**Multimodal Retrieval-Augmented Generation for Chest X-rays Using MIMIC-CXR and Groq LLM**

**Abstract**

This project presents a multimodal retrieval-augmented generation (RAG) system that integrates radiological image analysis with natural language generation. Utilizing the MIMIC-CXR dataset and a large language model (LLM) hosted by Groq, the system enables the generation of clinically relevant findings and impressions based on user-uploaded chest X-ray images and textual queries. Through the use of joint image-text embeddings, the system retrieves similar historical cases and incorporates them as contextual information to guide the generative process. This report outlines the dataset, methodology, tools employed, system design, and results, concluding with a discussion on limitations and future directions.

**Introduction**

The interpretation of chest radiographs is a fundamental task in clinical radiology, often requiring the synthesis of visual and contextual information to reach accurate diagnostic conclusions. Recent advancements in vision-language models have opened new possibilities for automating this process. The objective of this project is to develop an end-to-end system that retrieves semantically similar radiology reports from a dataset and utilizes them to produce coherent and clinically informed summaries in response to user queries. This is achieved by embedding both image and textual data using a vision-language model and employing a Groq-hosted LLM to generate final outputs.

**Dataset**

The project uses the MIMIC-CXR dataset, an openly available and de-identified collection of chest X-ray images and associated free-text reports. The dataset was accessed through Hugging Face's dataset hub, specifically from the repository identified as "itsanmolgupta/mimic-cxr-dataset." Each record in the dataset contains:

* A byte-encoded chest X-ray image
* A corresponding “Findings” section describing radiological observations
* An “Impression” section summarizing the clinical diagnosis

For the purposes of this prototype, a subset of 100 samples was extracted to balance computational efficiency and representativeness.

**Methodology**

**Multimodal Embedding**

To enable joint reasoning over image and textual data, each X-ray image and its associated "Findings" text were embedded using the CLIP (Contrastive Language-Image Pretraining) model. The image and text embeddings were normalized and concatenated into a single feature vector, forming the basis for multimodal comparison.

**Similarity Search**

A FAISS index was constructed using the embedded vectors of the dataset subset. FAISS (Facebook AI Similarity Search) provides fast vector similarity search, enabling efficient retrieval of cases most similar to a user-submitted query in both visual and textual dimensions.

**User Interaction**

Users interact with the system via a graphical web interface. They are prompted to upload a chest X-ray image and submit a textual query (e.g., "What abnormalities do you observe?"). The system embeds the user inputs and retrieves the top-k most relevant cases from the FAISS index. These retrieved cases are then formatted as contextual input for the language model.

**Language Model Generation**

The retrieved reports are provided to a large language model hosted by Groq. The LLM receives a structured prompt containing the user's query and relevant historical cases. It then generates two distinct outputs:

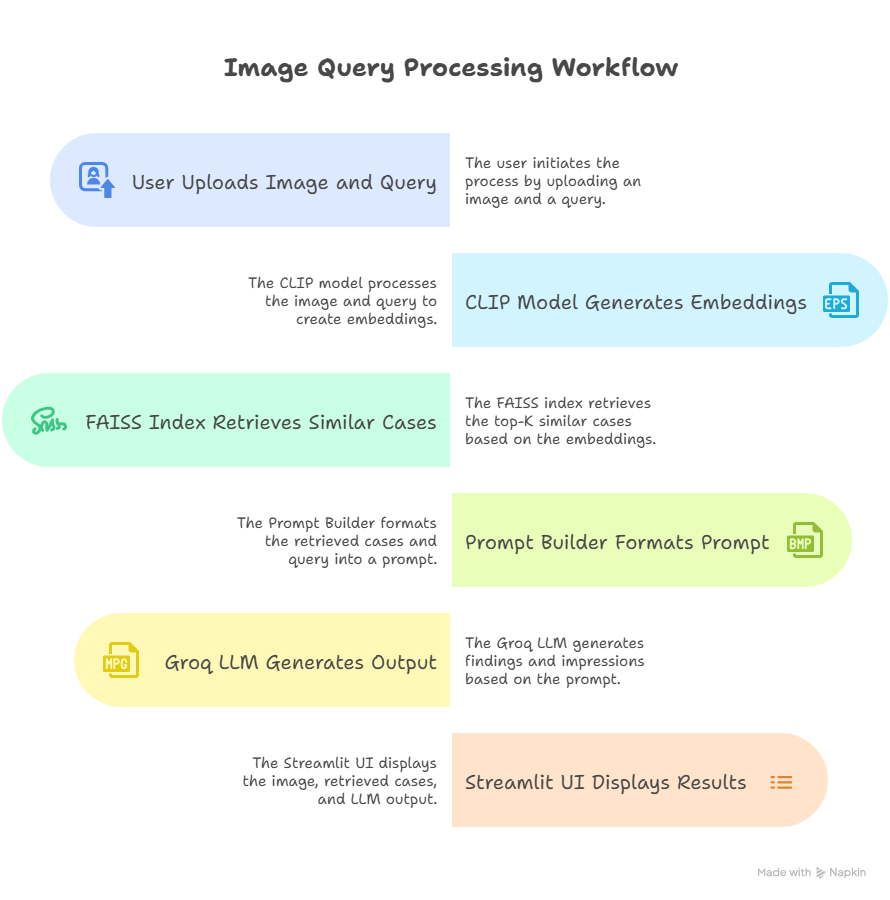
* **Findings**: A detailed, structured analysis of the X-ray
* **Impression**: A concise clinical summary

**System Architecture**

The system architecture comprises the following components:

* **Data Processing Module**: Responsible for loading, cleaning, and embedding the dataset
* **FAISS Index Module**: Stores and queries multimodal vectors
* **User Interface**: Developed using Streamlit for real-time user interaction
* **Language Model API Module**: Communicates with the Groq API to generate natural language outputs

The system is deployed locally and is dependent on GPU support for efficient embedding using CLIP.



**Results**

The prototype successfully demonstrated the feasibility of multimodal RAG for chest X-ray interpretation. Users were able to upload images and receive clinically plausible findings and impressions generated by the LLM, grounded in similar past cases. The retrieval system was capable of identifying semantically relevant examples, thereby enhancing the quality of the generated outputs.

Qualitative evaluation showed that the generated reports often captured the key features of the input images and reflected the style of formal radiological reporting. However, clinical validation by domain experts remains necessary for deployment in real-world settings.

**Conclusion**

This project presents a viable approach to combining image analysis and natural language generation for radiology using a multimodal RAG framework. By leveraging joint embeddings and contextual retrieval, the system is capable of generating structured diagnostic narratives from user-submitted chest X-rays and questions. While the current implementation is limited in scale, it demonstrates the foundational principles and practical utility of combining vision-language modeling with LLMs in clinical applications.