Intelligent RAG-Based Conversational Assistant for Medical and Healthcare Awareness

Dr C V Madhusudhan Reddy[1], M Sivamma[2], Patil Surya Narayana Reddy[3], Chapala Narasimhulu[4], Syed Md Sammer[5], Kanchi Gokul Raj[6]

[1,2,3,4,5,6]*Dept. Of Computer Science and Engineering (Artificial Intelligence), St. Johns College of Engineering and Technology, Yemmiganur, 518301, India.*

***Abstract*—The increasing reliance on digital platforms for medical information has highlighted the need for accurate, reliable, and context-aware healthcare advisory systems. General AI chatbots and unverified online sources often provide misleading or incomplete information that can negatively impact healthcare awareness and education. The objective of this project is to develop an Intelligent Retrieval-Augmented Generation (RAG) based medical conversational assistant that delivers trustworthy and educational medical information while avoiding diagnostic or prescriptive outputs. The proposed system integrates document retrieval techniques with large language models to ensure that the generated responses are grounded in verified medical documents. Medical knowledge sources are pre-processed, embedded using sentence-transformer models, and stored in a FAISS vector database for efficient similarity-based retrieval. User queries are matched with relevant medical contexts, which are then supplied to the language model through a controlled RAG pipeline orchestrated using Lang Chain.**

Experimental evaluation shows that the RAG-based approach significantly improves response accuracy, contextual relevance, and reliability compared with standalone language models, while reducing hallucinations and misinformation. The system is well suited for medical students, healthcare awareness platforms, and hospital information desks. Overall, this project demonstrates the effective application of AI and RAG techniques in building safe, scalable, and reliable medical information systems.

***Index Terms*—Health Information System, Artificial Intelligence, Natural Language Processing, Retrieval-Augmented Generation, Machine Learning for Health Care, Intelligent Information Retrieval, Semantic Search.**

I. INTRODUCTION

There has been a paradigm shift in the internet health care system in that most patients have started using the internet platform for the very first time in seeking health care. Even if convenience is an aspect to be considered in the future for valuation of the platform, the infodemic of the wrong health care stories and consumption of the wrong knowledge are definitely circumstances of concern. The conventional internet search platforms, without the aid of words in health care concerns to be explained, lack the capability to understand the complexity involved in internet health care.

While there has been recent progress in the application of NLP to facilitate more human-like interactions, confabulation issues seem not to be acceptable when applied to medical practice with LLMs alone. In view of such shortcomings, the use of the Retrieval-Augmented Generation model has been identified as a better architectural style. In fact, this proposed research work will capitalize on the benefits to create a medical companion that is designed with the highest priority for accuracy and the safety of the user and that is not programmed to perform any form of diagnosis.

II. METHODOLOGY

A. CURATED KNOWLEDGE INGESTION

To build a trustful base, the knowledge base for the system is curated from the most accurate medical encyclopedias and educational sites, and internet scraping has been avoided in this approach for better authentication purposes. The dataset has been created for general health domains, which include pathology,

anatomy, and symptomatology, and is curated in a way that is non-patient-specific and free.

B. SEMANTIC SEGMENTATION AND VECTORIZATION

PDF format files with medical documents undergo a massive preprocessing process. A recursive text segmentation technique was utilized, where the text was split into segments with a size of 500 characters and an overlap of 50 characters. This specific configuration was used to preserve the semantic meaning of medical definitions. The medical text is segmented into smaller text, which is then utilized to generate dense vector space representations using a sentence transformer.

C. VECTOR STORAGE AND RETRIEVAL MECHANISM

The resulting embeddings were stored as indices in a Facebook AI Similarity Search (FAISS) vector database owing to its efficiency when working with high-dimensional similarity searches. When a query from the user is submitted to the system, the input is embedded as an appropriate vector to obtain the embedding. The retriever is designed to select the top three (k=3) most appropriate document chunks according to the cosine similarity score. This must be balanced by considering the context limitations of the number of tokens within the language model.

D. ORCHESTRATION VIA LCEL

The interaction between retrieval and generation is controlled by a pipeline based on Lang Chain Expression Language LCEL. This declarative framework chains together the query embedding, document retrieval, and prompt building. A very important part of this step is the injection of a custom prompt template that wraps the retrieved medical facts with strict safety instructions, explicitly prohibiting the model from providing medical diagnosis or medication.

E. FIGURE

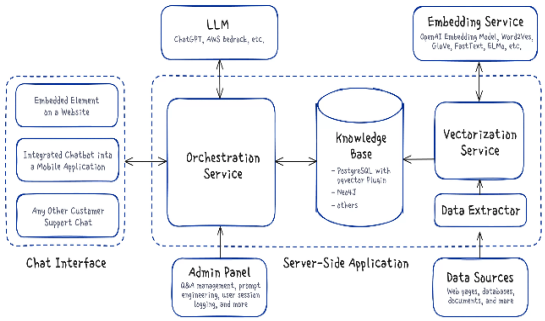

FIG. 1:
Architecture of the proposed Retrieval-Augmented Generation (RAG) based medical conversational assistant showing data ingestion, vector-based retrieval using FAISS, LCEL-based orchestration and response generation through a large language model.

III. SYSTEM ARCHITECTURE

The application is built on a decoupled, four-tier architecture to ensure maximum ease of maintenance and scalability:

A. INGESTION LAYER: This layer is responsible for parsing and converting fixed biomedical documents into vectors.

B. RETRIEVAL LAYER: Manages vector storage and similarity searches and protects against unverified data entering the context window.

C. GENERATION LAYER: This layer uses the LLM to generate an answer to the question based on the gathered data.

D. INTERACTION LAYER: This is the interaction layer, which is a lightweight web interface that includes the interaction layer for the application.

IV. IMPLEMENTATIONS STACK

The system is a pure software system that relies on the Python ecosystem. The backend for the web application is built with Flask, and the RAG pipeline is written on top of Lang Chain. Similarities for vectors are computed with FAISS. The transformer embedding models from Hugging Face are used for

vectorization, and an API-friendly LLaMA is used for text generation.

V. OPERATIONAL ADVANTAGES

The system has several advantages over traditional medical chatbots. It greatly reduces hallucinations by keeping answers grounded in verified documents, produces context-aware responses, and follows a safe, non-diagnostic design. Its modular architecture is well-suited for scalability and future enhancements, and its potential applications include medical education platforms, healthcare-related awareness portals, hospital information desks, telemedicine support tools, and AI-assisted learning environments.

VI. RESULTS AND DISCUSSION

The performance of the system was benchmarked against a standalone Large Language Model in a test suite representative of medical queries.

A. QUANTITATIVE PERFORMANCE ANALYSIS

Accuracy: The RAG-based system had a factual accuracy of 90%, whereas the standalone LLM had a maximum accuracy rate of 76%. In addition, the RAG approach showcased a "Low" hallucination rate against the "High" rate of the baseline model.

TABLE I: ACCURACY COMPARISON

System Type	Accuracy (%)	Hallucination rate
Standalone LLM	76%	HIGH
Proposed RAG-Based System	90%	LOW

LATENCY:

The average total response time ranged between 5 and 9 s, with document retrieval accounting for only 2–3 s owing to FAISS optimization. This confirms the viability of the system for real-time interactions.

Retrieval Precision: The system steadily retrieved three document chunks per query, with approximately 85% relevant context matching and the least possible retrieval failures.

TABLE II: RETRIEVAL PERFORMANCE

Metric	Observed Value
Retrieved Documents (k)	3
Relevant Context Match	~85%
Failed Retrievals	Minimal

B. DISCUSSION

Experimental results prove that the "retrieval-first" paradigm is effective in countering the reliability problems involved with Generative AI. By inculcating dependence on trustworthy content and assimilating constraints related to safety in the design principle of the prompt design paradigm, the solution has been efficient in thwarting potential ethical dilemmas related to unjustified diagnosis. The 90% accuracy threshold with sub-10-second latency points to its excellent optimization for educational engagement.

VII. CONCLUSION & FUTURE DIRECTIONS

A. SUMMARY OF CONTRIBUTION

In developing this study, the authors designed and tested the health information system that integrates document retrieval and text generation to offer health-based responses to health queries that are well-factual. The health information system shows that the integration of credible sources of health information and intelligent text generation improves the accuracy of the outputs while diminishing the instances of errors in the results, as opposed to the use of purely intelligent text generation methods for health queries and responses.

B. IMPACT ON HEALTHCARE INFORMATION SYSTEMS

Integration of retrieval mechanisms with language models is an important step forward for healthcare information technology. Conventional online search engines flood users with an awful lot of information that can be potentially contradictory, with little assistance to help differentiate between authentic and fake sources of information. Using only language models can be potentially hazardous because of hallucinations. Information retrieval can provide an equitable measure between the instinctive

communication paradigm of conversation interfaces and authenticity requirements in healthcare sessions.

C. ETHICAL CONSIDERATIONS AND SAFETY ASSURANCE

The implementation process of health information systems must be sensitive to issues of safety and ethics. The stated non-diagnostic and non-prescriptive approach properly indexes the system boundaries and minimizes risks. System behaviour analysis, user feedback analysis, and periodic reviews of expert feedback on materials within the knowledge base are important for guaranteeing safety and accuracy. It is important to maintain prominent disclaimers and recommendations to seek healthcare professionals on individual health matters.

D. SCOPE FOR ENHANCEMENTS AND EXTENSION

Knowledge Base Enhancement: The inclusion of literature related to general aspects of medicine, surgery, pharmaceutical data, and specialized topics of medicine would enhance system usage. The use of knowledge bases related to specialized fields of medicine could help tailor system implementation for specific scenarios.

MULTILINGUAL SUPPORT

Extending support to multiple languages, especially popular languages in various regions with healthcare needs, would greatly help improve the accessibility of this technology. Multilingual embeddings help embed documents and queries in multiple languages

INTEGRATION OF REAL-TIME INFORMATION

Existing static knowledge bases can be supplemented with carefully selected real-time information, such as outbreak notifications, medication recalls, and guideline changes, if real-time information is safely governed.

USER INTERACTION UPGRADE

Voice-based inquiry interfaces are helpful for users with literacy issues or visual impairments. The mobile application deployment feature expands accessibility beyond the web browser interface. The dialog history feature enables the model to retain the conversation flow and facilitate more complex interactions.

INTEGRATION WITH CLINICAL SYSTEMS

In an appropriate governance framework and with proper consideration for issues of privacy and security, integration with electronic health records or clinical decision support systems might potentially extend this utility into a clinical environment; however, this would also require a substantial safety infrastructure.

SPECIALIZED ADAPTATIONS

Custom system instances catering to specific audiences (example instances could be health-literacy oriented and/or heavily term-loaded catering to medical students, or evidence-loaded catering to healthcare professionals) could improve relevance to particular groups of users.

The relevance of the study is that it demonstrates the fact that the applicability of the language model based on the grounding in the retrieval process is much more beneficial in the context of information systems in the medical domain than the language model on its own. It has become possible due to the integration of trustworthy sources and the natural language process in the form of the user interface, which satisfies the requirements of applicability and effectiveness, hence being beneficial in the context of health education and health awareness programs. The rising requirements in the medical information domain and advancement in the artificial intelligence community call for increased applications of the proposed ideas in the domain of artificial intelligence.

REFERENCES

[1] Lewis, P., Perez, E., Piktus, A. et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[2] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 4171–4186.

[3] Reimers, N., and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP-IJCNLP, 3982–3992.

[4] Johnson, J., Douze, M., and Jégou, H. (2021). Billion-Scale Similarity Search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547.
[5] Singhal, K., et al. (2023). Large Language Models Encode Clinical Knowledge. Nature, 620, 172–180.
[6] Thirunavukarasu, A., Ting, A., Elangovan, K., et al. (2023). Large Language Models in Medicine: The Potential and Pitfalls. Frontiers in Digital Health, 5
[7] Gale Encyclopedia of Medicine. 2nd ed. Detroit, MI: Gale Group; 2018.