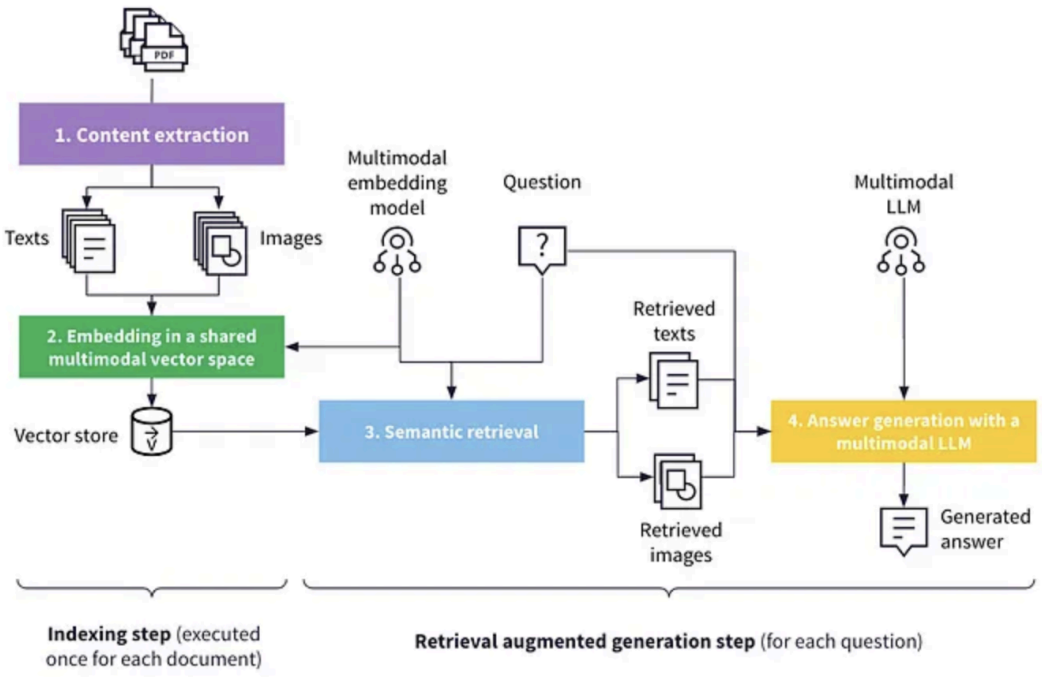
Task: Multimodal RAG application

Goal:
Implement a multimodal query processing workflow that handles both text and image inputs, without relying on OCR for image content extraction. Utilise a multimodal large language model (LLM) to process retrieved text and images for generating responses.

Objective:

- Build a complete pipeline following the given workflow that supports multimodal content retrieval and generation.
- Extract both text and image content from documents and embed them into a shared multimodal vector space.
- Perform semantic retrieval based on user queries to fetch relevant text and image data.
- Feed retrieved images and texts into a multimodal LLM that generates an answer based on both types of input.
- Avoid using OCR for image processing; instead, the images should be processed directly as part of the multimodal input to the LLM.
- The pipeline should be robust and capable of handling diverse user queries, reflecting the multimodal approach in every step.



Evaluation Criteria:

Pipeline Architecture:

- Follow the workflow, using multimodal embedding models and retrieval.

Multimodal Retrieval and Generation:

- Ensure accurate retrieval of text and images and integrate them into the LLM.

Error Handling & Feedback:

- Handle missing data gracefully and include a feedback loop for improving results.

Code Quality:

- Code should be readable, modular, and well-documented.

Deliverables:

GitHub Repository:

- Full code with a clear README on setup and usage.

Explanation Video:

- A 5-10 minute video explaining the workflow, key decisions, and a live demo.