PDFMASTER: LLM-DRIVEN PDF ANALYSIS AND EDITING

Project Reference No.: *48S_BE_2327*

College : Vidyavardhaka College of Engineering, Mysuru
Branch : Computer Science and Engineering
Guide(s) : Dr. Prasad M R
Student(s) : Ms. Bhumika A
Ms. Aishwarya S N
Ms. Ashwini C S
Ms. Bakthieshwari

Keywords:

LLM, OCR, AI, NLP, PDF EDITING

Introduction:

In today's digital era, the volume of documentation generated and consumed is growing rapidly across industries. From research reports and legal contracts to business proposals and manuals, the need to interact with, search through, and edit these documents efficiently has become increasingly important. However, manual handling of PDFs is often time-consuming, prone to errors, and lacks intuitive accessibility.

To address this challenge, our project presents an AI-driven interactive PDF platform that allows users to query, summarize, and edit documents using natural language. The system leverages the power of OpenAI's large language models (LLMs) and integrates with LangChain to handle prompt flow and retrieval. The front-end is developed using Streamlit, offering a simple, chat-style interface for user interaction.

The platform accepts PDF files, processes their contents using NLP techniques, and enables users to interact with them intelligently. Users can ask questions about the content, extract information, or edit the document directly—all in real-time. The integration of vector-based semantic search further enhances the system's ability to understand context and deliver accurate responses.

This project combines AI, NLP, and user interface technologies to transform the way we handle documents, making PDF interaction more efficient, intelligent, and user-friendly.

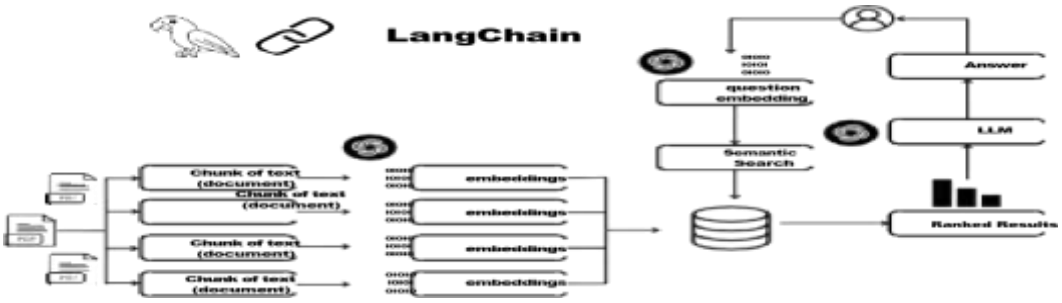

Figure 1: complex documents through natural language instructions.

Objectives:

- Develop a platform for real-time interaction with PDF documents.
- Integrate OpenAI's language models with LangChain for natural language processing.
- Enable querying, summarization, and contextual editing of PDFs.
- Use Streamlit for a clean and interactive user interface.
- Employ Python libraries (e.g., PyPDF2, ReportLab) for dynamic PDF manipulation.
- Automate document workflows and improve content accessibility.
- Add features like PDF splitting, merging, and AI-based suggestions.
- Enhance user experience through seamless, chat-based document editing.

Methodology:

The development of our interactive PDF framework follows a modular and layered methodology combining NLP, document processing, and web-based interaction. The primary components and their roles are described below:

- **System Architecture:** The system is built on a client-server architecture where the frontend is developed using Streamlit and the backend leverages Python with libraries like LangChain, OpenAI's API, PyPDF2, and ReportLab.
- **PDF Ingestion and Parsing:** The uploaded PDFs are read using PyPDF2, allowing extraction of text content, metadata, and structure. This raw text is the basis for further processing.
- **Text Chunking and Embedding:** The extracted content is divided into manageable chunks. These chunks are embedded using OpenAI’s embedding models and stored in a vector database (e.g., FAISS or Chroma) for efficient semantic search and retrieval.

- **Natural Language Querying:** LangChain acts as the orchestration framework, allowing LLMs to interpret user queries, retrieve relevant chunks, and provide summarized or direct responses using OpenAI's GPT models.
- **Dynamic Editing:** For real-time edits, ReportLab is used to create modified PDFs based on user input, allowing actions like text replacement, addition, deletion, or annotation generation.
- **Conversational Interface:** A Streamlit-based UI is built to offer a seamless, chat-like experience for users. Users can upload PDFs, ask questions, or request edits and see results in real-time.
- **Advanced Features(In Progress):** Features like splitting, merging, and AI-based suggestions for document improvement are being developed to extend functionality.

Result and Conclusion:

Our project successfully developed an AI-powered PDF platform that enables real-time querying, summarization, and editing through a user-friendly, chat-based interface. It integrates OpenAI's GPT models with LangChain and Streamlit, providing seamless interaction with PDF documents.

Key Results:

- Efficient extraction and understanding of content from various PDF types.
- Real-time, context-aware answers and dynamic editing features.
- Scalable performance even on large documents.

Observations & Conclusion:

- Natural language queries improve accessibility and user engagement.
- AI-generated suggestions enhance clarity and quality of document content.
- The system demonstrates LLMs' effectiveness in automating document workflows.

Industry Relevance:

- Useful in law, healthcare, finance, and education for streamlining documentation.
- Reduces manual effort and boosts productivity through automation.

Working Model:

- The platform is fully functional and operational, not a simulation or theoretical prototype.

Learnings:

- Gained practical knowledge in LLM orchestration, PDF manipulation, and conversational UI design.
- Understood modular design principles and semantic search using vector databases.

Future Scope:

The future scope of this project includes:

- **Multilingual Support :** Expanding language model integration to support document querying and summarization in multiple languages.
- **Voice-Enabled Interaction:** Incorporating speech-to-text and text-to-speech capabilities for voice-based document navigation and editing.
- **Collaborative Editing :** Adding real-time collaboration features, allowing multiple users to interact with and edit documents simultaneously.
- **Advanced Editing Tools :** Integrating advanced editing options such as font formatting, image insertion, and digital signature support.
- **Mobile Compatibility :** Developing a mobile-friendly version of the platform for easy document interaction on smartphones and tablets.

These enhancements will broaden the project's utility across industries and create a comprehensive, intelligent document handling ecosystem.