1. Approach Used

1.1. Data Storage & Retrieval

- The medical knowledge base is stored in **FAISS (Facebook AI Similarity Search)**, a vector database optimized for fast similarity search.
- **Google Generative AI Embeddings** are used to encode medical text into numerical vectors for efficient search and retrieval.

1.2. Conversational AI with Memory

- The chatbot uses **LangChain's LLMChain** to generate responses from **Google Gemini 1.5 Pro**.
- It maintains **session-based chat history** to ensure smooth, natural conversations.
- A **custom prompt** ensures that responses are **concise, engaging, and context-aware**.

1.3. User Interaction via Streamlit

- The chatbot is deployed using **Streamlit**, allowing a simple, interactive interface.
- Users can enter health-related queries, and the system responds based on stored medical data.

2. Challenges Faced

2.1. Maintaining Context in Conversations

- Initially, the chatbot explicitly mentioned, *"Given our previous conversation,"* which made interactions less natural.
- To improve this, the prompt was adjusted to **infer context without explicitly stating it**.

2.2. Handling Ambiguous Queries

- When users used pronouns like "it" or "this," the chatbot struggled to infer context correctly.
- A rule was added to **assume references point to the most recent topic** unless specified otherwise.

2.3. FAISS Search Accuracy

- The chatbot initially retrieved **only the top 1 result**, sometimes missing relevant context.
- Increasing **k=5** for FAISS similarity search improved retrieval accuracy.

3. Model Performance & Improvements

3.1. Response Relevance

✅ Improved retrieval by **increasing FAISS similarity search depth**.
✅ Enhanced **context retention** using **session-based chat history**.

3.2. Conversational Flow

✅ Adjusted prompts to remove robotic phrasing and create **natural conversations**.
✅ Ensured chatbot handles **pronouns and references smoothly**.

3.3. Future Improvements

🔷 Fine-tune **retrieval ranking** to prioritize the most relevant medical information.
🔷 Integrate **differential response styles** (e.g., detailed vs. summary mode).
🔷 Expand **multi-turn memory** beyond a single session for longer conversations.