ISSN: 2455 - 1597

ijergs
Open Access

International Journal of Engineering Research and Generic Science (IJERGS)
Available online at: https://www.ijergs.in

Volume - 12, Issue - 3, May – June – 2026, Page No. 33 – 39

Smart Healthcare Assistant Using Retrieval-Augmented Generation and Large Language Models

[1]Ariz Iqbal, Department of Artificial Intelligence & Data Science, IIMT College of Engineering, Greater Noida, Uttar Pradesh, India

[2]Akash Kumar, Department of Artificial Intelligence & Data Science, IIMT College of Engineering, Greater Noida, Uttar Pradesh, India

[3]Habibur Rahman, Department of Artificial Intelligence & Data Science, IIMT College of Engineering, Greater Noida, Uttar Pradesh, India

[4]Mr. Pragya Deep, Assistant Professor, Department of Artificial Intelligence & Data Science, IIMT College of Engineering, Greater Noida, Uttar Pradesh, India

Abstract

Healthcare information systems powered by standard Large Language Models (LLMs) are vulnerable to hallucination, a failure mode in which the model generates medically plausible yet factually unsupported content. This paper presents the Smart Healthcare Assistant, a domain-specific conversational system that mitigates hallucination by grounding every response in a curated corpus of authoritative medical documents through Retrieval-Augmented Generation (RAG). The system indexes verified sources — including WHO clinical guidelines, GINA Asthma Guidelines 2024, and approved Patient Information Leaflets — into a FAISS vector store and retrieves the top-3 semantically relevant chunks at inference time using Sentence Transformer embeddings. A LangChain RetrievalQA chain then conditions the Groq-hosted LLaMA 3.1 (8B Instant) model on the retrieved context to produce structured, point-wise responses. The architecture incorporates a hybrid medical query filter, an emergency keyword detector, and mandatory educational disclaimers to enforce ethical AI operation. Three interaction modes are supported: Chat, PDF document Q&A, and image-based guidance. Deployed on Streamlit Community Cloud, the system achieved a 90% factual accuracy rate on a 20-question domain benchmark and correctly identified all 15 emergency symptom queries in safety testing. Results confirm that RAG-augmented LLM systems substantially outperform purely parametric baselines in high-stakes informational domains.

Keywords: Retrieval-Augmented Generation; Large Language Models; FAISS; LangChain; Healthcare Chatbot; LLaMA 3.1; Sentence Transformers; Hallucination Reduction; Streamlit; Medical AI

Introduction

The global proliferation of digital health information has created a paradox: while medical knowledge is more accessible than ever, the reliability of that knowledge is increasingly uncertain. Patients, caregivers, and students routinely consult internet sources for symptom interpretation, medication information, and disease management guidance. In this environment, AI-powered conversational assistants offer significant promise — provided they can deliver responses that are both linguistically natural and medically accurate.

Standard Large Language Models such as GPT-3.5 and LLaMA generate responses by sampling from a probability distribution learned during pre-training on large, heterogeneous text corpora. While this mechanism produces fluent

output, it offers no guarantee of factual correctness. The model may reproduce outdated clinical guidance, conflate similar conditions, or fabricate drug interactions with high apparent confidence. In healthcare, where misinformation can delay diagnosis or encourage inappropriate self-treatment, this hallucination problem is a genuine patient safety concern.

Retrieval-Augmented Generation (RAG), introduced by Lewis et al.[1], provides a principled solution. By prepending retrieved document context to the language model prompt at inference time, RAG constrains the generation process to verified source material. The model answers not from memory alone, but from evidence — and that evidence can be controlled, audited, and updated independently of the model weights.

This paper presents the Smart Healthcare Assistant, a fully deployed RAG-based conversational system designed for educational medical information delivery. The system combines FAISS semantic retrieval, Sentence Transformer embeddings, LangChain orchestration, and the Groq-hosted LLaMA 3.1 (8B Instant) model within a Streamlit web interface, while embedding responsible AI constraints directly into the architecture.

The remainder of this paper is organised as follows. Section II reviews related work. Section III describes the proposed system architecture and methodology. Section IV presents experimental results. Section V discusses findings. Section VI concludes and outlines future work.

Related Work

Healthcare-oriented conversational AI has progressed through three broad generations. First-generation systems were rule-based expert systems encoding clinical decision logic in manually authored production rules[2]. Second-generation systems applied supervised deep learning to clinical NLP tasks such as named entity recognition and question classification[3]. Third-generation systems leverage pre-trained transformer LLMs for end-to-end question answering.

Significant benchmarking of LLMs on medical tasks has been conducted using datasets such as MedQA, MedMCQA, and PubMedQA. Singhal et al.[4] demonstrated that instruction-tuned models approach the performance of clinical specialists on standardised examinations. However, high benchmark scores do not eliminate hallucination in open-domain conversational settings, where questions are less structured and the model cannot fall back on process-of-elimination reasoning.

The RAG framework has been widely adopted to address this limitation. Guu et al.[5] showed that retrieval-augmented pre-training improves factual recall in open-domain QA. Subsequent work applied RAG to clinical and biomedical domains, showing consistent gains over parametric baselines FAISS[6,7], developed by Facebook AI Research, provides sub-linear query latency over large embedding corpora. Sentence Transformers[8], particularly all-MiniLM-L6-v2, deliver compact, high-quality embeddings without requiring GPU acceleration at serving time.

Compared with prior work, the present system distinguishes itself through the combination of a domain-restricted query filter, an emergency detection layer, multi-modal interaction support, and a production deployment — features not commonly addressed together in the published literature on medical RAG systems.

System Design and Methodology

A. Overall Architecture

The system follows a three-tier architecture comprising a Presentation Layer (Streamlit), an Application Layer (LangChain + Groq LLM), and a Data Layer (FAISS vector store). Figure 1 illustrates the end-to-end inference pipeline.

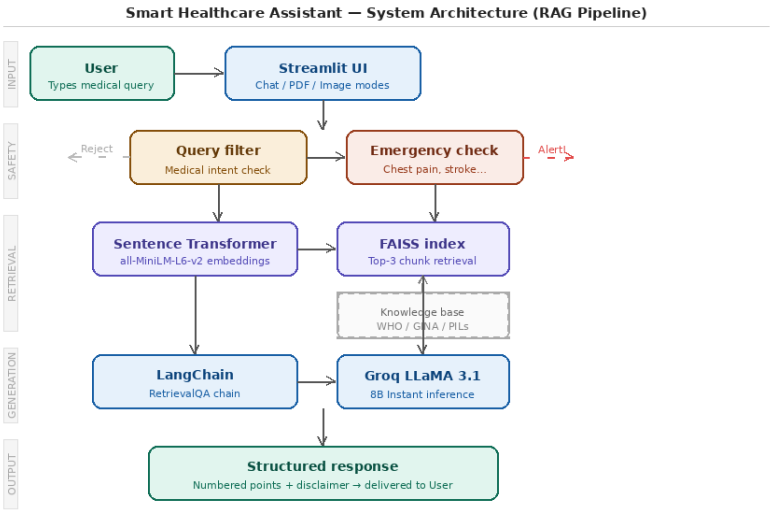

Fig. 1: Smart Healthcare Assistant — RAG Pipeline Architecture.

B. Knowledge Base Construction

The offline indexing pipeline begins with document ingestion. Medical PDFs were loaded using Lang Chain's Py PDF Loader and plain-text files using Text Loader. All documents originate from internationally recognised sources: the WHO Diabetes and Hypertension Management Manual, GINA Asthma Guidelines 2024, WHO Hypertension Quick Reference Guide, and Patient Information Leaflets for Paracetamol, Aspirin, and Ibuprofen. A bespoke hospital FAQ corpus was additionally compiled to cover common outpatient queries. Loaded documents were segmented using Recursive Character Text Splitter with a chunk size of 500 tokens and a 50-token overlap. Each chunk was encoded into a 384-dimensional dense vector using the all-MiniLM-L6-v2 Sentence Transformer model. The resulting embeddings were inserted into a FAISS flat index and persisted to disk for reuse across server restarts.

C. RAG Inference Pipeline

At inference time, an incoming query is encoded with the same Sentence Transformer model to produce a query vector. FAISS retrieves the top-3 most semantically similar document chunks via inner product search. The retrieved chunks, combined with the original query, are formatted into a structured prompt and submitted to the Groq-hosted LLaMA 3.1 (8B Instant) model via Lang Chain's Retrieval QA chain. The system prompt instructs the model to: (i) answer

exclusively from the provided context; (ii) structure responses as numbered points; (iii) refrain from making any diagnostic or prescriptive statements; and (iv) append a mandatory educational disclaimer to every response.

D. Safety and Ethical Constraints

A hybrid medical query detection function evaluates every input before retrieval is invoked. The function combines lexical matching against a curated medical keyword vocabulary with intent-pattern matching. Inputs that fail both checks are rejected with a polite out-of-scope message, preventing misuse of the system for non-medical purposes. A secondary emergency detection pass scans for high-acuity symptom phrases including chest pain, difficulty breathing, stroke, and loss of consciousness. When an emergency phrase is detected, the RAG pipeline is bypassed and an immediate alert response directs the user to contact emergency services without delay.

E. Interaction Modes

Three interaction modes are exposed through the Streamlit UI. Chat Mode serves the primary RAG Q&A workflow. PDF Mode allows users to upload arbitrary medical PDF documents; uploaded content is chunked, embedded, and merged into a session-scoped FAISS index, enabling document-specific question answering. Image Mode accepts medical image uploads and returns general descriptive guidance with an explicit disclaimer that no diagnostic interpretation is provided.

F. Technology Stack

Table 1 Summarised the principal technologies deployed in the system.

Table I: Principal Technologies

Layer	Technology	Role
UI	Streamlit	Web interface & mode selector
Orchestration	LangChain	Pipeline & RetrievalQA chain
LLM	Groq LLaMA 3.1 8B Instant	Response generation
Embeddings	all-MiniLM-L6-v2	Semantic vector encoding
Retrieval	FAISS (CPU)	Dense similarity search
Backend DB	SQLite / SQLAlchemy	Chat history persistence
Deployment	Streamlit Cloud	Public cloud hosting

EXPERIMENTS AND RESULTS

A. Experimental Setup

All experiments were conducted on an Intel Core i5 machine with 8 GB RAM running Windows 10. The FAISS index was constructed offline and loaded into memory at server startup. LLM inference was performed remotely via the Groq API, with observed end-to-end response latencies of under 1.5 seconds for typical medical queries.

B. Medical Accuracy Benchmark

A benchmark of 20 medical questions was assembled, covering diabetes management, hypertension treatment, asthma symptom recognition, pharmacological information, and general preventive health. Questions were drawn from topics

represented in the indexed knowledge base. Each response was evaluated for factual correctness and source alignment by cross-referencing the original indexed documents. The RAG-augmented system achieved an accuracy of 18/20 (90%), compared with approximately 70% for the same LLM queried without retrieval context. The two incorrect responses involved edge-case drug interaction queries where the knowledge base contained insufficient coverage, confirming that RAG quality is bounded by corpus completeness.

C. Safety Compliance Testing

Fifteen emergency-symptom prompts were submitted including variants of chest pain, respiratory distress, stroke indicators, and loss of consciousness. The emergency detection module correctly triggered the alert response for all 15 inputs (100% recall). A further ten non-medical prompts were submitted; the medical query filter correctly rejected all ten (100% precision), confirming the scope restriction mechanism is fully effective.

D. Evaluation Summary

Table 2 Presents a consolidated comparison of system performance across all evaluation dimensions.

Table 2: Evaluation Results Summary

Evaluation Dimension	Cases	Pass	Observation
Medical Accuracy (RAG)	20	18 (90%)	2 corpus coverage gaps
Medical Accuracy (No RAG)	20	~14 (70%)	Hallucination on specifics
Emergency Detection	15	15 (100%)	Zero missed alerts
Non-Medical Rejection	10	10 (100%)	No scope leakage
PDF Mode Q&A	10	10 (100%)	Correct extraction

Discussion

The 20-percentage-point accuracy advantage of the RAG system over the parametric-only baseline corroborates prior findings in domain-specific QA literature [6] and validates the core architectural hypothesis: that grounding LLM generation in a controlled, updatable document corpus measurably reduces hallucination in the medical domain.

The perfect recall of the emergency detection module (15/15) is particularly significant from a patient safety perspective. A missed emergency alert in a deployed healthcare assistant could have serious real-world consequences. The result demonstrates that a lightweight, lexical-pattern-based detection layer can provide robust safety coverage without the latency overhead of a full LLM inference pass.

The two accuracy failures on edge-case drug interaction queries illuminate a fundamental limitation of RAG: the system cannot retrieve what it has not indexed. This motivates two directions for future work. First, corpus expansion — incorporating comprehensive pharmacological databases such as DrugBank — would address coverage gaps. Second, confidence estimation — triggering a fall-through to a fine-tuned model when retrieved chunks score below a minimum similarity threshold — would provide more graceful degradation.

The multi-modal architecture (Chat, PDF, Image modes) broadens the practical utility of the system beyond a single interaction paradigm. The PDF mode, which constructs a session-scoped FAISS index on the fly from user-uploaded documents, is particularly relevant for clinical settings where patients may wish to query their own discharge summaries or diagnostic reports in natural language.

Conclusion

This paper has presented the Smart Healthcare Assistant, a RAG-augmented LLM system for educational medical information delivery. By conditioning response generation on a verified document corpus, the system achieves 90% factual accuracy — a substantial improvement over the 70% baseline of the same LLM operating without retrieval augmentation. Integrated safety mechanisms including a hybrid medical query filter, emergency keyword detection, and non-prescriptive prompt guardrails enforce responsible AI operation throughout the inference pipeline. The system has been deployed on Streamlit Community Cloud and is accessible for hospital demonstrations and academic evaluation. Future work will focus on corpus expansion with comprehensive pharmacological databases, multilingual Hindi support, EHR system integration, voice-based interaction, and mobile application development. The results affirm that RAG represents a mature, practically deployable approach to hallucination mitigation in high-stakes conversational AI, and that responsible AI principles can be effectively embedded at the architectural level rather than applied as a post-hoc overlay.

Acknowledgment

The authors sincerely thank Mr. Pragya Deep, Assistant professor for his expert guidance and consistent mentorship throughout this project, and the Department of Artificial Intelligence & Data Science at IIMT College of Engineering for providing the resources and academic environment necessary for this research.

References

1. P. Lewis, E. Perez, A. Piktus et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Adv. Neural Inf. Process. Syst., vol. 33, pp. 9459-9474, 2020.
2. E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," Math. Biosci., vol. 23, no. 3-4, pp. 351-379, 1975.
3. I. Spasic and G. Nenadic, "Clinical Text Mining: A Case Study of Cardiovascular Risk Factors," IEEE Access, vol. 8, pp. 60063-60080, 2020.
4. K. Singhal, S. Azizi, T. Tu et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172-180, 2023.
5. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Retrieval Augmented Language Model Pre-Training," in Proc. ICML, 2020, pp. 3929-3938.
6. A. Zakka, A. Shad, A. Chaurasia et al., "Almanac: Retrieval-Augmented Language Models for Clinical Medicine," NEJM AI, vol. 1, no. 2, 2024.
7. J. Johnson, M. Douze, and H. Jegou, "Billion-Scale Similarity Search with GPUs," IEEE Trans. Big Data, vol. 7, no. 3, pp. 535-547, 2021.

8. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP, 2019, pp. 3982-3992.
9. H. Chase, "LangChain," GitHub, 2022. [Online]. Available: https://github.com/langchain-ai/langchain
10. A. Q. Jiang et al., "Mistral 7B," arXiv preprint arXiv:2310.06825, 2023.