Abstract

This report presents the design and implementation of a medical question-answering system based on Retrieval-Augmented Generation (RAG). Users upload medical PDFs or text documents, and the system uses semantic search with ChromaDB and a local language model (LLaMA3 via Ollama) to retrieve relevant context and generate informative answers. We discuss the workflow, document indexing, retrieval pipeline, and natural language generation, along with challenges and future scope.

Contents

1 Introduction **3**

2 Objective **3**

3 Methodology **3**

3.1 Architecture Overview 3

3.2 Document Processing and Vector Indexing 4

3.3 Retrieval and Prompt Construction 4

3.4 Local LLM Inference 4

3.5 Frontend Interface 4

4 Results and Discussion **5**

4.1 Performance and Observations 5

4.2 Example Questions and Responses 5

4.2.1 Document-Based Questions and Answers 6

4.2.2 General Medical Questions and Answers 7

4.3 Strengths 7

4.4 Limitations 7

5 Conclusion **7**

6 References **8**

1 Introduction

As healthcare data grows rapidly, extracting useful knowledge from medical reports becomes increasingly important. Traditional search fails to understand context, while modern language models often hallucinate when queried without grounding. To overcome these limitations, we adopt Retrieval-Augmented Generation (RAG), which enhances generative models with document retrieval. This approach is applied to create a web-based assistant that answers medical queries using uploaded files.

2 Objective

- Implement a document-aware QA assistant using the RAG framework.
- Integrate PDF and TXT ingestion with vector storage (ChromaDB).
- Use semantic embeddings and retrieval to provide relevant chunks.
- Leverage a local LLM (LLaMA3 via Ollama) for answer generation.
- Provide a user-friendly Streamlit interface for document upload and Q&A.

3 Methodology

3.1 Architecture Overview

The architecture follows a modular pipeline:

1. **Document Upload:** Users upload PDFs or TXT files via Streamlit.
2. **Text Extraction:** Text is parsed using PyPDFLoader or TextLoader.
3. **Embeddings:** Embeddings are created using `OllamaEmbeddings`.
4. **Vector Store:** ChromaDB persists the document embeddings.
5. **Retriever:** A similarity-based retriever fetches top-k relevant chunks.
6. **RAG Chain:** Retrieved context and user query are passed to LLaMA3.
7. **Response Generation:** LLM generates an answer which is shown on the UI.

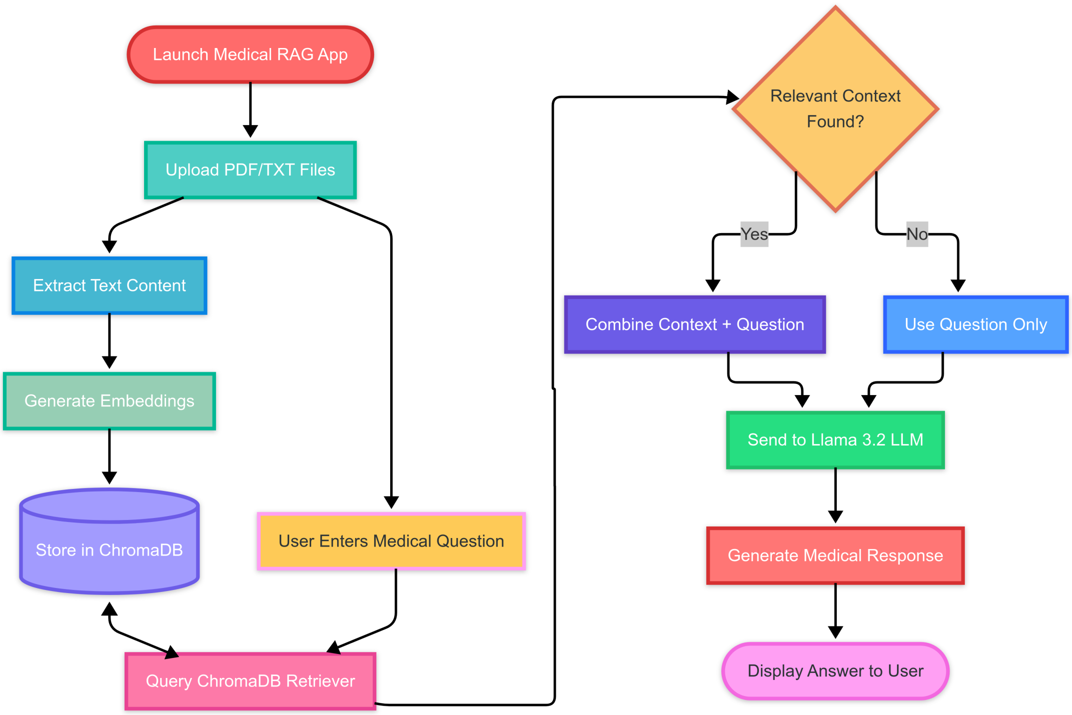


Figure 1: Workflow of the Medical RAG Assistant

3.2 Document Processing and Vector Indexing

Uploaded files are saved in a local directory. For PDFs, each page is treated as a document chunk. The text is converted into embeddings using a sentence transformer model via Ollama. These embeddings are stored in ChromaDB for persistent vector search.

3.3 Retrieval and Prompt Construction

When a user enters a question, the retriever searches for semantically similar chunks using cosine similarity. If relevant context is found, it is injected into a prompt template along with the user question. Otherwise, a fallback prompt is used. This prompt is sent to the LLM using LangChain's chain abstraction.

3.4 Local LLM Inference

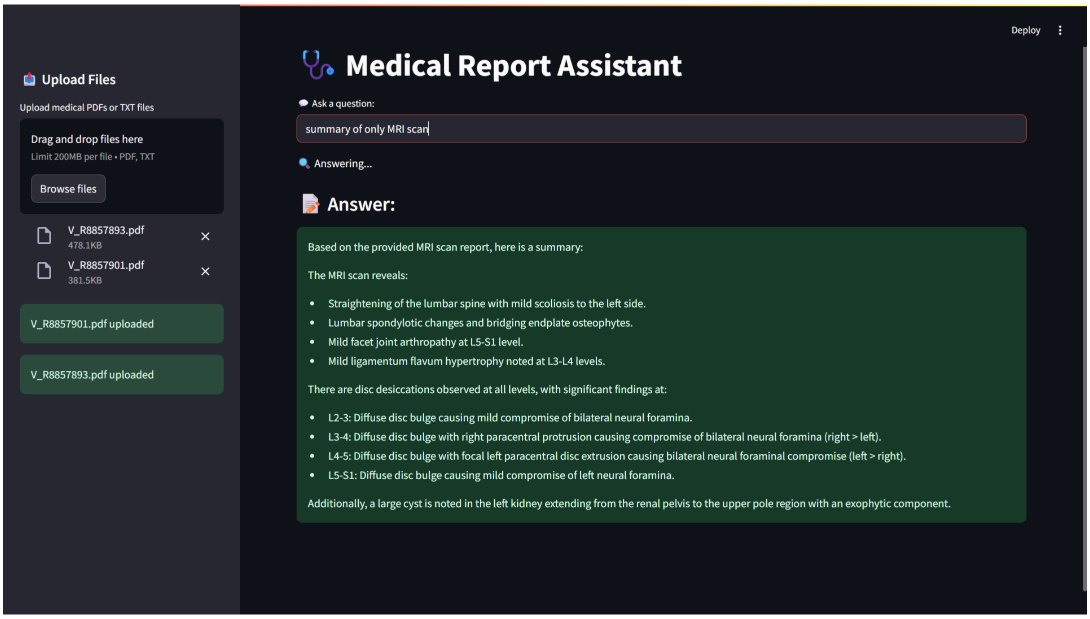
The LLM used is `llama3.2` served via Ollama. It operates locally and avoids the latency and cost of cloud APIs. The LLM uses the input prompt to generate responses grounded in the provided context.

3.5 Frontend Interface

Streamlit is used to create a clean, interactive interface:

- File uploader (PDF, TXT)
- Sidebar progress and feedback
- Question textbox

- Answer output area



V_R8857901.pdf uploaded

V_R8857893.pdf uploaded

Ask a question: summary of only MRI scan

Based on the provided MRI scan report, here is a summary:

The MRI scan reveals:

- Straightening of the lumbar spine with mild scoliosis to the left side.
- Lumbar spondylotic changes and bridging endplate osteophytes.
- Mild facet joint arthropathy at L5-S1 level.
- Mild ligamentum flavum hypertrophy noted at L3-L4 levels.

There are disc desiccations observed at all levels, with significant findings at:

- L2-3: Diffuse disc bulge causing mild compromise of bilateral neural foramina.
- L3-4: Diffuse disc bulge with right paracentral protrusion causing compromise of bilateral neural foramina (right > left).
- L4-5: Diffuse disc bulge with focal left paracentral disc extrusion causing bilateral neural foraminal compromise (left > right).
- L5-S1: Diffuse disc bulge causing mild compromise of left neural foramina.

Additionally, a large cyst is noted in the left kidney extending from the renal pelvis to the upper pole region with an exophytic component.

Figure 2: Streamlit Interface of Medical RAG Assistant

4 Results and Discussion

4.1 Performance and Observations

- Indexing is done only for new documents (avoiding re-embedding).
- LLaMA3 produces fluent and relevant answers with short latency.
- ChromaDB vector store ensures efficient retrieval and scalability.

4.2 Example Questions and Responses

This section presents ten example interactions with the Medical RAG Assistant. The first five questions are based on uploaded radiology reports of Mr. Sanjay Kumar Paul. The latter five are general medical queries not found in the uploaded documents.

4.2.1 Document-Based Questions and Answers

Q1	What is a diffuse disc bulge and how does it affect the thecal sac?
A1	A diffuse disc bulge refers to abnormal protrusion of disc tissue that can compress the thecal sac, leading to symptoms like pain, numbness, or weakness. In Mr. Paul's MRI, multiple disc bulges caused mild neural foraminal compromise.
Q2	What does "abutting the L4 nerve root" mean?
A2	It means the disc bulge at L3–L4 is touching or pressing against the L4 nerve root, possibly causing nerve irritation or compression without full displacement.
Q3	What does "mild scoliosis to the left" indicate?
A3	It indicates a slight curvature of the lumbar spine to the left, commonly seen in aging spines and often asymptomatic.
Q4	What is ligamentum flavum hypertrophy?
A4	It is a thickening of a spinal ligament that can narrow the spinal canal. Mild hypertrophy at L3–L4 was noted in the report, likely related to age-related degeneration.
Q5	What does "desiccation of discs" imply?
A5	It refers to drying and degeneration of intervertebral discs, leading to reduced disc height and cushioning. This was noted throughout the lumbar spine in Mr. Paul's case.

Table 1: Questions and Answers Based on Uploaded Medical Reports

4.2.2 General Medical Questions and Answers

Q6	What is the normal range for white blood cell (WBC) count?
A6	The normal adult WBC count ranges from 4,500 to 11,000 cells per microliter. It may vary slightly across labs.
Q7	What are common symptoms of diabetes?
A7	Common symptoms include frequent urination, excessive thirst, fatigue, blurred vision, slow healing, and numbness in extremities.
Q8	How is high blood pressure treated?
A8	Treatment includes lifestyle changes (diet, exercise), and medications like ACE inhibitors, beta blockers, and diuretics based on patient needs.
Q9	What is the function of the liver in the human body?
A9	The liver detoxifies blood, produces bile for digestion, stores glycogen, and regulates metabolism and hormones.
Q10	What are common side effects of antibiotics?
A10	Side effects may include nausea, diarrhea, allergic reactions, yeast infections, and increased risk of C. diff infections.

Table 2: General Medical Questions and Answers

4.3 Strengths

- Full local deployment: no reliance on OpenAI or external APIs.
- Modularity: Easily replace LLM, embeddings, or retriever.
- Usability: Drag-and-drop uploads and natural question interface.

4.4 Limitations

- No conversational memory—stateless interaction.
- No OCR support (scanned PDFs not handled).
- Context length limited by model input size.

5 Conclusion

This project demonstrates a lightweight yet powerful RAG-based medical assistant built with Streamlit, LangChain, and Ollama. By combining document retrieval with local language generation, it allows users to interactively query custom datasets. The system shows how open-source tools can be combined to create explainable, document-grounded AI systems for healthcare.

6 References

References

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al., ”Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”, NeurIPS 2020.

[2] LangChain documentation. Available: `https://docs.langchain.com`

[3] Ollama - Run open-source LLMs locally. Available: `https://ollama.com`

[4] Chroma - The AI-native open-source vector database. Available: `https://www.trychroma.com`

[5] Streamlit Documentation. Available: `https://docs.streamlit.io`