JSPM’s
Rajarshi Shahu College of Engineering, Tathawade, Pune
(An Empower Autonomous Institute under Savitribai Phule Pune University)

Department of Computer Science and Business Systems
Artificial Intelligence Mini Project
AY-2025-26

Medical Chatbot

RAG for Medical Information Retrieval

Project Team

- Pratik Rajendra Patil (RBT23CB026)
- Atharv Pramod Barge (RBT23CB034)

Faculty Guide

Prof. Parul Rajwade

Problem & Objective

Addressing LLM Hallucination in Medical Domains

THE CHALLENGE

Standard Large Language Models (LLMs) are prone to **hallucination**—generating plausible but incorrect or fabricated information. In medical domains, this poses a serious risk to user safety and trust.

Risk: Inaccurate medical information can mislead users and compromise their health decisions.

OUR SOLUTION

- ✓ Build a factually-grounded Q&A chatbot using RAG architecture
- ✓ Ground all answers exclusively in The Gale Encyclopedia of Medicine
- ✓ Eliminate hallucination through retrieval-augmented generation
- ✓ Provide a user-friendly, voice-enabled interface

KEY FEATURES

- ▸ Context-Aware Answers: Strictly based on encyclopedia content
- ▸ Voice-Enabled UI: Hands-free interaction with speech-to-text and text-to-speech
- ▸ Interactive Interface: Clean, responsive Streamlit application
- ▸ Medical Disclaimer: Prominently displayed for responsible use
- ▸ Fast Retrieval: Efficient FAISS vector database search

SCOPE & LIMITATIONS

- ▸ Knowledge Source: The Gale Encyclopedia of Medicine (Static)
- ▸ No Live Internet: Strictly limited to provided knowledge base
- ▸ Informational Only: Not a substitute for professional medical advice
- ▸ Stateless: No user data or chat history stored

Aligned with

SDG 3: Good Health & Well-being

RAG Architecture

Retrieval, Augmentation, and Generation Pipeline

Core Concept: RAG enhances a Large Language Model (LLM) by integrating an external knowledge retrieval system. This hybrid approach ensures factually accurate answers grounded in trusted sources.

1

Retrieve

When a user asks a question, the system embeds the query and searches the FAISS vector database to retrieve the most relevant text chunks from The Gale Encyclopedia of Medicine.

Technology: FAISS (Facebook AI Similarity Search) for fast, high-speed document retrieval

2

Augment

The retrieved text chunks are injected into the prompt as context, alongside the user's original question, creating a rich, contextual input.

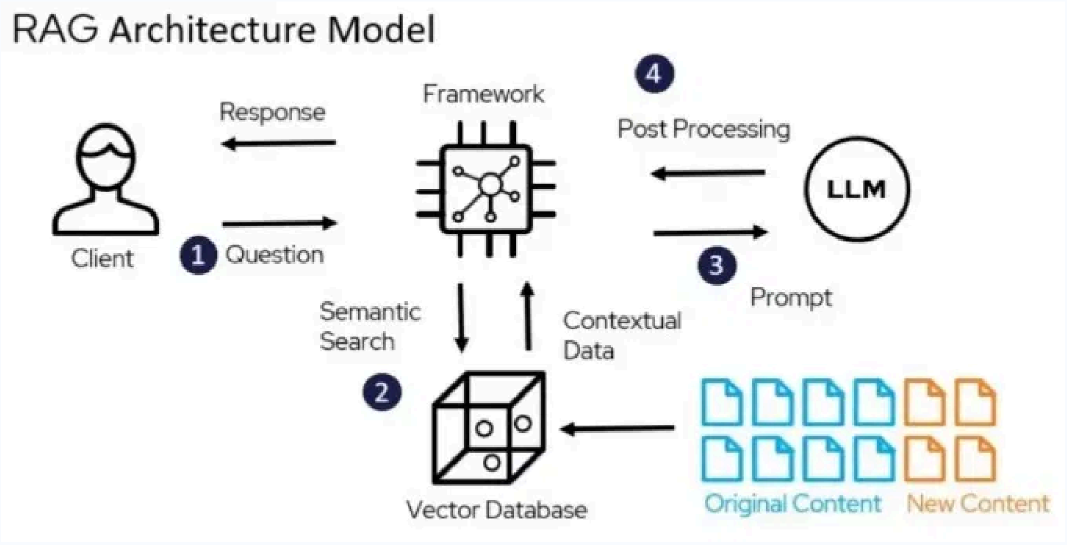
The augmented prompt ensures the LLM has access to verified, source-based information

3

Generate

The Mistral-7B LLM processes the augmented prompt and generates a factually-grounded answer based exclusively on the provided context.

Custom prompt enforces: "If you don't know, say so. Don't make up answers."



WHY RAG FOR MEDICAL DOMAINS?

- → Eliminates hallucination through source grounding
- → Enables source verification and transparency
- → Knowledge base updates without LLM retraining
- → Fast, efficient retrieval at scale

Technical Implementation

Three-Module Architecture

1 Data Ingestion & Vectorization

Purpose: Process source PDF and create searchable knowledge base

→ Load: Ingest The Gale Encyclopedia of Medicine PDF
→ Chunk: Split document into overlapping text pieces
→ Embed: Convert chunks to vectors using all-MiniLM-L6-v2
→ Store: Save vectors in FAISS database

Tech: LangChain, HuggingFace Embeddings, FAISS

2 Core RAG Pipeline

Purpose: Orchestrate query, retrieval, and answer generation

→ Retrieve: Find relevant text chunks from FAISS database
→ Augment: Inject retrieved chunks into prompt as context
→ Generate: Send augmented prompt to Mistral-7B LLM
→ Return: Deliver factually-grounded answer to user

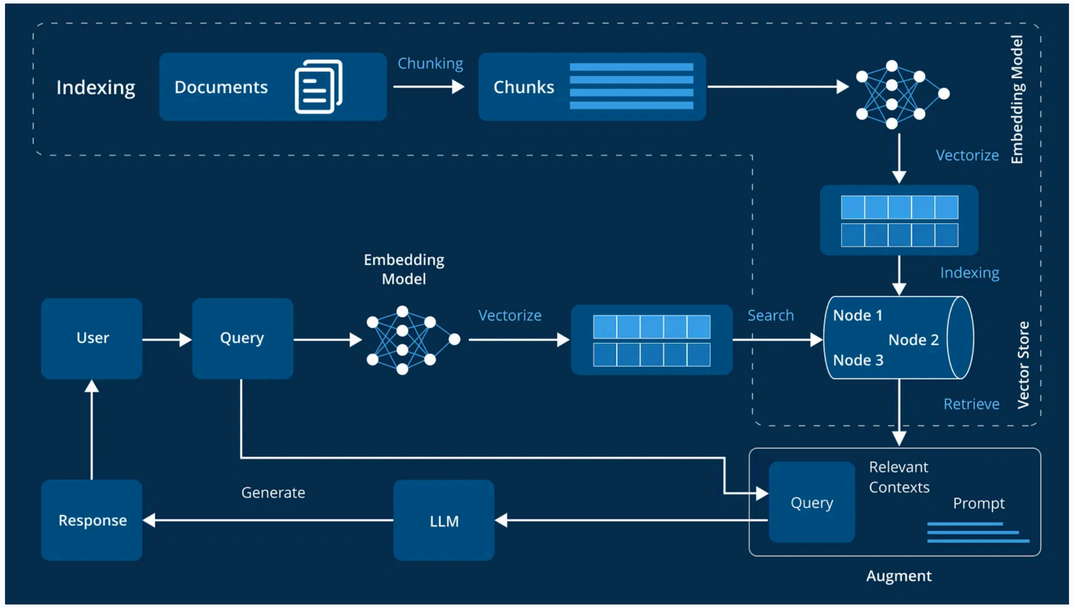
Tech: LangChain, HuggingFace Endpoint, Mistral-7B

3 User Interface (Streamlit)

Purpose: Provide interactive front-end for users

→ Chat Interface: Web-based conversation management
→ Voice Input: Speech-to-text via streamlit_mic_recorder
→ Voice Output: Text-to-speech via gTTS
→ Display: Show chat history and responses in real-time

Tech: Streamlit, gTTS, streamlit_mic_recorder



Complete RAG Pipeline: From Data Ingestion to Response Generation

Results, Conclusion & Future

Project Outcomes and Strategic Enhancements

KEY RESULTS

FUNCTIONAL PROTOTYPE

Successfully deployed "MediChat," a fully operational voice-enabled medical Q&A chatbot with text and speech interfaces.

HALLUCINATION ELIMINATION

RAG pipeline consistently produces answers grounded exclusively in The Gale Encyclopedia of Medicine, preventing fabricated information.

PERFORMANCE

Fast query response times achieved through efficient FAISS vector database retrieval and optimized embedding search.

ACCURACY

All generated answers verified against source text chunks, ensuring factual correctness and traceability.

CONCLUSION

The Retrieval-Augmented Generation (RAG) architecture has proven to be a highly effective and reliable solution for building domain-specific, factually-grounded AI assistants from static knowledge bases.

By combining the generative capabilities of Large Language Models with the accuracy of retrieval systems, we have successfully transformed a general-purpose LLM into a specialized medical expert system.

> **Key Insight: RAG is the optimal approach for applications requiring high factual accuracy and source verifiability, particularly in safety-critical domains like healthcare.**

FUTURE ENHANCEMENTS

SOURCE DISPLAY

Enhance UI to display retrieved source text chunks alongside answers for user verification and transparency.

FORMAL EVALUATION

Implement quantitative metrics (e.g., RAGAs framework) for rigorous performance analysis and benchmarking.

DYNAMIC KNOWLEDGE

Integrate with live database systems to enable real-time knowledge base updates without system downtime.

USER FEATURES

Add conversation history, export options, and personalized user profiles for enhanced engagement.

References & Technology Stack

Core Technologies and Knowledge Sources

RAG ORCHESTRATION

- LangChain

  Framework for building and orchestrating the RAG pipeline with modular components

VECTOR STORAGE & RETRIEVAL

- FAISS (Facebook AI Similarity Search)

  High-speed, efficient vector database for fast document retrieval at scale

LANGUAGE MODEL

- Mistral-7B-Instruct-v0.2

  Open-source LLM for generating factually-grounded answers

WEB INTERFACE

- Streamlit

  Python framework for building interactive, responsive web applications
- streamlit_mic_recorder & gTTS

  Speech-to-text and text-to-speech capabilities for voice interaction

MODEL ACCESS & EMBEDDINGS

- Hugging Face Hub

  Access to LLMs and embedding models; all-MiniLM-L6-v2 for text embeddings

CORE RESEARCH

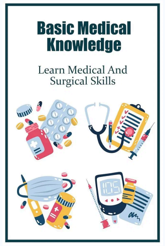
Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
Lewis, P., et al. (2020). NeurIPS.

Mistral 7B
Jiang, A. Q., et al. (2023). Mistral AI.

Sentence-Transformers: Semantic Textual Similarity
Reimers, N., & Gurevych, I. (2019).

KNOWLEDGE SOURCE

The Gale Encyclopedia of Medicine
Second Edition
Longe, J. L. (Ed.). (2002). Gale Group.



We extend our gratitude to **Prof. Parul Rajwade** for her invaluable guidance and support throughout this project. Special thanks to JSPM's Rajarshi Shahu College of Engineering for providing the resources and infrastructure to develop this innovative solution.