LLM, RAG – Chosen Components

Framework – We have decided to go with LangChain because it brings multiple tools together to create a broader framework. It is also suitable for building intelligent agents capable of performing multiple tasks simultaneously.

Text Splitter – We have decided to go with recursive chunking because it divides chunks of text in an iterative manner. After the first iteration, if we are not satisfied with the chunk size, the function will repeatedly call itself on the rest of the chunks with a different separator until the desired chunk size is achieved.

Embedding Model – We have chosen mxbai-embed-large since it is a completely free and open-source model that outperforms closed-source, costlier models such as OpenAI's own embedding model while also using significantly less memory than better performing embedding models

Vector Store – We have chosen Facebook AI Similarity Search (FAISS) since it is retrieves faster than other vector stores without hurting its accuracy and is able to search multimedia documents and match them which SQL-based vector stores can't do

LLM – We have picked Llama 3 as it is completely free and open-source and is easier to integrate into our pipeline than GPT4 or Gemini. allows developers to modify and integrate the model into unique use cases like language translation tools and efficient chatbots

Metadata Extractor - We decided to go with Doctran because of its flexibility with different types of metadata, and we have different types of files.