Weekly Plan

Week 1: Text Extraction, Image Text Extraction, Tabular Data Extraction, Vectorization

1. Consider different methods for text extraction depending on the format of the original data. For image text extraction, you may need to use Optical Character Recognition (OCR) tools (Common pdf image based text extraction packages will not work with some image based pdfs). For tabular data extraction, check the structure of the data to ensure it is correctly extracted. Also, the vectorisation process must take into account the dimensionality (angular or any other dimenstion) of your data. Try to use index based vectorisation to improve latency.

2. Automate the program such that each PDF is transformed into its respective vector space representation, named after the original PDF. If an additional PDF is introduced, the system should automatically convert this specific PDF into a corresponding vector representation.

3. Select a LLM model - (Try to create a sample question answering chatbot).

Week 2: Parallel Processing, Asynchronous Processing, Caching - Latency

1. When implementing parallel and asynchronous processing, ensure the tasks are independent to avoid any data inconsistency issues. For caching, decide on what data or results are important to cache based on their retrieval time and usage frequency.

2. Try to start with a small pdf based chatbot using chunking - Understand how much time it takes to respond

Week 3: Vectorization for Questions, Similarity Search - Find the Best Similarity Search Method

1. When vectorising questions, consider the method that best captures the semantics of the question. For similarity search, experiment with different methods such as cosine similarity or Jaccard index to see which works best with your data.

2. Check with different dimensional vector storage and index based vector storage and computational time.

Week 4: Selection of Models - Fix LLM model and show the answer with the source

1. Based on your knowledge from 1st and 2nd week (Based on the observation of tested LLM models and sample pdf based chatbot you tried in the 1st and 2nd week) fix a LLM model.

2. And pass the questions and best matched paragragh to get a absolute answer with source (Name of the pdf). Integrate Questions with Best Matched Paragraph (from vectors).

Week 5: Integrate Backend and Frontend

1. Consider using APIs for communication between frontend and backend.

Week 6: Testing and Additional aspects, Modification

1. Testing
2. Check any additional modification is needed
3. Check the latency
4. UI modification - If needed