Real Time medical Chatbot using NLP+RAG.

This project involves the design and development of a real-time medical chatbot that uses Natural Language Processing (NLP) techniques to answer health-related questions intelligently. The system combines a retriever-generator (RAG) architecture, ensuring that the responses are both accurate and context-aware.

Technology Stacks:

- Programming Language: Python
- Frontend: Streamlit
- Model: Google FLAN-T5 (transformers library)
- Retrieval Technique: TF-IDF with cosine similarity (Scikit-learn)
- Deployment (Real-time): Pyngrok + Streamlit in Google Colab
- Dataset: Custom-built Medical Q&A (medical_qa.txt)

Working flow:

User Query Input:

The user types a health-related question via the Streamlit interface.

Retriever Module (TF-IDF):

- Uses TF-IDF vectorization to search the top 3 relevant questions from a medical Q&A knowledge base.
- Applies cosine similarity to measure relevance.

Generator Module (FLAN-T5):

- Combines the user's query with the retrieved context.

- Generates a dynamic, fluent answer using the transformer-based FLAN-T5 model.

Response Display:
The chatbot shows a natural-language response in real time along with the source question for transparency.