

The full pipeline results show a significant difference compared to the basic retrieval approach and results difference is almost neglected in the full pipeline compared to the pipeline without the classifier. However, the classifier remains important for query optimization, as it reduces the search space by approximately half.

4.3 LLM

Used llama-3.3-70b-versatile with temperature = 0.2

5 Experimental Results

We compared a baseline naive RAG with our advanced RAG as a whole system, both evaluated on 100 mixed-format test questions.which gave the results show in the table below:

Table 5: BERTScore comparison between naïve and advanced RAG systems.

System	BERTScore
Naïve RAG	0.8287
Advanced RAG	0.8620

6 Conclusion

The proposed approach benefits the large dataset in an advanced RAG system. It reduces the search space for each query, lowering total inference time. It also decreases the chance of retrieving irrelevant data. However, the system still runs sequentially with an LLM, adding extra processing. Additionally, since the dataset is web-scraped, it may require additional pre-processing by an LLM to fix typos and improve text quality before model fine-tuning. The next step is to fine-tune a model using this cleaned dataset and compare its results considering not only the answer quality but the computational power in inference in both approaches.

7 Acknowledgments

We gratefully acknowledge the opportunity provided by Nile Universitys School of Information Technology and Computer Science. Our sincere thanks go to Dr. Ensaf Hussein Mohamed (Supervisor) and Eng. Gannat Ibrahim (Teaching Assistant) for their invaluable guidance and support. This work was carried out as part of our participation in AraHealthQA 2025.

References

Mohammed Abdelhay, Ammar Mohammed, and Hesham Hefny. 2023. Deep learning for arabic healthcare: Medicalbot. *Social Network Analysis and Mining*, 13:71.

Abdelhay and Mohammed. Maqa: Medical arabic q&a dataset. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/Y2JBEZ. Accessed 2025.

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 e-prints*, pages arXiv–2505.

Nashwan Ahmed Al-Majmar, Hezam Gawbah, and Akram Alsubari. 2024. Ahd: Arabic healthcare dataset. *Data in Brief*, 56:110855.

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, and 1 others. 2025. Arahealthqa 2025 shared task description paper. *arXiv preprint arXiv:2508.20047*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, LREC 2020*, Marseille, France.

Jalil Labadi. Altibbi: Arabic online medical forum. https://altibbi.com/. Accessed 2025.

(Abdelhay et al., 2023) (Al-Majmar et al., 2024) (Antoun et al., 2020) (Abdelhay and Mohammed) (Labadi) (Abu Daoud et al., 2025) (Alhuzali et al., 2024) (Alhuzali et al., 2025)

A Code Repository

The GitHub repository for this project is available at: Advancing-Arabic-Medical-QA.

B Software Versions

All experiments were conducted using the following software versions:

Software	Version
Python	3.11.13
sentence-transformers	4.1.0
langchain	0.1.20
chromadb	1.0.17