



DocBot: A Multi-Modal Question-Answering System for PDF's



-Aridaman
-Vayun
-Achal
-Rishabh
-Chetan



Problem Statement

The DocBot challenge aims to develop an intelligent chatbot-based assistance system capable of efficiently processing documents containing both text and images. Participants are required to create a user-friendly interface allowing users to input documents and interact with the chatbot to seek information related to the document content. Upon receiving queries, the chatbot should retrieve or generate meaningful textual answers along with the most relevant image(s) from the document, specifying the corresponding page number(s) where they are located.

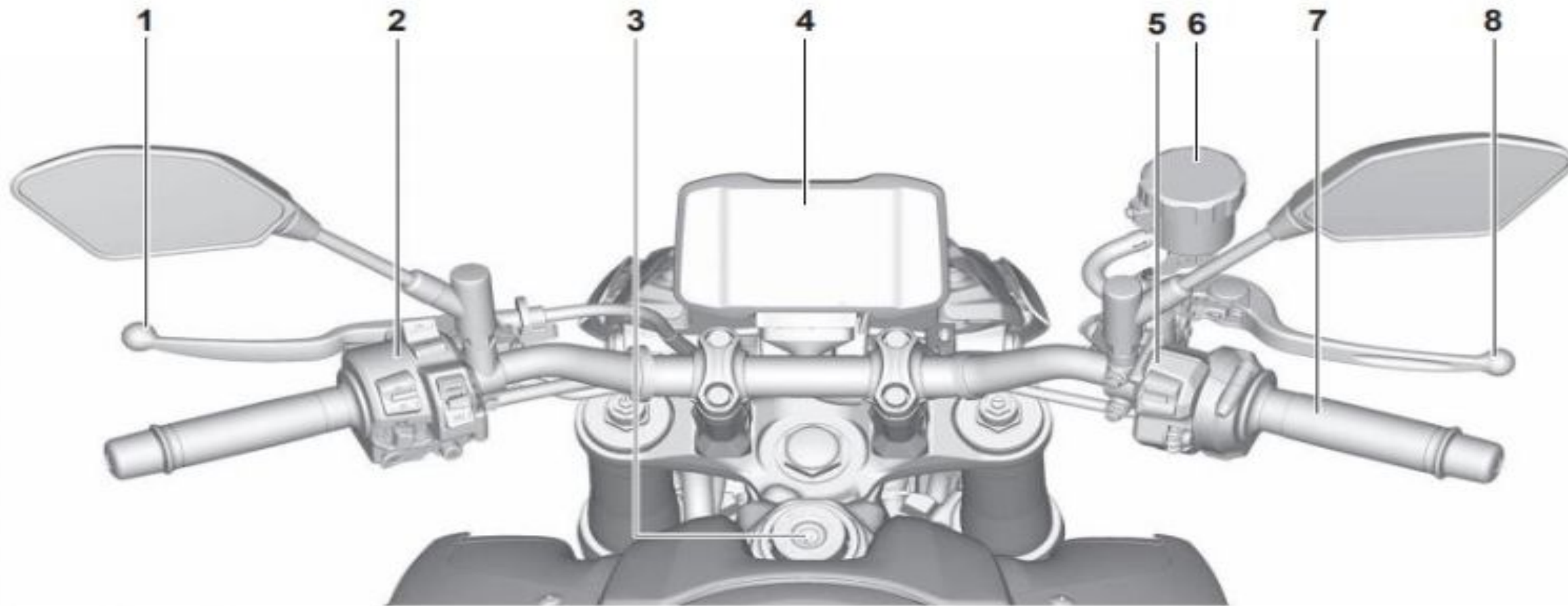
Prompt: Where is the clutch Lever of bike?

Reply:

The clutch lever is located at the left-hand side of the steering. Refer to page 13.

Controls and instruments

EAU10431



1. Clutch lever (page 4-18)
2. Left handlebar switches (page 4-3)
3. Main switch/steering lock (page 4-2)
4. Instrument panel (page 4-5, 4-9)
5. Right handlebar switches (page 4-3)
6. Front brake fluid reservoir (page 7-19)
7. Throttle grip (page 7-23)
8. Brake lever (page 4-19)



APPROACH

Multi-modal RAG

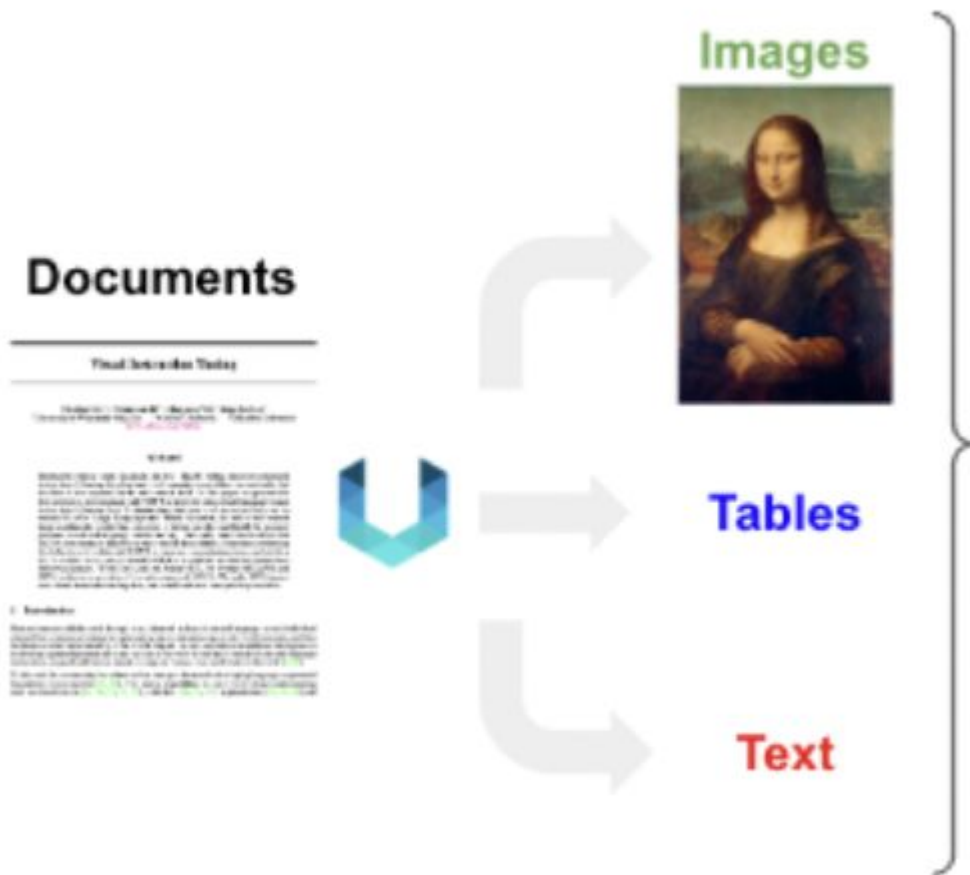
Many documents contain a mixture of content types, including text and images.

Yet, information captured in images is lost in most RAG applications.

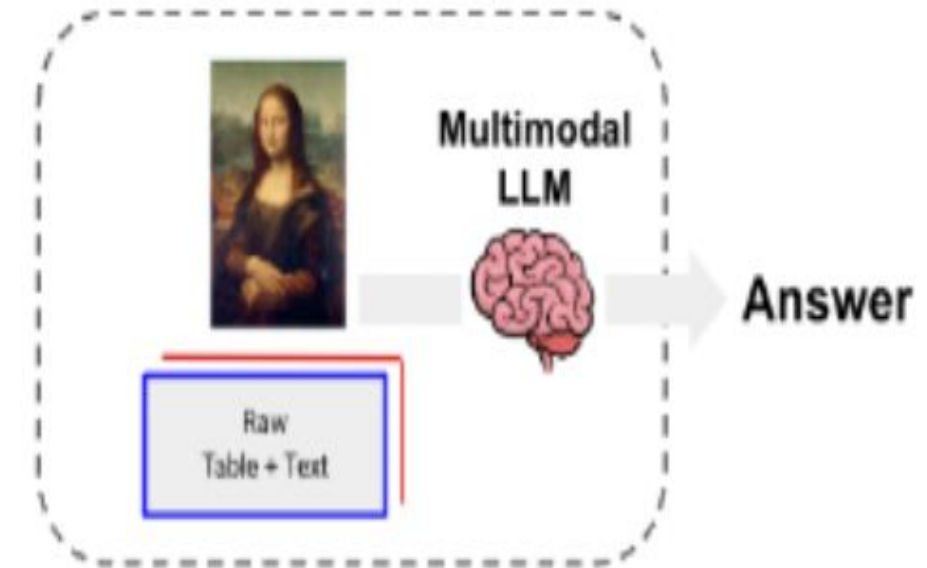
With the emergence of multimodal LLMs, like LLaVa, it is worth considering how to utilize images in RAG:

By leveraging Retrieval Augmented System and LangChain Framework, we can enhance user experience and accessibility.

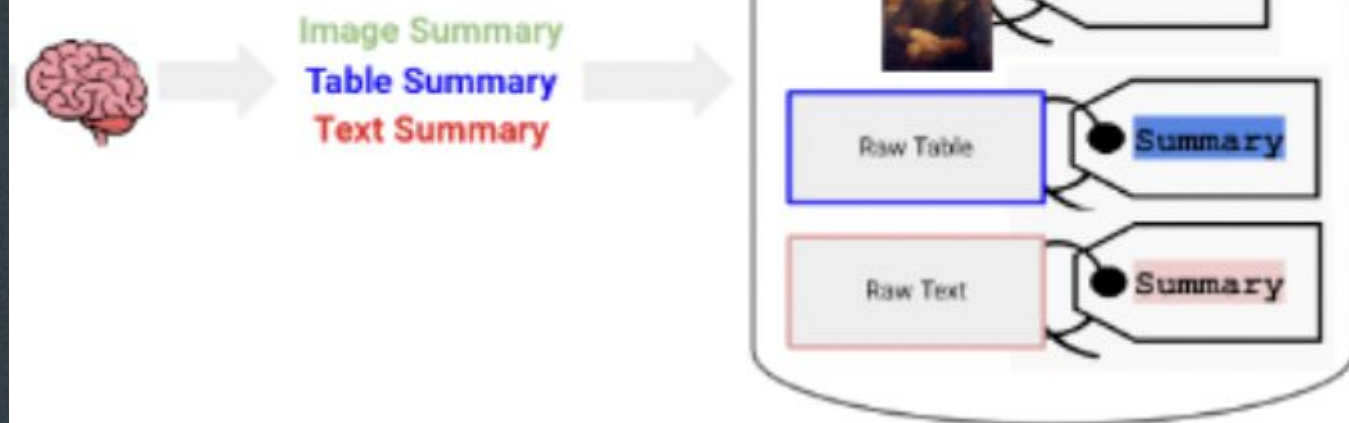




Retrieve image summary but pass raw image to LLM for synthesis



Text Embedding





PIPELINE



What we will be Using / Doing

- Use a multimodal LLM (such as Claude Opus 3 and LLaVA) to produce text summaries from images
 - Embed and retrieve image summaries with a reference to the raw image
 - Pass raw images and text chunks to a multimodal LLM for answer synthesis
 - We will use Unstructured to parse images, text, and tables from documents (PDFs).
- We have used OpenClip embeddings for extracting image embeddings from the image space
- We will use the multi-vector retriever with Chroma/Mongo to load/store raw text, tables and images along with their summaries for retrieval.





The PDF partitioning used by Unstructured will use:

- tesseract for Optical Character Recognition (OCR)
- poppler for PDF rendering and processing
- **Add to vectorstore**
- Add raw docs and doc summaries to Multi Vector Retriever.
- Store the raw texts and tables in the docstore.
- Store the table summaries in the vectorstore for semantic retrieval.
- Use the tables in answer synthesis.
- Store the images in the docstore.
- Store the image summaries in the vectorstore for semantic retrieval.
- Using the image in answer synthesis with a multimodal LLM.
- **Check retrieval**
- The mult-vector retriever will return base64-encoded images or text documents.
- Confirm we can get images back based on natural language search.



Retrieval Augmented System



We format the inputs using a `RunnableLambda`.

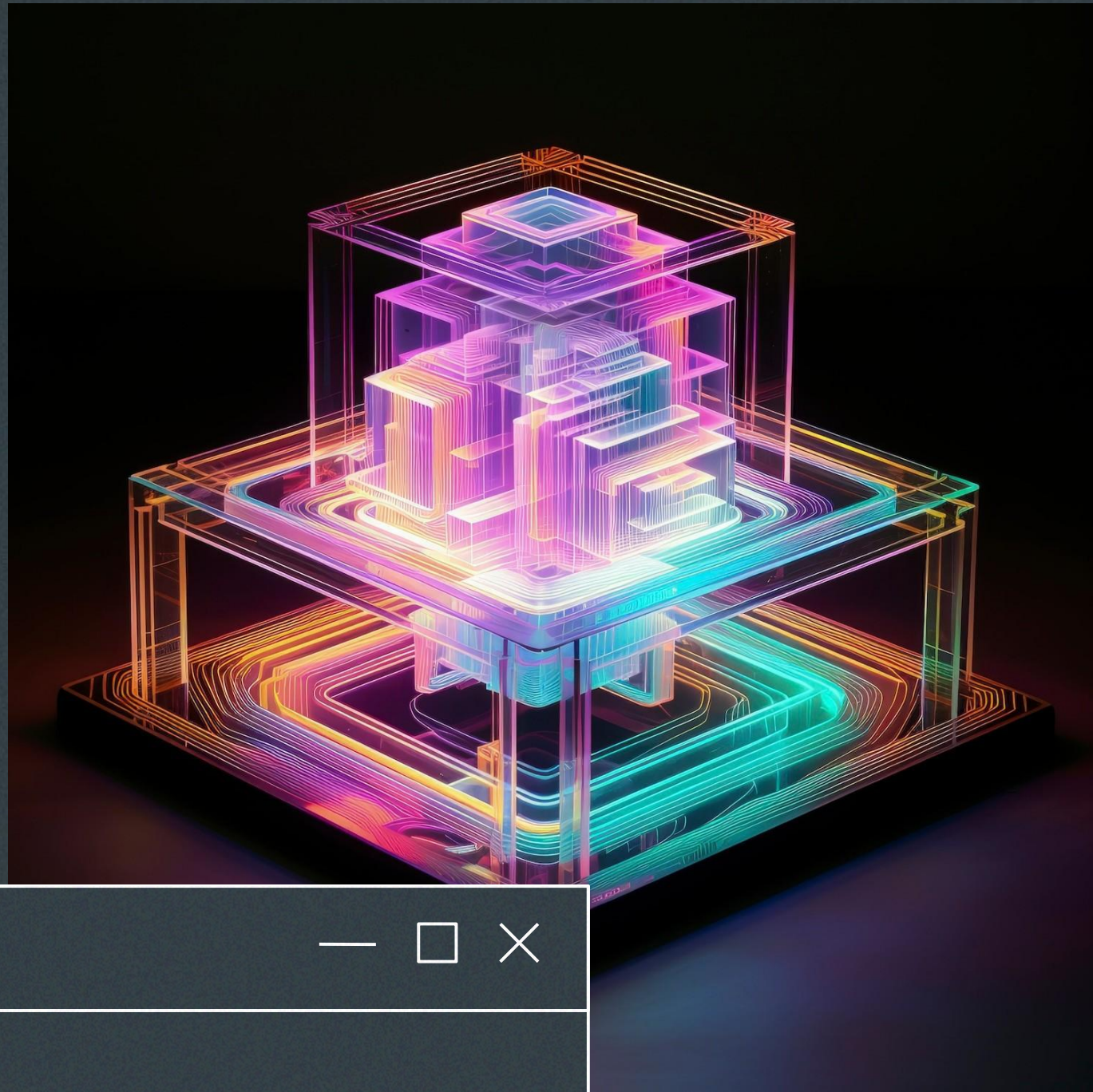
`Runnable` follows the classic RAG flow:

- We first compute the context (both "texts" and "images") and the question (just a `RunnablePassthrough` here)
- Then we pass this into our prompt template, which is a custom function that formats the message for the Claude Opus 3 model.
- And finally we parse the output as a string.



Multi-Functionality

- Capable of handling complex composite queries across various PDF's.
- Provides output on Chain of Thought basis which is visible to user.
- Can provide query-relevant images/tables.
- Our model is capable of handling inputs/outputs both in japanese and english which can be easily toggled.
- Can dynamically create tools for different PDF's.
- Incorporated with various LLM tools such as math, human tool, etc.
- Text-to-voice and voice-to-text.





New Chat

Past Chats

American Wildfires

Yamaha AI Hackathon



Give a summary of American wildfires



In 2018, the total wildfire acreage burned was 8.8 million acres. Of this, 4.6 million acres (52%) burned on federal lands, with 2.3 million acres each on Forest Service and Department of Interior lands. The remaining 4.1 million acres (47%) burned on nonfederal lands.

Some key disaster management strategies for wildfires include:

1. Building community resilience through training, partnerships and capacity building
2. Reducing vulnerabilities of at-risk communities through microfinance, retrofitting buildings, and stockpiling reconstruction materials

Enter the Prompt





Thanks!

