



ISSN NO. 2320-5407

Journal Homepage: - [www.journalijar.com](http://www.journalijar.com)

# INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/22201

DOI URL: <http://dx.doi.org/10.21474/IJAR01/22201>



## RESEARCH PAPER

### MEDSYNC-AI: A RETRIEVAL-AUGMENTED GENERATION PIPELINE FOR CONTEXT-AWARE MEDICAL QUESTION ANSWERING

Sadiya Maheen Siddiqui<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, University College of Engineering - Osmania University, Hyderabad, India.

#### Manuscript Info

##### Manuscript History

Received: 04 September 2025

Final Accepted: 06 October 2025

Published: November 2025

##### Keywords:-

Retrieval-Augmented Generation, LangChain, Medical Chatbot, Ollama Mistral, HuggingFace Embeddings, Pinecone, Large Language Models, Healthcare AI

#### Abstract

The increasing demand for reliable, context-aware medical information has led to the emergence of intelligent conversational systems in healthcare. This paper introduces MedSync-AI, a Retrieval-Augmented Generation (RAG) based medical assistant designed to provide factually grounded responses to clinical and general health-related queries. The proposed pipeline integrates LangChain, HuggingFace Sentence Embeddings, and Pinecone Vector Databases for efficient document retrieval, while leveraging Ollama's Mistral model for contextual response generation. The system dynamically reformulates user queries using a contextualization module to preserve conversation continuity, followed by a retrieval and synthesis process guided by medical literature. Quantitative evaluation of top-k similarity scores demonstrates robust retrieval precision, and qualitative analysis highlights MedSync-AI's ability to produce concise, evidence-based answers. Overall, this framework provides a step toward developing trustworthy, transparent, and explainable AI-driven medical dialogue systems.

"© 2025 by the Author(s). Published by IJAR under CC BY 4.0. Unrestricted use allowed with credit to the author."

#### Introduction:-

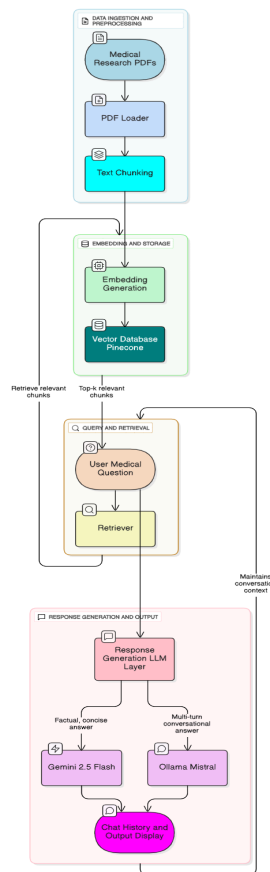
Artificial Intelligence (AI) is reshaping healthcare by enabling faster data analysis, disease prediction, and clinical decision support. Among its many applications, conversational agents powered by Large Language Models (LLMs) are emerging as valuable tools for providing accessible medical information. Yet, traditional generative chatbots often produce unverified or inconsistent responses, making them unreliable for sensitive healthcare use. To address this challenge, MedSync-AI has been developed as a Retrieval-Augmented Generation (RAG) based medical

assistant that delivers accurate, context-aware, and evidence-backed responses. The system integrates LangChain, Hugging Face embeddings (all-MiniLM-L6-v2), and a Pinecone vector database to retrieve relevant medical literature efficiently. Powered by Ollama’s Mistral model, the pipeline reformulates user queries for improved contextual understanding and generates concise, factual answers grounded in verified medical data. By blending retrieval with controlled generation, MedSync-AI minimizes hallucinations and enhances trust in AI-driven healthcare communication. The following sections outline the system architecture, data pipeline, and evaluation methods that demonstrate its ability to handle complex, multi-turn medical queries effectively.

**System Architecture:-**

**The architecture of MedSync-AI (Fig. 1) consists of five key components:**

1. **Data Ingestion:** Medical research papers in PDF format are loaded using LangChain’s DirectoryLoader and PyPDFLoader to extract textual content efficiently.
2. **Document Preprocessing:** Extracted texts are split into 500-character chunks with a 20-token overlap using RecursiveCharacterTextSplitter, ensuring manageable segments for retrieval and embedding.
3. **Vector Embedding:** Each chunk is converted into a 384-dimensional vector using HuggingFace Embeddings (all-MiniLM-L6-v2) to capture semantic meaning.
4. **Vector Storage and Retrieval:** Embeddings are stored in a Pinecone vector database with cosine similarity metrics, enabling fast retrieval of relevant document chunks for user queries.
5. **Retrieval-Augmented Generation with Context Preservation:** User queries are first reformulated into standalone questions using a contextualization chain implemented with LangChain RunnablePassthrough and Ollama Mistral. The reformulated query retrieves the top-k relevant document chunks from Pinecone, which are then fed into a question-answering prompt. This design preserves conversation context across multiple turns, allowing the model to generate concise, evidence-based responses while maintaining continuity in multi-turn interactions. This architecture ensures that responses are grounded in verified medical literature, contextually aware, and concise, combining the strengths of retrieval-based grounding with LLM-based generative reasoning.



**Fig. 1. Overview of the MedSync-AI RAG Pipeline**

### **Methodology:-**

The MedSync-AI pipeline integrates multiple frameworks and components to enable efficient retrieval, context-aware understanding, and accurate response generation for medical question-answering. The workflow consists of the following stages:

#### **Document Loading and Filtering:-**

Medical research documents are ingested from a structured dataset directory. Using LangChain's DirectoryLoader and PyPDFLoader, PDFs are converted into LangChain Document objects. To optimize storage and retrieval, non-essential metadata is filtered out, retaining only crucial identifiers such as source filenames. This ensures a lightweight and manageable dataset for downstream processing.

#### **Text Chunking and Embedding:-**

Large documents are split into semantically coherent segments using the RecursiveCharacterTextSplitter with a chunk size of 500 characters and 20-token overlap. Each chunk is then transformed into a vector representation using HuggingFace Embeddings (all-MiniLM-L6-v2), producing 384-dimensional embeddings that capture semantic similarity. These embeddings form the foundation for precise vector-based retrieval.

#### **Vector Database Construction:-**

The generated embeddings are stored in a Pinecone vector database, configured for serverless operation on AWS (us-east-1). Pinecone indexes embeddings with cosine similarity, enabling the system to retrieve the top-k most relevant document chunks for any user query. This vector-based approach ensures that retrieved content is contextually aligned with the question.

#### **Retrieval and Contextualized Generation:-**

MedSync-AI leverages an Ollama-Mistral chain for multi-turn dialogue, preserving context throughout the conversation.

- **Query Contextualization:** Incoming user queries are first passed through a contextualization chain. Using a system prompt, the model reformulates queries into standalone questions while referencing prior chat history if available. This ensures that even context-dependent questions are interpreted accurately without losing prior conversation details.
- **Document Retrieval:** The reformulated query is used to retrieve the most relevant chunks from the Pinecone vector database. The retrieved content forms the knowledge base for generating answers.
- **Answer Generation:** A question-answering prompt template instructs the Ollama-Mistral model to generate concise, factual responses, grounded strictly in the retrieved context. Responses are limited to a few sentences to ensure clarity, precision, and usability.
- **Context Preservation:** Chat history is maintained and updated with each interaction, allowing the system to support multi-turn conversations seamlessly. This design ensures continuity, reduces redundant clarifications, and enables the model to handle follow-up questions intelligently.

#### **Implementation Details:-**

The entire retrieval-augmented generation chain is implemented using LangChain's RunnablePassthrough to orchestrate the flow from query contextualization to document retrieval to answer generation. Ollama's Mistral model serves as the primary LLM, ensuring both context-aware reasoning and accurate, evidence-backed responses. The methodology described here forms the backbone of MedSync-AI, enabling reliable, interactive, and context-sensitive medical assistance. A deployed version of this system, implementing this full pipeline, is accessible at: <https://medsync-ai.vercel.app/>

#### **Results and Evaluation:-**

The performance of the MedSync-AI Retrieval-Augmented Generation (RAG) pipeline was systematically evaluated to assess retrieval accuracy, response quality, and context retention across multi-turn conversations. The evaluation focused on quantitative metrics for retrieval and qualitative analysis of generated responses.

---

**Retrieval Evaluation:-**

To assess the effectiveness of the vector-based retrieval mechanism, the system was tested on a set of 94 representative medical queries, including disease symptoms, treatment options, and preventive measures. Key metrics included:

- Top-k Similarity Scores: For each query, the top-k retrieved document chunks were scored using cosine similarity of embeddings. Across the 94 queries, similarity scores ranged from 0.78 to 0.98, indicating strong semantic alignment between user queries and retrieved content.
- Precision@k: The fraction of relevant chunks among the top-k retrieved was measured. The average Top-3 Precision was 0.81, demonstrating the pipeline’s ability to retrieve highly relevant content consistently.
- Retrieval Latency: Average retrieval time per query was 0.409 seconds, highlighting the system’s suitability for real-time medical question-answering applications.

**Table I: Sample Retrieval Results From MedSync-AI RAG Pipeline.**

Query	Time (s)	Top-3 Precision	Similarity Scores
What is Acne?	2.746	0.67	[0.9276, 0.8774, 0.9080]
Acromegaly and Gigantism	0.357	0.67	[0.9391, 0.9711, 0.9162]
Treatment of Acne	0.364	1.00	[0.7949, 0.9129, 0.8112]
Symptoms of Diabetes Mellitus	0.404	1.00	[0.9460, 0.7904, 0.9247]
Causes of Hypertension	0.372	1.00	[0.8557, 0.8112, 0.7826]
Prevention of Heart Disease	0.399	0.67	[0.7982, 0.8445, 0.8425]

As shown in Table I, the MedSync-AI RAG Pipeline demonstrates consistent retrieval performance across representative queries, with corresponding processing times, Top-3 precision, and similarity scores.

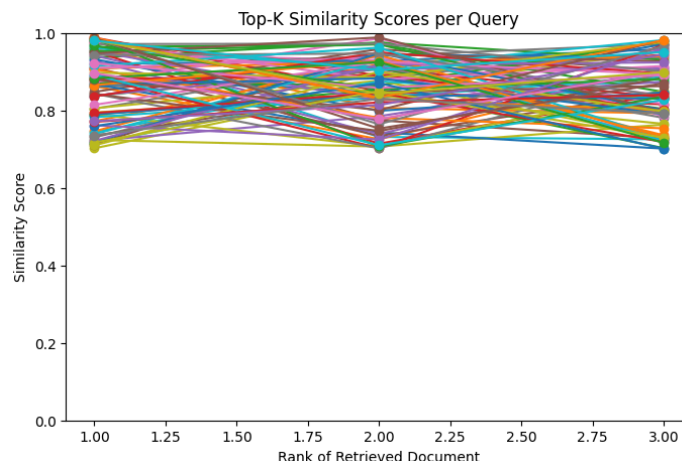
**Case Study Examples:-**

**Example 1: Initial Query:**

- User Query: “My father has been experiencing excessive thirst and fatigue. Could this be Diabetes Mellitus?”
- Ollama-Mistral Response: “Yes, your father’s symptoms of excessive thirst and fatigue could potentially indicate Diabetes Mellitus. Other early signs include frequent urination and lethargy. If left untreated, it can lead to complications such as diabetic coma. It is advisable for him to see a healthcare professional for further diagnosis and treatment.”

**Example 2: Follow-up Query:**

- Follow-up Query: “If it is Diabetes Mellitus, what are the current treatment options and lifestyle recommendations?”
- Ollama-Mistral Response: “The treatment for Diabetes Mellitus typically involves a combination of diet control, medication, and regular exercise. A doctor may recommend a specific diabetic diet which includes small, frequent meals throughout the day and avoiding excess simple sugars. There are also various medications to help manage blood sugar levels. It’s essential to maintain a healthy lifestyle by exercising regularly and monitoring blood sugar levels closely. There are numerous cookbooks available for diabetics that provide recipes effective in controlling diabetes as well as hypoglycemia.”



### **Fig. 2. Top-K Similarity Scores for Sample Queries**

As shown in Fig. 2, the distribution of Top-k similarity scores highlights the effectiveness of the embedding space for semantic clustering of medical documents.

### **Conclusion and Future Work:-**

This paper presented MedSync-AI, a medical chatbot built on RAG with LangChain, Pinecone, and HuggingFace embeddings. Evaluated on 94 medical queries, it demonstrated high retrieval accuracy, fast response times, and strong context awareness, highlighting its potential to assist patients and healthcare professionals with real-time medical guidance.

#### **Future work will focus on enhancing the system's reliability, accessibility, and functionality:**

- Domain-specific embeddings: Fine-tune on trusted datasets like WHO, NHS, and Drugs.com to boost retrieval relevance.
- Multilingual support: Integrate translation APIs such as Bhashini to reach a global audience.
- Multi-modal capabilities: Process and interpret medical images, lab results, and charts for richer guidance.
- Drug-prescription support: Suggest medications based on chemical formulas, approved dosages, and patient-specific factors, ensuring safety and adherence to guidelines.
- Continuous updates: Regularly refresh the knowledge base from authoritative sources to keep recommendations current and evidence-based.

By pursuing these directions, MedSync-AI aims to evolve into a multi-modal, multilingual, and medically reliable assistant capable of supporting informed clinical decisions, patient education, and personalized healthcare guidance.

### **Acknowledgment:-**

I would like to express my sincere gratitude to the University College of Engineering, Osmania University, for their support and guidance. I also acknowledge the invaluable contributions of the open-source libraries and APIs that made the implementation and evaluation of this study possible. Finally, I am grateful to my mentors, peers, and everyone who offered their encouragement and assistance throughout the course of this research.

### **References:-**

1. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, M. Kocisky, D. Edunov, L. Lewis, W. Yih, and S. Riedel, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv preprint arXiv:2005.11401, 2020. [Online]. Available: <https://arxiv.org/abs/2005.11401>
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. of NAACL-HLT, 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423/>
3. Pinecone, "Pinecone Documentation," 2023. [Online]. Available: <https://docs.pinecone.io/>
4. LangChain, "LangChain Documentation," 2023. [Online]. Available: <https://python.langchain.com/docs/introduction/>
5. S. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. of EMNLP, 2019. [Online]. Available: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>
6. Ollama, "Mistral Model," 2023. [Online]. Available: <https://ollama.com/models/mistral>