e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:02/February-2025 Impact Factor- 8.187 www.irjmets.com

AI-POWERED DOCUMENT PARSING FOR REAL-TIME PDF KNOWLEDGE BASE INTEGRATION

Mrs. K Swathi*[1], D. Deepthi Reddy*[2], K. Mahesh Babu*[3], M. Samuel*[4], S. Sai Ram*[5]

*[1]Internal Guide, ACE Engineering College of Computer Science and Engineering, Ghatkesar, Telangana, India.

*[2,3,4,5]Student, ACE Engineering College of Computer Science and Engineering, Ghatkesar, Telangana, India.

DOI: https://www.doi.org/10.56726/IRJMETS67101

ABSTRACT

This explores the development of an advanced AI-powered document parsing system designed to revolutionize the way static PDF documents are processed and utilized. Existing solutions like Adobe Acrobat and Mendeley, while functional, lack the contextual understanding and semantic capabilities necessary for efficient information retrieval, particularly in academic, corporate, and research-intensive environments. The proposed system, leverages cutting-edge AI technologies, including Vector Databases (Vector DB) and OpenAI's semantic search engine, to transform traditional PDF documents into intelligent, searchable knowledge bases. The architecture focuses on three main processes: document ingestion, contextual information extraction, and semantic retrieval. Users can query the system in natural language, receiving accurate and context-aware results, even when the exact keywords are not present in the document.

By integrating advanced algorithms for vector embeddings and utilizing AI for contextual understanding, It addresses the limitations of keyword-based systems, offering improved precision, reduced search times, and enhanced user satisfaction. This paper presents the system's design, implementation, and evaluation, demonstrating its potential to set a new benchmark in document management and retrieval systems. Future enhancements, including multi-language support and real-time collaboration, are also discussed, highlighting the system's scalability and adaptability.

I. INTRODUCTION

The extensive information contained within PDF documents often creates challenges in retrieving relevant knowledge efficiently. Whether for academic, corporate, or personal purposes, traditional tools such as Adobe Acrobat and Mendeley frequently fail to deliver context-aware results, relying instead on basic keyword-based searches that are limited in scope and precision. These methods struggle with queries that require understanding the relationships between terms or when precise keywords are unknown. Such limitations lead to inefficiencies, particularly in environments where time and accuracy are critical.

This paper presents an AI-driven approach to document parsing that employs **Vector Databases** and **semantic search technologies** for precise, contextually-aware information retrieval. The system processes natural language queries to deliver relevant results, even when query terms differ from those used in the document. By overcoming the constraints of traditional systems, this approach improves search accuracy, speeds up retrieval, and enhances user interaction with document collections, making information management more efficient and accessible.

II. PROBLEM STATEMENT

Traditional PDF document management and retrieval systems rely heavily on keyword-based search techniques, which are often inadequate for extracting meaningful and contextually relevant information. Existing solutions, such as Adobe Acrobat and Mendeley, struggle with understanding the relationships between terms, leading to inefficiencies when users are unaware of the exact keywords needed for retrieval. This limitation is particularly problematic in academic, corporate, and research-intensive environments where quick and precise access to information is essential.

Moreover, these conventional systems do not leverage advanced AI-driven techniques, such as vector embeddings and semantic search, resulting in slow retrieval times, low accuracy, and reduced user satisfaction. The absence of contextual understanding makes it difficult to retrieve information effectively, especially when documents contain complex, domain-specific terminology.

To address these challenges, there is a need for an AI-powered document parsing system that transforms static PDF documents into intelligent, searchable knowledge bases. By utilizing Vector Databases and semantic search technologies, such a system can enhance information retrieval, improve precision, and significantly reduce search times, thereby optimizing document management for users across various industries.

III. PROPOSED SYSTEM

The proposed AI-powered document parsing system is designed to enhance, rather than replace, existing document management workflows. This hybrid approach ensures that users accustomed to traditional tools like Adobe Acrobat and Mendeley can continue using familiar interfaces while benefiting from advanced AI-driven search capabilities. By integrating seamlessly with current systems, the solution minimizes disruption and eliminates the need for extensive retraining, making adoption more efficient and user-friendly.

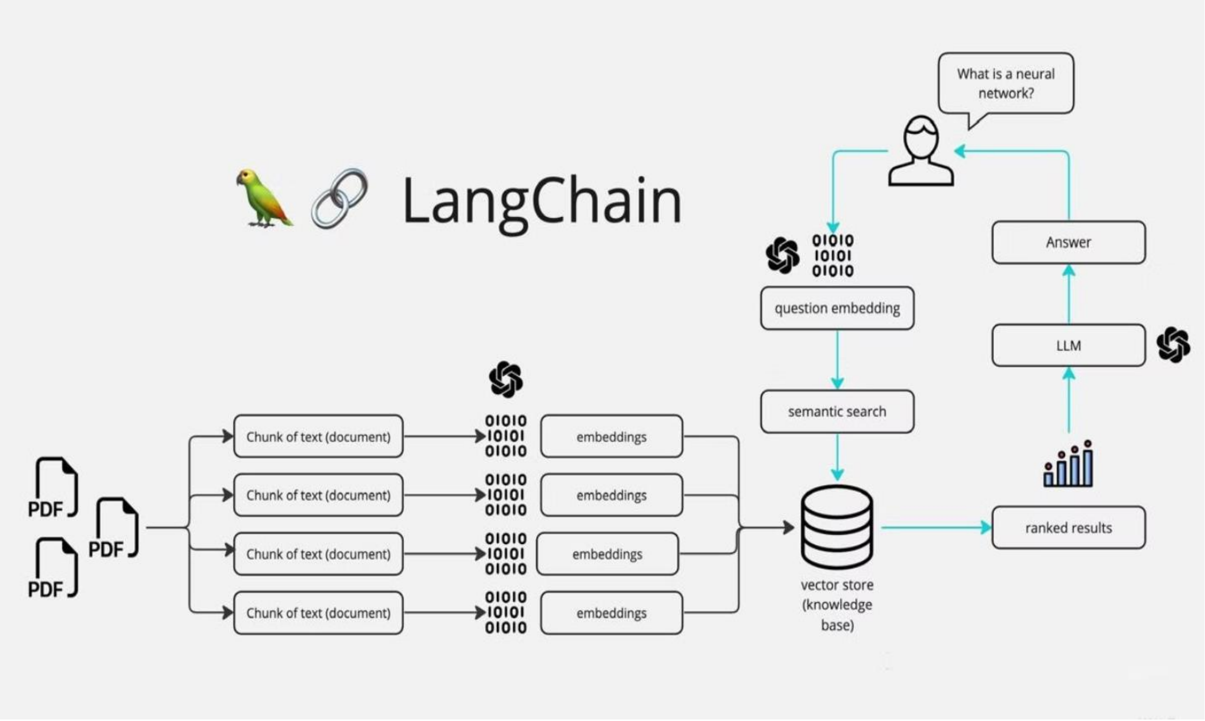
A key limitation of conventional document retrieval methods is their reliance on keyword-based searches, which often fail to capture the contextual meaning of queries. This leads to inaccurate or incomplete results, especially in domains where terminology is complex and varied. The proposed system addresses this challenge by incorporating Vector Databases and semantic search technologies, enabling it to understand the relationships between concepts within a document. As a result, users can enter natural language queries and receive highly relevant results, even if the exact keywords do not appear in the document.

This enhanced search capability significantly improves information retrieval by reducing search times, increasing accuracy, and enhancing user satisfaction. By offering context-aware search and intelligent document parsing, the system transforms static PDFs into dynamic, searchable knowledge bases, optimizing information management across academic, corporate, and research-intensive environments.

IV. SOFTWARE REQUIRMENTS

- Operating System : Windows 7 (Min)
- Front End : Streamlit
- Back End : Python
- Database : Microsoft Excel
- HAM10000 dataset
- Processor : Intel® Pentium® G4560 (Min)
- Speed : 2.9 GHz (Min)
- RAM : 2 GB (Min) Hard Disk : 2 GB (Min).

V. SYSTEM ARCHITECTURE



The architecture depicted in the image represents an AI-powered document retrieval system using **LangChain**, integrating **Vector Databases** and **Large Language Models (LLMs)** for intelligent querying and contextual search. Here's a breakdown of the workflow:

1. Document Ingestion

- **Input:** The system starts with PDF documents as input.
- **Chunking:** The documents are split into smaller chunks of text to facilitate efficient processing and retrieval.

2. Embedding Generation

- Each chunk of text is passed through an embedding model (likely an OpenAI embedding model).
- The embedding model converts the text into numerical vector representations.
- These embeddings capture the semantic meaning of the text rather than just keyword-based information.

3. Vector Store (Knowledge Base)

- The generated embeddings are stored in a **vector database** (knowledge base).
- This vector store enables efficient **semantic search**, meaning it can retrieve relevant content even if the query uses different words than those in the original document.

4. Query Processing

- A user inputs a **natural language query**, such as *"What is a neural network?"*.
- The query is converted into an **embedding representation**, similar to the document embeddings.

5. Semantic Search & Retrieval

- The query embedding is compared against the stored embeddings in the vector database.
- The system retrieves **ranked results** based on semantic similarity rather than just keyword matching.

6. LLM Processing & Answer Generation

- The retrieved ranked results are passed to a **Large Language Model (LLM)**.
- The LLM generates a final **answer** based on the most relevant retrieved documents.
- The user receives an accurate, context-aware response.

VI. OUTPUT

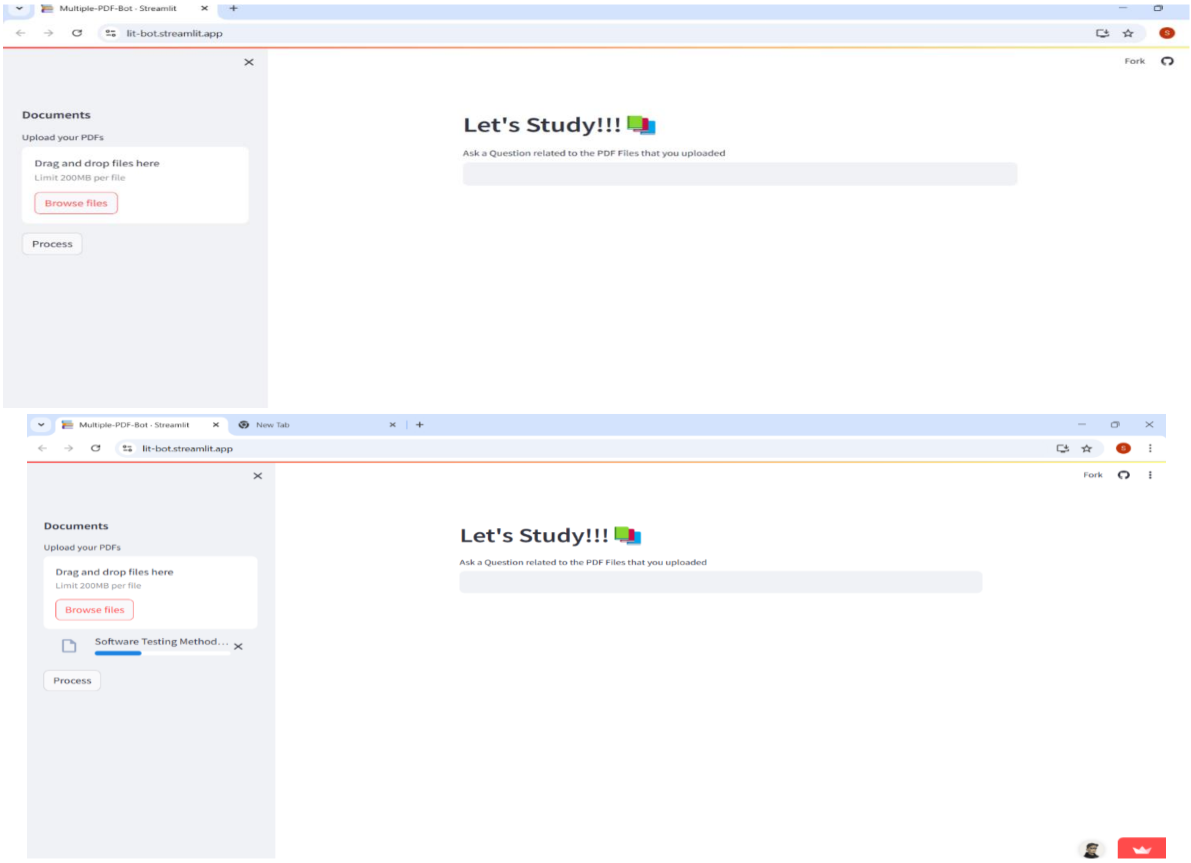
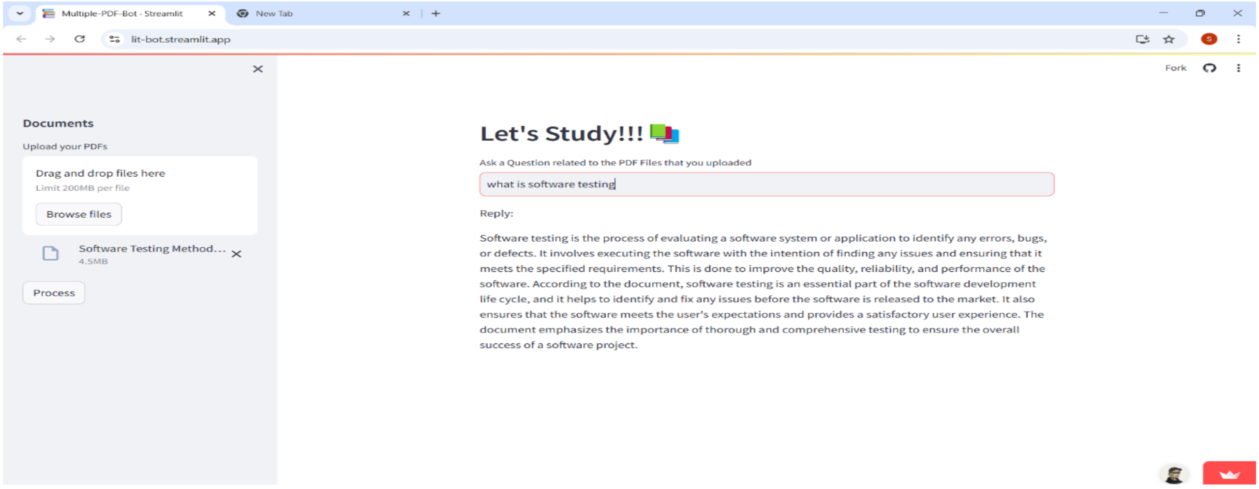


Fig: Analysing the Pdf Document



Let's Study!!!

Ask a Question related to the PDF Files that you uploaded

what is software testing

Reply:

Software testing is the process of evaluating a software system or application to identify any errors, bugs, or defects. It involves executing the software with the intention of finding any issues and ensuring that it meets the specified requirements. This is done to improve the quality, reliability, and performance of the software. According to the document, software testing is an essential part of the software development life cycle, and it helps to identify and fix any issues before the software is released to the market. It also ensures that the software meets the user's expectations and provides a satisfactory user experience. The document emphasizes the importance of thorough and comprehensive testing to ensure the overall success of a software project.

Fig: Generation of Answer for prompt provided

VII. CONCLUSION

The proposed AI-powered document parsing system, integrated with Vector Databases and semantic search technologies, offers a significant improvement over traditional keyword-based search tools. By leveraging the power of text embeddings and semantic search, the system enables users to retrieve information with greater accuracy and contextual relevance, addressing the limitations of conventional systems.

The architecture ensures that the system is scalable and can handle large volumes of documents, making it ideal for applications in academia, research, corporate environments, and personal knowledge management. Its ability to integrate with existing workflows without replacing them entirely ensures that users can adopt the system with minimal disruption.

By facilitating more intuitive and efficient interactions with document repositories, the system has the potential to transform the way individuals and organizations search for and manage information. The proposed work lays the foundation for further advancements in document retrieval, helping users to uncover critical knowledge faster, ultimately enhancing productivity and decision-making.

VIII. REFERENCES

[1] Bojar, O., et al. "Multilingual BERT for Document Retrieval: A Comparative Study." Proceedings of the 27th International Conference on Computational Linguistics (COLING), 2020.

[2] Vaswani, A., et al. "Attention is All You Need." Proceedings of NeurIPS 2017, 2017. doi:10.5555/3295222.3295349.

[3] Rajbhandari, S., et al. "Faiss: A library for efficient similarity search and clustering of dense vectors." Facebook AI Research, 2020. https://github.com/facebookresearch/faiss.

[4] Pinecone Team. "Pinecone: A Vector Database for Scalable Search and Retrieval." Pinecone, 2021. https://www.pinecone.io/.

[5] LangChain Documentation. "LangChain: Framework for building applications with LLMs." Lang Chain Documentation. https://langchain.com/docs/.

[6] Schuster, M., et al. "AI-Powered Semantic Search for Knowledge Retrieval in Large-Scale Document Repositories." Journal of Artificial Intelligence and Data Science, 2022.