

Mini LLM-Powered Question Answering System Using RAG

Problem Statement

In healthcare, legal, and compliance-heavy industries, professionals must extract specific information from large, dense PDF documents like diagnostic codes, clinical reports, and regulatory guidelines. Manual searching is inefficient and error-prone. This project addresses this by creating a lightweight, retrieval-augmented generation (RAG) based question-answering system using large language models (LLMs).

Step-by-Step Implementation

The project starts by installing some important tools like PyMuPDF (to read PDFs), FAISS (to search similar text), LangChain (to help build the QA system), and Gradio (to make a simple web interface).

Next, we bring in useful components:

- **PyMuPDF** helps us pull out text from PDF files.
- **FAISS** helps find pieces of text that are similar to a user's question.
- **SentenceTransformers** turns text into numbers (embeddings) that the computer can understand.
- **Transformers** gives us a language model (FLAN-T5) that can answer questions.
- **LangChain** helps us connect all the parts together.
- **Gradio** is used to build a user-friendly webpage for uploading PDFs and asking questions.

We use two main models:

- **all-MiniLM-L6-v2** to convert text into embeddings (vectors).
- **google/flan-t5-small** to generate answers based on the user's question and related text.

When a user uploads a PDF, the program reads all the text from the file. Then, it breaks that text into smaller overlapping parts, so it can better understand what each part means.

Each of those parts is turned into vectors using the embedding model and saved in a special searchable format called a FAISS index.

When the user asks a question, the system turns the question into a vector too, and then searches for the most similar pieces of text from the PDF.

Finally, it sends those pieces, along with the question, to the FLAN-T5 model to generate a clear answer using a method called **Retrieval-Augmented Generation (RAG)**.

Tools & Models Used

Tool/Library	Purpose
PyMuPDF (fitz)	PDF text extraction
FAISS	Vector similarity search
SentenceTransformers	Embedding generation
Transformers (Hugging Face)	LLM model (FLAN-T5-small)
LangChain	Modular LLM chaining (RAG integration)
Gradio	Web UI deployment

Use of AI Tools

- **ChatGPT:** Generated and debugged initial boilerplate code, FAISS logic
- **Hugging Face:** Provided pretrained models (FLAN-T5) and tokenizer
- **LangChain:** Integrated embedding retrieval and prompt handling
- The Gradio interface shown in your screenshot is a simple **Mini LLM QA system using RAG (Retrieval-Augmented Generation)**

The screenshot shows a web application titled "Mini LLM QA using RAG". The interface includes a file upload section with a text input showing a file path, a "Drop File Here" area with an upload icon, and a "Click to Upload" link. Below this is a text input for a query: "Give me the correct coded classification for the following diagnosis?". At the bottom are "Clear" and "Submit" buttons. On the right, there is an "output" text area and a "Flag" button. A vertical scrollbar is visible on the right side of the interface.

Sample Output

Mini LLM QA using RAG

Upload a PDF document and ask questions about its content

/content/9241544228_eng.pdf

9241544228_eng.pdf

3.7 MB ↓

Give me the correct coded classification for the following diagnosis?

Recurrent depressive disorder, currently in remission

Clear

Submit

output

Diagnostic guidelines for a definite diagnosis: (a) the criteria for recurrent depressive disorder (F33. -) should have been fulfilled in the past, but the current state should not fulfil

Flag

Mini LLM QA using RAG

Upload a PDF document and ask questions about its content

/content/9241544228_eng.pdf

9241544228_eng.pdf

3.7 MB ↓

Give me the correct coded classification for the following diagnosis?

What disorders are included under mood [affective] disorders?

Clear

Submit

output

a change in mood or affect, usually accompanied by a change in the overall level of activity. The persistent affective disorders are classified here rather than with the personality disorders because of evidence from family studies that they are genetically related to the mood disorders, and because they are sometimes amenable to the

Flag

Mini LLM QA using RAG

Upload a PDF document and ask questions about its content

/content/9241544228_eng.pdf

9241544228_eng.pdf

3.7 MB ↓

Give me the correct coded classification for the following diagnosis?

Explain the features of Somatization Disorder

Clear

Submit

output

body dysmorphic disorder dysmorphophobia (nondelusional) hypochondriacal neurosis hypochondriasis nosophobia Differential diagnosis. Differential diagnosis. Differentiation from the following disorders is essential: 165 MENTAL AND BEHAVIOURAL DISORDERS Somatization disorder. Emphasis is on the presence of the disorder itself and its future consequences, rather than on the individual symptoms as in somatization disorder. In hypochondriacal disorder, Includes: psychogenic confusion twilight state F44.9 Dissociative [conversion] disorder, unspecified mZikm Somatoform disorders

Flag

Conclusion

This mini-project showcases an efficient and lightweight RAG-based QA system using Python and open-source models. It supports real-time semantic search and LLM-based question answering for user-uploaded documents. With future enhancements such as larger LLMs and multi-document indexing, it can evolve into a production-ready enterprise-grade search assistant.

Google Cloab Link

https://colab.research.google.com/drive/1YWxaIsW0rHUUIYlmA0Hcxn3SIDderUdS?usp=s_haring