The system accepts question in a natural language (English) from the end user for the implied knowledge base to answer and resolve query from the horde of pdf documents collected from open sources and has been feed to system, the architecture supports any computer readable pdf document that is feed to it and provides smart answers in natural language. The system provides answer to any query that is part of the knowledge base and does not give outside knowledge that is not part of the knowledge base.

Architecture:

1. Pdf_doc=> |Txt_Extractor| => Raw_Txt + URL to the Images (also embeds the images and associates with the related text)

2. Raw_Text + Image_URL => |Embedding_FAISS_Indexing| => Semantic_Vector_Store

3. User_Query => |LLM| => Summerized_txt_for_Semantic_Search

4. Summerized_TXT => |Vector_Query_Top_K| => Top_Related_Documents

5. Top_Related_Documents => |LLM| => Human_Friendly_Response

6. Evaluation Set creation.

7. Testing for proper knowledge extraction and knowledge base exclusive response extraction

8. Finding the Numerical evaluation matrices for evaluating system capability.

Architectural Features: Supports both online API calls to the LLM(GEMINI) and offline LLM(OLLAMA) for privacy of sensitive documents. We can have multiple different databases simultaneously on the system and each will have options to choose the LLM preference.

User Centric Architecture:

1. Select Database.
2. Select LLM source.
3. Select option to add more document or query the existing knowledge base.
4. If selected query, then natural language response is returned.
5. If selected add document, then gives option to upload documents.

Testing: We utilize personally handcrafted documents containing imaginary information for verification of the fact that the end response is indeed only extracted from the knowledge base no outside information can invade the generated response from the system even after sugar coating the raw semantically similar response returned from the FAISS indexed database by using a SOTA LLM.

Evaluation Matrices: We find the numerical matrices (MRR, Recall@2, Recall@5, ROUGE-1, and BLEU) for evaluating the system performance.