# AUTOMATING PDF INTERACTION USING LANGCHAIN AND CHATGPT

***Project Reference No.:*** *47S_BE_1393*

***College*** ***:*** *K.L.E Society's K.L.E. College of Engineering and Technology, Chikodi*
***Branch*** ***:*** *Department of Computer Science and Engineering*
***Guide(s)*** ***:*** *Prof. Shilpa B. Hosagoudra*
***Student(S)*** ***:*** *Mr. Rushikesh Joshi*
*Mr. Sarthak Shrikant Tavadar*
*Mr. Nitin Patil*
*Mr. Sharanbasu Kalburgi*

**Keywords:**

PDF chatbot, LangChain, Pinecone, LLM Model, natural language processing, large language model

**Introduction of project**:

PDFs are a popular format for storing documents. They are often used in business, education, and research. However, PDFs can be difficult to read and understand. This is because they are not designed to be interactive. Chatbots are a type of artificial intelligence (AI) that can be used to interact with users in a natural way. They can be used to provide information, answer questions, and complete tasks.

Here we demonstrate a method that used the LangChain and LLM Model to create a PDF chatbot. A framework called LangChain makes it simpler to create chatbots and scalable AI/LLM applications. The LLM Model is a huge language model that may be used to create text, translate across languages, create various types of creative material, and provide user with helpful answers to user inquiries. The Methodology used Pinecone to store the vectors of the PDF files. Pinecone is a vector store for storing embeddings and user PDF in text to later retrieve similar docs. React JS was used for the front end to develop a webpage to interact with the chatbot.

The chatbot was able to answer questions about the PDF files, and it was also able to generate text that was similar to the text in the PDF files. The chatbot was tested on a variety of PDF files, and it was able to achieve a high accuracy rate.

**The proposed system is implemented using the following technologies:**

**LangChain:** A framework for building scalable AI/LLM apps and chatbots.

**Large Language Model (LLM):** A large language model that can be used to generate text, translate languages, write various types of creative material, and provide informative answers to user queries.

**Pinecone:** Its vector repository for embeddings and user PDF in text so use may access related documents later.

**Scope and Objectives of the project:**

**Scope:**

The LangChain project introduces a Language Learning Module (LLM) to redefine how individuals and organizations interact with PDF documents. The primary focus is on integrating an efficient tool that not only manages and manipulates PDF content but also facilitates language learning within the document context. By leveraging advanced natural language processing, LangChain aims to empower users to extract, analyze, and learn from PDFs through intuitive language-driven commands. The project's multifaceted approach enhances document management, boosts productivity, and creates a unique learning experience, making LangChain a versatile tool across diverse sectors.

**Objectives:**

1) **Develop and Integrate LLM:**
2) **Enable Language-Driven Commands:**
3) **Enhance PDF Content Management:**
4) **Faster Language Learning Experience:**
5) **Ensure Versatility and Productivity:**

**Methodology:**

1. **Data Gathering and Preparation:**
    - Collect diverse PDFs from various domains.
    - Structure text into paragraphs, tables, figures (chunks).
    - Annotate documents with relevant tags (person, concepts).
    - Clean and normalize text.

2. **Question Answering and Chatbot:**
    - Train a question-answering system using annotated data and embeddings.
    - Develop a conversational chatbot interface to interact with PDFs.
    - Process user queries, match them to relevant document embeddings, and retrieve answers.
3. **LangChain Integration:**
    - Design and implement LangChain architecture for knowledge management.
    - Enable knowledge reasoning and inference.
4. **Evaluation and Deployment:**
    - Evaluate NLP and question answering performance.
    - Gather user feedback and iterate.
    - Analyse LangChain performance and optimize.
    - Develop user interface and APIs for integration.

**Future Scope:**

The future scope for automating PDF interaction and integrating it with ChatGPT is vast, offering numerous possibilities across various industries. Here are some potential developments and applications:

**1. Enhanced Data Extraction and Analysis**

- Advanced OCR and NLP Techniques: Leveraging advanced Optical Character Recognition (OCR) combined with Natural Language Processing (NLP) to accurately extract data from complex PDF layouts, including tables, charts, and handwritten notes.
- Contextual Understanding: Developing models that not only extract data but also understand the context, enabling more accurate and meaningful information retrieval.

**2. Intelligent Document Summarization**

- Automated Summarization: Creating tools that can summarize lengthy documents into concise and relevant abstracts, making it easier for users to quickly grasp key points.
- Custom Summaries: Allowing users to request summaries based on specific sections or topics within a document.

**3. Interactive Document Search and Query**

- Semantic Search: Implementing search capabilities that understand the semantics of queries, providing more relevant results even with vague or complex questions.
- Voice and Chat Interfaces: Enabling users to interact with PDFs through voice commands or chat interfaces, asking questions and getting answers in real-time.

**4. Automated Compliance and Auditing**

- Regulatory Compliance: Automating the process of checking documents for compliance with industry regulations, highlighting areas that need attention.
- Audit Trails: Creating detailed audit trails of interactions and modifications made to documents for better tracking and accountability.

**5. Personalized Learning and Knowledge Management**

- Custom Study Guides: Generating personalized study guides or training materials from academic papers, textbooks, or corporate documents.
- Knowledge Bases:*Building intelligent knowledge bases that can automatically update and organize information extracted from new documents.

**6. Enhanced Collaboration and Workflow Automation**

- Collaborative Editing: Allowing multiple users to collaboratively edit and annotate PDFs in real-time, with ChatGPT assisting in suggestions and corrections.
- Workflow Integration: Integrating with enterprise workflow systems to automate

document handling processes, such as approvals, routing, and digital signatures.

**7. Improved Accessibility**

**-** Assistive Technologies: Developing tools that convert PDFs into accessible formats for individuals with disabilities, including audio descriptions and Braille conversions.

- Language Translation: Providing real-time translation of PDF content into multiple languages, making documents accessible to a global audience.

**8. Security and Privacy**

**-** Data Protection: Implementing robust security measures to protect sensitive information extracted from PDFs, ensuring compliance with data privacy regulations.

- Access Controls: Providing fine-grained access controls and encryption to secure document interactions.

**9. Industry-Specific Applications**

**-** Healthcare: Automating the extraction and analysis of patient records, medical research papers, and clinical trial data.

- Legal: Streamlining the review of legal documents, contracts, and case files, providing quick summaries and key insights.

- Finance: Enhancing the automation of financial reports, invoices, and compliance documents, enabling faster and more accurate processing.

**Integration with ChatGPT**

**-** Natural Interaction: Combining ChatGPT's conversational abilities with PDF automation to create a seamless user experience, where users can interact naturally with documents.

- Learning and Adaptation: Continuously improving the system by learning from user interactions, making it more intuitive and efficient over time.

**Future Research and Development**

**-** AI and ML Advances: Continuing research in AI and machine learning to improve the accuracy, efficiency, and capabilities of automated PDF interaction tools.

- User Feedback: Incorporating user feedback to refine and enhance features, ensuring the tools meet evolving needs and preferences.

By advancing these areas, the integration of automated PDF interaction with ChatGPT can significantly enhance productivity, accessibility, and efficiency across numerous domains, paving the way for smarter, more responsive digital document management systems.

**Result and Outcome of the project:**

In conclusion, the development of the automated PDF interaction system represents a significant step forward in the realm of document management and user interaction. Through rigorous testing and validation, we have demonstrated the system's ability to efficiently extract text from PDF files, process user queries, and provide precise responses, thereby enhancing productivity and user experience. Despite encountering some challenges during the development and testing phases, such as accuracy limitations and performance bottlenecks, the system has showcased notable strengths in functionality, scalability, and adaptability. Overall, the project has laid a solid foundation for future advancements in AI-driven document management solutions, with the potential to revolutionize how users interact with PDF files across various domains and industries.