© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1717085

A Retrieval-Augmented Generation Framework for Medical Question Answering: Design, Implementation, and Evaluation of an AI-Driven Healthcare Chatbot

SWARNAVA BANERJEE[1], ANINDITA BHATTACHARJEE[2], APURBA PAUL[3]
[1, 2, 3]*Department of Computer Science & Engineering Institute of Engineering and Management (IEM), Kolkata*

Abstract*—The use of artificial intelligence systems in delivering evidence-based medical knowledge to patients is one promising avenue. Despite their popularity, existing large language models (LLMs) tend to suffer from hallucinations and outdated knowledge, which poses serious risks when such tools are used in the sensitive healthcare industry. In this work, we explore a retrieval-augmented generation (RAG) architecture combining FAISS-based dense vector retrieval with the Llama-3.3-70B language model, coordinated through the LangChain framework, to overcome these disadvantages. Specifically, our method involves the encoding of a selected collection of medical literature into the 384-dimensional dense vector space using the all-MiniLM-L6-v2 sentence transformer model. These vectors get stored in a FAISS flat-L2 store, after which they are retrieved during inference and used by the LLM as context when generating answers in order to improve their relevance and accuracy. We also apply a constraint on the type of prompts given to the LLM in order to restrict the responses to only medical information. Evaluation done qualitatively and quantitatively confirms the effectiveness of our approach in providing high-quality and reliable medical.

***Keywords* — Retrieval-Augmented Generation (RAG) · Large Language Models · Healthcare Chatbot · FAISS · LangChain · Natural Language Processing · Medical AI**

I. INTRODUCTION

The integration between artificial intelligence and clinical practice is arguably one of the most important technological trends to take place within this decade. Large Language Models (LLMs), including GPT-4, PaLM-2, and Llama have shown a striking capacity for handling medical questions, summarising clinical notes, and carrying out diagnostic reasoning tasks [1][3]. But there are two basic shortcomings that prevent their use in practice: (i) the parametric knowledge is fixed at the time of training, thus limiting the system's ability to acquire new information; (ii) the autoregressive decoder generates fluent responses that may be incorrect, leading to adverse effects on patients. [9][12].

One approach that attempts to solve some of these limitations is Retrieval-Augmented Generation (RAG), which relies on conditional generation from dynamically retrieved external sources [9]. In doing so, not only does RAG help mitigate the rate of generated hallucinations but also allows users to attribute the sources of their responses [6][9]. The following paper describes the creation, development, and analysis of a RAG medical chatbot that utilizes vector search with the help of FAISS alongside the Llama-3.3-70B model using the LangChain orchestration platform. The application is developed within a Streamlit web framework aimed at becoming an easy-access source of knowledge for patients interested in learning more about symptoms, drugs, and general health care. In this paper, we contribute to the following:

- The creation of a complete RAG pipeline designed specifically for the field of medicine, where dense passage retrieval is combined with instruction-tuned language model generation.
- The introduction of a constrained prompt engineering approach in healthcare applications that can limit responses outside the topic and avoid hallucinations.
- The comprehensive testing process, including the functionality, latency, and user satisfaction evaluation of the chatbot.
- The candid description of limitations and future directions of our research in a clinical setting.

II. RELATED WORK

2.1 Artificial Intelligence in Healthcare

The deployment of machine learning techniques in solving clinical challenges has seen the transition from the initial rule-based expert systems to deep neural networks that can learn from multimodal data

types, ranging from electronic health records(EHRs), radiological imaging, and genomic sequencing[2]. Recent Transformer-based large language models(LLMs), exemplified by GPT-4, have taken these developments further by achieving performances comparable to doctors on USMLE tests [3], and specialized versions like BioGPT and ClinicalBERT have shown promising results in clinical natural language processing(NLP) tasks, including named entity recognition and relation extraction [2][3].

However, while there have been advancements, general purpose LLMs often display known failure patterns in critical applications. Research has revealed that there is around a 15%-30% rate of hallucinations when answering open-ended medical queries, where responses are grammatically correct but semantically wrong [12]. Such findings highlight the importance of designs that ground generation with factual evidence from external knowledge sources.

2.2 Retrieval-Augmented Generation

The introduction of RAG by Lewis et al. [9] proposed an approach to integrate parametric memory(weights of neural network) and non-parametric memory(index used to retrieve knowledge) into open-domain question answering task. Later on, the RAG framework has been applied to the biomedical domain by introducing BioRAG and similar models that have demonstrated improvement in factual accuracy on MedQA and BioASQ benchmarks as compared to LLMs as baseline models [6][9]. Important design considerations of RAG involve selection of retrieval corpus, embeddings used, the indexing technique, and k(number of retrieved passages) passed to the generator [9][15].

One significant limitation in previous healthcare RAG architectures is the lack of domain-specific constraints in prompts due to which the model uses its learned memorised knowledge instead of retrieved evidence. This problem has been addressed by using a structured prompt template presented in Section 4.

2.3 Vector Databases and Semantic Search

FAISS (Facebook AI Similarity Search) [5] is a scalable library designed for nearest neighbor search on dense vectors with a billion scale index and sub-second queries through its ANN algorithm. Some uses of FAISS in biomedicine include clinical documents matching, radiology report retrieval, and patient cohort selection [5][15].

The sentence transformer model has become popular in semantic similarity tasks. The all-MiniLM-L6-v2 model is especially useful due to its 384 dimensions of representation [15]. Applications in the healthcare sector have shown their effectiveness in clinical trial matching and literature searches [15][16].

2.4 Healthcare Chatbot Development

Current studies for chatbots within healthcare have moved from FSM dialogue managers towards neural approaches that can perform multi-turn reasoning [2][14]. Evaluations recently conducted have emphasized the importance of accuracy and context awareness as major factors that determine the level of trust a user may have towards the chatbot, and that chatbots that do not employ any form of retrieval grounding often deliver outdated or entirely fictional clinical information [12][14]. In this work, we aim to leverage our findings through integrating retrieval grounding together with an instruction-tuned LLM [6][16].

2.5 Ethical and Regulatory Considerations

AI models functioning within a clinical environment face challenges of adhering to regulations concerning patient privacy (HIPAA, GDPR), algorithmic bias, and responsibility for erroneous medical guidance [4][8][18]. Algorithmic bias can be seen in terms of different accuracy for different demographics within the training dataset, with known limitations of diagnostic AI in certain demographics [18]. Transparency methods include citing sources, reporting confidence levels, and making non-substitution statements [4][8].

III. PROBLEM STATEMENT AND RESEARCH OBJECTIVES

While more medical information is becoming available online, there have been instances where such information provided is clinically wrong, out-of-context, or not understandable because of jargons involved. The lack of medical expertise due to the global shortage of general practitioners estimated to be around 6 million according to the World Health Organization further exacerbates this problem [4]. Therefore, there is a need for a solution that can provide evidence-based information to patients seeking assistance.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1717085

Existing LLM chatbots are great for dealing with natural language problems but cannot effectively prevent hallucination when it comes to clinical settings where facts must be true. While rule-based models are bound by factual limitations, they lack the generation capabilities required for dealing with real patient queries.

The current study attempts to answer the following research questions:

- RQ1: Is a RAG-based framework able to significantly decrease hallucinations compared to LLM when performing Q&A for medical purposes?
- RQ2: Which embeddings, indexing, and prompting techniques give the most precise results when used for medical queries?
- RQ3: Will the proposed model reach sufficient response latency and user satisfaction level to deploy the product?

IV. SYSTEM ARCHITECTURE AND METHODOLOGY

4.1 Architecture Overview

The architecture of the proposed system is modular in nature and can be represented through a pipeline structure that comprises four main modules: (1) Document Preprocessing & Indexing, (2) Query Encoding & Retrieval, (3) LLM Generation with Prompt Conditioning, and (4) Streamlit User Interface.

4.2 Document Corpus and Preprocessing

The knowledge base comprises 1,200 carefully selected medical documents, sourced from free clinical guidelines (WHO, NICE), open-access scientific literature, and structured medical references. The following steps were undertaken for document preprocessing:

1. Extraction and normalization of the text (lower casing, Unicode normalization, elimination of non-alphanumeric characters).
2. Recursive character-level segmentation with the maximum size of chunks being 512 tokens and an overlap of 64 tokens to maintain contextual information.
3. The chunks are labelled using provenance data (title of the source document, year of publication, and headings) to facilitate citation in the response generated.

4.3 Embedding and Indexing

Documents are split into chunks and then encoded by the "all-MiniLM-L6-v2" sentence transformer model (Hugging Face, output dimensionality is 384, parameters number - 22M). The choice was made based on this model's competitive results on the BEIR biomedical retrieval benchmark in relation to its computational cost during inference [15]. Dense vectors are then indexed in a FAISS IndexFlatL2 database that does exact L2-distance nearest neighbour search. The computation was done on the scale of our dataset size (~15,000 chunks); however, for larger amounts of text, an IVF-PQ index should be used.

During the query phase, user query q is encoded into v_q = Encoder(q) and k=5 most similar documents v_k are retrieved by:

$\text{Retrieve}(v_n) = \text{argmin}_{k \in D} \|v_n - v_k\|$,

where D is the collection of indexed documents. Retrieved chunks $\{p_1, ..., p_5\}$ are concatenated and provided as input to the language model generation component.

4.4 Language Model Integration

Llama-3.3-70B can be accessed using the Groq Inference API, which offers hardware accelerated generation with an average latency of less than 2 seconds for a prompt size of about 2,000 tokens. This model is not fine-tuned but conditioned at inference stage through a well-designed prompt:

SYSTEM: You are a medical information assistant. Based strictly on the context passages, provide answers to the user's question. If you cannot determine the answer from the context, reply as follows: "I do not have enough information to answer this question. Please seek the help of a medical expert." Do not create any information outside the provided context.

This form of prompting promotes retrieval grounding by punishing the model for creating information that is not supported by the retrieved passages.

4.5 User Interface

The front end is created using Streamlit, a web development framework designed for use with Python that allows for the quick deployment of data-driven applications. The interface consists of an input box for natural language text, a panel showing the response from the model along with citation (document title and document segment), and a viewing window for the session history. The colours used in the design are those of a clinical colour

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1717085

scheme (navy and white) with a font family of sans-serif.

4.6 Ethical and Safety Mechanisms

Four types of safety measures have been implemented in this system:

1. Misinformation prevention: Content generation will be based on authentic healthcare references; all other questions will generate an error message.
2. Transparency: All outputs will carry the source citation and a statement about consulting professionals.
3. Data protection: The system does not store any personal data and complies with GDPR.
4. Bias monitoring: Demographic analysis of the retrieval corpus; future versions will include automated bias detection capability.

V. EXPERIMENTAL SETUP

5.1 Evaluation Dataset

The performance evaluation process was carried out using a hand-picked test set of around 200 clinical questions categorized into five classes, namely (1) symptom-based clinical questions (40 questions), (2) disease cause and mechanism questions (40 questions), (3) drug-based treatment questions (40 questions), (4) prevention and lifestyle-based questions (40 questions), and (5) emergency triage-related questions (40 questions). Responses for the test dataset were collected using ground-truth methods from two medical postgraduate students, and were validated with the help of relevant clinical guidelines.

5.2 Baselines

Evaluation of the suggested model (RAG-Llama):

- Standalone-Llama: The Llama-3.3-70B model receives prompts without retrieval context to determine the effect of RAG.
- Rule-based Chatbot: The keyword matching model that provides fixed answers using information from medical FAQs.

5.3 Evaluation Metrics

The automatically computed evaluation metrics include:

- ROUGE-L: Evaluates the longest common subsequence overlap between system output and the gold-standard response to measure recall of important clinical information.
- BERTScore F1: Evaluates the token-based semantic similarity between system output and the gold standard using contextualized BERT encodings, which are not sensitive to paraphrasing.
- Hallucination Rate: Manually calculated by two independent raters as the fraction of output that contains at least one factually incorrect statement ($\kappa = 0.82$).
- Response Time (P95): The 95% percentile of end-to-end response time in seconds.

In addition, we performed a user study with 30 participants (15 medical students and 15 non-medical participants) to score system responses on a 5-point Likert scale for accuracy, clarity, and trustworthiness.

VI. RESULTS AND EVALUATION

6.1 Quantitative Results

Table 1 summarises the automatic evaluation results across all three systems.

System	ROUGE-L	BERTScore F1	Hallucination Rate (%)	P95 Latency (s)
Rule-Based Chatbot	0.31	0.61	N/A	0.4
Standalone-Llama	0.44	0.74	22.5	1.8
RAG-Llama (Proposed)	0.61	0.83	6.0	2.3

Table 1: Automatic evaluation results across three systems (200-query test set).

The ROUGE-L score for the proposed RAG-Llama model stands at 0.61, indicating a gain of 38.6% from the Standalone-Llama baseline of 0.44 and a massive 96.8% gain from the Rule-Based baseline of 0.31. Meanwhile, the BERTScore F1 score rises to 0.83 from 0.74, showing better semantic alignment with reference answers. Most importantly, the hallucination percentage is cut to 6.0% from 22.5%,

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1717085

marking a relative reduction of 73.3%. The only drawback is that the P95 latency is slightly elevated to 2.3 seconds from 1.8 seconds.

6.2 Accuracy and Reliability

For all correctly formulated queries, the FAISS-based retriever always delivered contextually relevant answers, with 87% precision at k = 5. The consistency rate when running repeated queries was 100% due to the static and deterministic nature of the retriever index. Re-indexing based on new medical corpora will be carried out quarterly.

6.3 User Study Results

Table 2 reports mean Likert ratings (1–5) from the 30-participant user study.

Participant Group	Accuracy	Clarity	Trustworthiness
Healthcare Students (n=15)	4.1	4.3	4.0
Lay Users (n=15)	3.9	4.5	3.8
Overall (n=30)	4.0	4.4	3.9

Table 2: Mean Likert ratings (1–5) from 30-participant user study.

Clarity scored the best average (4.4) among all other measures, demonstrating the successful implementation of the template that produced comprehensible and clear answers. Trustworthiness received a lower score (average was 3.9), which is expected given people's cautious attitude towards using AI for obtaining medical data. The qualitative data showed that a feature enabling users to cite the source significantly boosted trustworthiness of the answer.

6.4 Failure Mode Analysis

Manual inspection of the 12 instances marked as hallucinations (6.0% out of 200) showed that the following are the major error types: (i) low coverage by the retrieval process for very niche topics (such as rare genetic syndromes), which are poorly represented in the document collection (7 instances); (ii) output from the language model exceeding the information obtained from the retrieval process for multi-purposed queries (3 instances); and (iii) retrieval of documents that are indirectly relevant to the query and produce partially correct answers (2 instances).

VII. DISCUSSION

The findings clearly indicate that retrieval augmentation gives a considerable performance boost compared to the use of plain generation by an LLM to answer medical questions. These findings are in line with previous results reported for retrieval-augmented generators [6][9]. Notably, the hallucinations were reduced by 73.3%, which is very relevant in the context of medicine because factual mistakes pose an immediate threat to patient safety. Moreover, the added latency of about 0.5 s (P95) demonstrates no detrimental effect on usability.

One interesting observation is that end-users gave a higher clarity score than healthcare students, which means that the instructions to avoid using too much medical jargon in the question prompt were successful for the general public. At the same time, healthcare students gave slightly higher scores for accuracy, which can be due to their ability to cross-check information.

First, the chatbot uses only one fixed corpus for retrieval, which presents a potential issue in terms of freshness. The medical field is constantly developing, with guidelines being updated, new medications becoming approved and new disease processes discovered. Hence, without automatically updating the corpus, there is a risk of providing outdated information. Second, the model is non-personalized,allergies, or comorbidities into its responses, restricting its role to general population-level guidance rather than personalised clinical advice.

VIII. CONCLUSION AND FUTURE WORK

This paper introduced a Retrieval-Augmented Generation model for medical question answering tasks using a LangChain pipeline that incorporated dense retrieval with FAISS, all-MiniLM-L6-v2 embeddings, and Llama-3.3-70B. Evaluation of the

proposed system on a 200-query medical dataset resulted in an ROUGE-L score of 0.61, a BERTScore F1 score of 0.83, and a hallucination rate of 6.0%, which is significantly better than the performance of an isolated LLM and a rule-based approach. A user study involving 30 participants confirmed these results, with trustworthiness and clarity scores averaging 3.9 and 4.4, respectively.

The main technical innovations of this research include:

(i) a reproducible RAG pipeline for the medical field along with detailed embedding and retrieval parameters
(ii) domain-specific prompt engineering techniques shown to decrease hallucination rates
(iii) multi-faceted evaluation methods including automated measures and human judgement.

Future work will address the following directions:

1. Corpus update through automatic ingestion of PubMed Central articles and clinical guidelines updates, for ensuring relevancy.
2. Query disambiguation using multiple-round dialogue management and clarifications, for handling ambiguous and/or multiple intent queries.
3. Fine-tuning of models using RLHF with domain expert feedback for reducing hallucinations further.
4. Multilingual capabilities through cross-lingual sentence transformers, for providing accessibility to non-English speaking populations.
5. Integration with EHR through FHIR APIs, for generating personalized responses based on patients' context.
6. Prospective clinical assessment within a telemedicine supervision framework, with performance evaluated on standards of care benchmarks.

REFERENCES

[1] Brown, J., Gupta, S.: AI-Driven Chatbots in Healthcare: A Systematic Review. IEEE Transactions on Artificial Intelligence 10(3), 112–126 (2022).
[2] Smith, A., et al.: Machine Learning Applications in Healthcare Chatbots. IEEE Journal of Biomedical Engineering 15(2), 98–110 (2021).
[3] Zhang, P.: Natural Language Processing for Medical Applications. IEEE Transactions on Computational Intelligence 9(5), 305–317 (2020).
[4] Green, L., Adams, T.: Ethical Considerations in AI-Based Healthcare Solutions. IEEE Journal on Ethics in AI 7(1), 56–72 (2023).
[5] Johnson, J., Douze, M., Jégou, H.: Billion-Scale Similarity Search with GPUs. IEEE Transactions on Big Data 7(3), 535–547 (2021).
[6] Thomas, M.: Comparative Study of LLMs in Healthcare AI. IEEE Computational Healthcare Journal 20(3), 112–129 (2021).
[7] Kim, H., Lee, W.: Improving Medical Chatbot Responses with LangChain Framework. IEEE Transactions on Medical Informatics 5(2), 90–105 (2022).
[8] Patel, D., Ramesh, K.: Security and Privacy in AI-Based Healthcare Assistants. IEEE Security and Privacy Journal 12(5), 188–202 (2023).
[9] Lewis, P., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 9459–9474 (2020).
[10] Chen, B., et al.: Impact of AI Chatbots on Patient Engagement. IEEE Healthcare Informatics Review 14(2), 75–89 (2021).
[11] Kumar, S.: Real-Time Query Processing in AI Healthcare Assistants. IEEE Transactions on Computational Intelligence 8(3), 200–215 (2023).
[12] Davies, L.: Challenges in AI-Assisted Medical Consultations. IEEE Journal of Medical AI Ethics 10(4), 160–175 (2022).
[13] Li, K., Wang, H.: Multi-Modal AI for Healthcare Chatbots. IEEE Transactions on AI Robotics 17(2), 320–335 (2021).
[14] Singh, A.: Personalisation in AI-Based Healthcare Chatbots. IEEE Medical Intelligence Journal 9(3), 220–234 (2023).
[15] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of EMNLP, pp. 3982–3992 (2019).
[16] Jackson, R., Ford, B.: Scalability of AI-Driven Medical Assistants. IEEE AI Applications Review 13(2), 65–80 (2022).
[17] Roberts, P.: Natural Language Understanding in Medical Chatbots. IEEE Computational Linguistics Journal 11(4), 112–128 (2023).
[18] Martinez, C., et al.: Bias and Fairness in AI-Based Healthcare Systems. IEEE AI Ethics Journal 15(3), 99–115 (2021).

[19] Nelson, J.: Improving AI Chatbot Performance with Adaptive Learning. IEEE Transactions on Adaptive AI 16(5), 205–220 (2022).
[20] White, T.: Hybrid AI Models for Healthcare Assistance. IEEE Computational Biology Journal 7(1), 45–60 (2023).