Project Proposal

# MediBot

## — AI Medical Chatbot using Retrieval-Augmented Generation (RAG)

Submitted by
SHAHBAZ NOOR (FA24-MSDS-0014)
KEEMAT RAI (FA24-MSDS-0024)

# MediBot — AI Medical Chatbot using Retrieval-Augmented Generation (RAG)

## Project Idea / Problem Statement

The exponential growth of medical literature, research papers, and clinical documents presents a major challenge: how can healthcare professionals and researchers quickly retrieve accurate, contextually relevant information from unstructured medical data?

MediBot addresses this problem by enabling natural language interaction with medical documents. It is a Retrieval-Augmented Generation (RAG) based chatbot that allows users to ask medical questions and receive precise answers grounded in uploaded medical documents. Unlike traditional chatbots, MediBot does not rely on predefined intents or static responses—it dynamically retrieves relevant content using vector similarity and augments a powerful LLM to generate context-specific responses.

## Dataset Source

MediBot does not use a traditional labeled dataset. Instead, it uses:

- User-uploaded PDFs or predefined documents as the knowledge base.
- These documents are chunked and embedded using sentence-transformers into a vector database (FAISS).
- During inference, relevant chunks are retrieved from this vector store based on the user's query and fed into a language model.

## Models to be Used

1. Embedding Model:
   - SentenceTransformers / MiniLM-L6-v2
   - Purpose: Convert document chunks into high-dimensional vectors for similarity search.

2. Vector Store:
   - FAISS (Facebook AI Similarity Search)
   - Purpose: Efficient similarity search over document embeddings to retrieve the most relevant chunks.

3. Large Language Model (LLM):
   - Mistral 7B Instruct (via HuggingFace Inference API)
   - Purpose: Generate answers using both the retrieved context and the user’s query, based on a prompt template.

4. RAG Pipeline (Implemented via Langchain):
  - Combines retrieval and generation to ensure factual, context-grounded responses.

## Reason for Choosing Specific Models

- MiniLM-L6-v2 Embedding Model: Offers a strong balance of performance and efficiency. It enables high-quality semantic similarity search at low computational cost—ideal for near real-time RAG use cases.

- FAISS Vector Store: State-of-the-art for vector similarity search; it scales well and provides fast indexing and retrieval for large document corpora.

- Mistral-7B (via HuggingFace): Chosen for its superior instruction-following capabilities at a manageable size and cost. It allows secure, scalable generation without requiring local GPU resources.

- Langchain Framework: Greatly simplifies the implementation of complex RAG workflows. It provides modular integration between LLMs, prompts, retrievers, and user interfaces.

## Conclusion

MediBot represents a modern AI-driven solution for document-grounded medical information retrieval. Its RAG architecture ensures that responses are accurate, traceable to source documents, and up-to-date. It can be extended for use in healthcare education, research support, or internal hospital knowledge management systems.

Future enhancements include:
- Multi-document ingestion
- UI-level user authentication
- Upload support for end users
- Integration with private clinical data sources under secure conditions