1. Data Ingestion

Data Sources:

- **Curated Documents:** The system uses 30 pre-selected documents covering various medical topics. These documents provide a rich, diverse base of medical information.
- **Web Scraping:** In addition to the curated documents, content is also collected through web scraping from 5 trusted websites. This ensures that the system includes the most recent and relevant information from reliable online sources.

Models Used: PubMedBERT

- Utilizes a domain-specific BERT model pretrained on biomedical literature, ensuring effective comprehension of medical terminology and context.

Key Components:

- **Document Chunking (512 Token Size):**
  Splits both the curated documents and web-scraped content into manageable semantic chunks that adhere to the model's 512-token processing limit.
- **Metadata Extraction:**
  Extracts vital metadata such as titles, authors, publication dates, and source details, which is crucial for later reference and traceability.
- **FAISS Index Creation:**
  Constructs a FAISS index to store vector embeddings, facilitating rapid and efficient similarity search over the ingested data.
- **JSON Storage for Chunks:**
  Stores each document chunk along with its metadata in JSON format, ensuring that the information is well-organized and easily retrievable.

2. Information Retrieval

- **Models & Tools:**

  FAISS (Facebook AI Similarity Search):

  Provides an efficient framework for quickly locating similar vector embeddings.

  PubMedBERT Embeddings:

  Uses embeddings generated from PubMedBERT, ensuring that the medical context is preserved in the similarity search.

  Cosine Similarity Ranking:

  Applies cosine similarity for a finer ranking of search results based on the angular similarity of the vectors.

3. LLM Integration

- **Model: Google's Gemini**

Employs an advanced language model designed to understand complex queries and provide context-aware, evidence-based responses.

- **Features:**

  Medical Context Awareness:

  Maintains the medical context throughout the response generation.

  Evidence-Based Responses:

  Generates answers grounded in verified medical literature and data.

  Source Attribution:

  Tracks and cites the origins of the information provided.

  Accuracy Verification:

  Incorporates mechanisms to ensure the accuracy of medical information.

3. References, Disclaimers, and Follow-Up Questions

This integrated component ensures that every response is not only informative and accurate but also responsibly presented with verifiable sources, necessary legal disclaimers, and dynamic follow-up questions to encourage deeper exploration.

References:

Automatic Citation: Extracts and formats metadata from sources (curated documents and trusted websites) to generate verifiable citations with document links and page/section references.

Source Verification: Ensures that all referenced materials are credible and reliable.

Medical Disclaimers:

Automatic Inclusion: Every response includes a disclaimer stating that the information is not a substitute for professional medical advice.

Context-Specific Warnings: Provides emergency guidance and advises consultation with healthcare professionals when needed, while ensuring HIPAA compliance.

Follow-Up Questions:

Dynamic Suggestions: Generates follow-up questions based on the initial query and response, exploring related topics, symptoms, treatment options, and risk factors to encourage deeper engagement.

4. Integration Points

- **API Layer: FastAPI**

  Facilitates efficient communication between different system components through a robust API interface.

- **UI: Streamlit**

  Provides a user-friendly interface that allows easy interaction and data visualization.

- **Storage:**

  - **FAISS Vector Database:** Manages and stores high-dimensional embeddings for similarity search.
  - **JSON Document Store:** Organizes and stores document chunks and their metadata in a structured JSON format.

- **Security:**

  API Key Management: Ensures that only authorized users can access the system via secure API keys.

  HIPAA Compliance: Incorporates necessary security measures to handle sensitive medical data.

Architecture diagram