Multimodal RAG System

1. What is RAG?

2. Architecture of RAG.

3. 3 main component of RAG.

4. What is MultiModel RAG?

5. MultiModel Embedding and MultiModel Generation

6. Different Approaches to solve this MultiModel RAG.

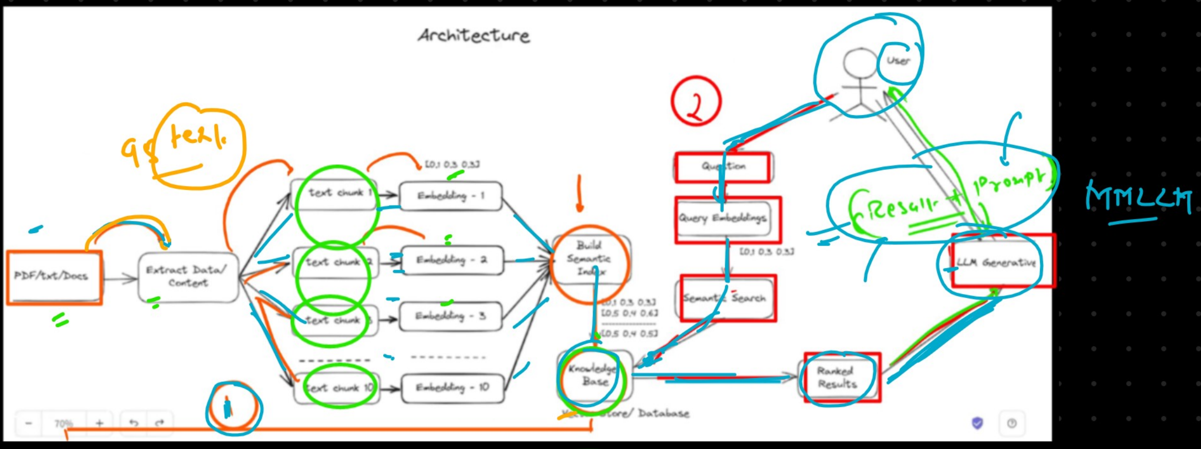
7. Use case of MultiModel RAG

8. Framework for Building MultiModel RAG

1. What is RAG(Retrieval-Augmented Generation)?

(RAG is the process of optimizing the output of a large language model. RAG process references a knowledge base outside of its training data sources before generating a response.)

2. Architecture of RAG.



Data

Doc → text img

MME

=

LanceDB
=
Weviate

MMR

MultiModel RAG

3. 3 main component of RAG.

A) Ingestion
B) Retrieval
C) Generation or Synthesis

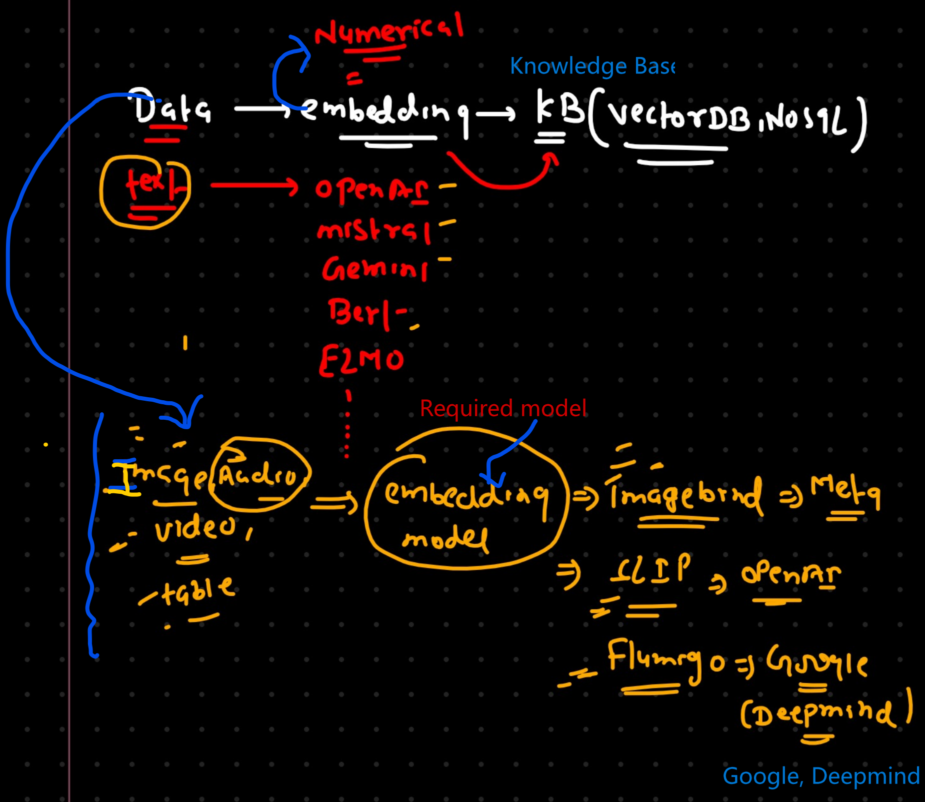
NOW Finally We are Start MultiModel RAG Syst

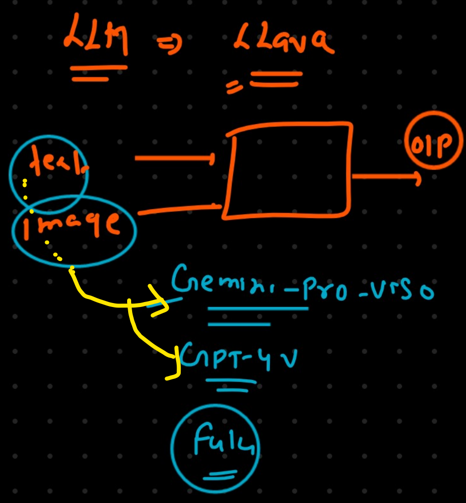
4. What is MultiModel RAG?

Multimodal RAG is AI systems that can understand different modalities like images, audio, video, and text.

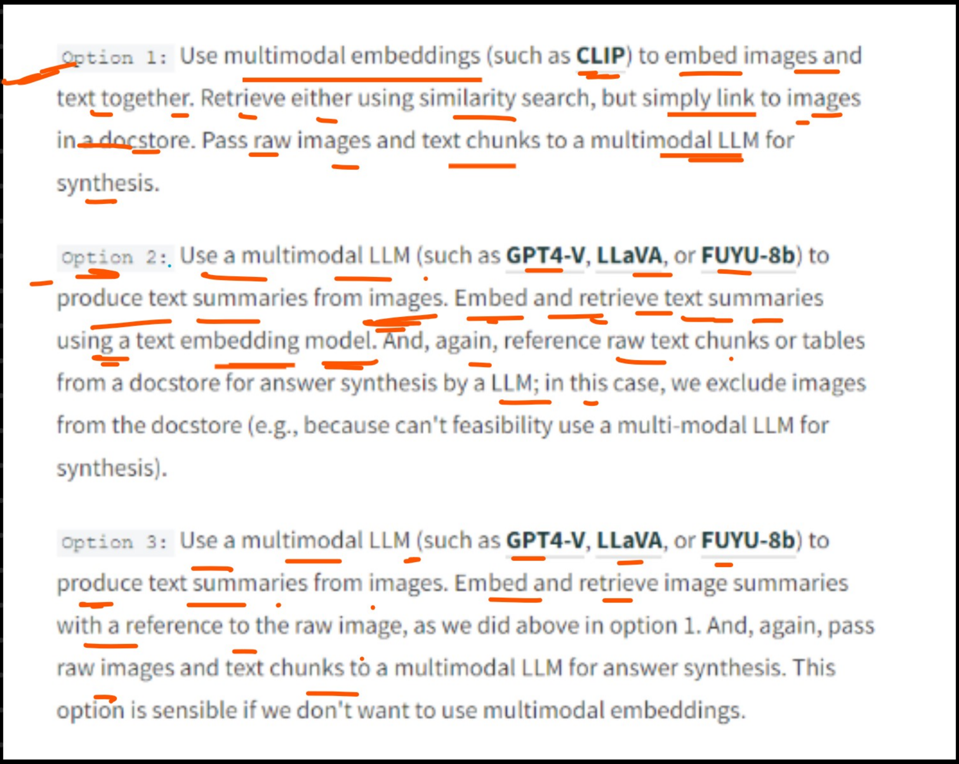
It enables AI systems to process and integrate data from numerous modalities, such as text, images, audio, and video.

5. MultiModel Embedding and Generation





6. Different Approaches to solve this MultiModel RAG.



`Option 1:` Use multimodal embeddings (such as **CLIP**) to embed images and text together. Retrieve either using similarity search, but simply link to images in a docstore. Pass raw images and text chunks to a multimodal LLM for synthesis.

`Option 2:` Use a multimodal LLM (such as **GPT4-V**, **LLaVA**, or **FUYU-8b**) to produce text summaries from images. Embed and retrieve text summaries using a text embedding model. And, again, reference raw text chunks or tables from a docstore for answer synthesis by a LLM; in this case, we exclude images from the docstore (e.g., because can't feasibility use a multi-modal LLM for synthesis).

`Option 3:` Use a multimodal LLM (such as **GPT4-V**, **LLaVA**, or **FUYU-8b**) to produce text summaries from images. Embed and retrieve image summaries with a reference to the raw image, as we did above in option 1. And, again, pass raw images and text chunks to a multimodal LLM for answer synthesis. This option is sensible if we don't want to use multimodal embeddings.

7. Use Case of MultiModel RAG.

1 Video Processing

2 Doc-Processing (text, img, table)

3 Image.captioning

4 Image. Resoning

8. Frame Work to Build MultiModel RAG.

A) Llama Index

B) Langchain