

AI-Driven Document Question Answering and Content Retrieval Using Large Language Models and Vector Embeddings

Sukrutha S Bhat¹, Prashanth Ankalkoti²

¹PG Student, Department of MCA, Jawaharlal Nehru New College of Engineering, Shivamogga, Karnataka, India.

²Assistant Professor, Department of MCA, Jawaharlal Nehru New College of Engineering, Shivamogga, Karnataka, India.

sukruthasbhat@gmail.com, psankalkoti@gmail.com

Abstract: The quick growth of digital documents has made efficient information retrieval a major challenge for both organizations and individuals. Traditional keyword-based search techniques often fail to deliver complete and contextually relevant results. To address this limitation, this study proposes an intelligent system titled AI-Driven Document Question Answering and Content Retrieval Using Large Language Models and Vector Embeddings. The system leverages the capabilities of large language models (LLMs) to enable accurate, context-aware, and interactive exploration of document content. The proposed approach processes unstructured documents by transforming them into semantic representations using embedding techniques and storing them in a vector database for efficient similarity-based retrieval. When a natural language query is submitted, the system identifies the most relevant document segments based on semantic meaning rather than simple keyword matching. The LLM then generates precise and human-like responses based on the retrieved information, ensuring both accuracy and readability. Moreover, the proposed framework enhances search efficiency, reduces manual effort, and improves user experience through a conversational interface. The system is designed to be scalable and supports multiple document formats, making it applicable across various domains such as academic research, enterprise knowledge management, customer support, and technical documentation. Overall, this work demonstrates how AI-driven retrieval systems, powered by large language models, can transform traditional document search into an intelligent knowledge discovery platform.

Keywords —Artificial Intelligence, Large Language Model, Streamlit, Queries, Groq, Machine Learning.

I. INTRODUCTION

Digital documents such as PDFs are widely used in education, research, and business environments. Although these documents store valuable information, retrieving specific content from large files is often time-consuming and inefficient. Traditional search methods rely on keyword matching, which cannot understand the meaning or intent behind user queries. This limitation leads to irrelevant results and increased manual effort. Additionally, users often struggle when they do not know the exact keywords present in the document. To overcome these challenges, this research introduces an AI-powered PDF chatbot that allows users to interact with documents using natural language. By combining semantic search and large language models, the system provides accurate and context-aware answers directly from the document content. A brief review of related literature is presented below. Li et al. [1] present a deep learning-based recruitment system that enhances candidate-job matching by analysing resumes and job descriptions. Chen and Huang [2] apply data-driven machine learning techniques to talent acquisition. Mehrabi et al. [3] survey different types of bias in machine learning models and propose mitigation strategies. Jain and Gupta [4] introduce machine learning techniques for

automatically classifying resumes. Kumar and Singh [5] propose an NLP-based system that ranks resumes by analysing skills, experience, and qualifications. Zhang et al. [6] present an automated resume screening system that extracts key information using NLP and ranks candidates based on job requirements. Goyal and Singh [7] explore predictive analytics techniques for recruitment, focusing on forecasting candidate success. Sajjadi et al. [8] demonstrate how machine learning models can analyse candidate work history to predict job performance. van Esch et al. [9] discuss the impact of AI on recruitment, emphasising automation and improved candidate engagement. Alghamdi and Alfalqi [10] review topic modelling techniques used in text mining, which are useful for extracting meaningful insights from resumes. Liem et al. [11] focus on fairness and ethical considerations in automated hiring systems. Bogen and Rieke [12] examine the risks of bias in algorithmic hiring systems and highlight the need for transparency. Paparrizos et al. [13] present early machine learning methods for resume classification. Fernandez and Fernandez-Mateo [14] analyse traditional hiring practices and identify biases related to social networks. Malinowski et al. [15] introduce a recommendation-based approach for matching candidates with job roles.

II. RESEARCH PROBLEM AND OBJECTIVE

A. Problem Statement

Traditional document retrieval systems present several key limitations:

- Depend on keyword matching
- Fail to understand context
- Require manual searching
- Provide incomplete or irrelevant results

B. Objective

The main objective of this research is to develop an intelligent document retrieval system that:

- Understands natural language queries
- Retrieves semantically relevant content
- Generates accurate and human-like responses
- Reduces manual effort and improves efficiency

III. PROPOSED METHODOLOGY

A. Research Design

This system follows a build-and-test approach, where a web-based application is developed to enable intelligent document retrieval.

B. Materials and Tools

- Python
- Streamlit (Frontend)
- LangChain (Workflow orchestration)
- HuggingFace Embeddings
- FAISS (Vector database)
- Groq API (LLM access)

C. Procedure (Steps)

- Upload PDF documents through the interface
- Extract text using PyPDFLoader
- Split text into smaller chunks
- Convert chunks into embeddings
- Store embeddings in FAISS
- Accept user query in natural language
- Retrieve relevant document chunks
- Pass data to LLM
- Generate and display answer

4.1 ARCHITECTURE DIAGRAM

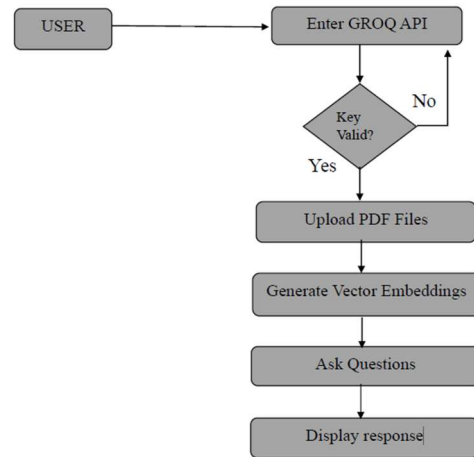


Fig 4.1: Working of PDF Chatbot

Fig. 1. Working of PDF Chatbot

TABLE I. TECHNIQUE COMPARISON TABLE

Technique	Evaluation Method	Strength	Limitation
Manual Search	Human reading	Low	Time-consuming
Keyword Search	Exact match	Medium	No context understanding
AI-Based Retrieval (Proposed)	Semantic + LLM	High	Depends on model

TABLE II. PARAMETER TABLE

Parameter	Meaning
Query	User question
Embeddings	Vector representation
Context	Retrieved document chunks

IV. RESULTS / FINDINGS

The system was tested using multiple PDF documents and queries. The system successfully retrieved relevant content. Responses were generated accurately and quickly, and chat-based interaction improved usability.

A. Display of the User Interface

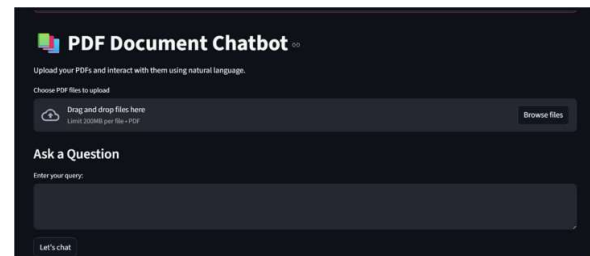


Fig. 2. Display of the User Interface

Once the Groq key is entered, the user is prompted to upload files for processing. There is a file size constraint of 200 MB.

B. Uploading the PDF Files and Query Entering

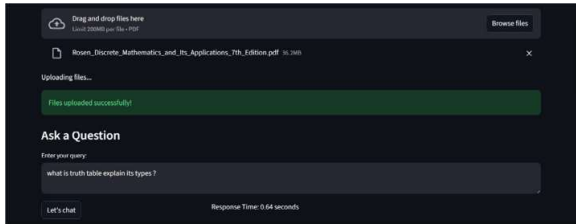


Fig. 3. Uploading the PDF Files and Query Entering

Example: Query: "What is truth table?" Output: Correct contextual explanation within seconds.

The system demonstrated: faster response time, improved accuracy, and better user interaction compared to traditional methods.

C. Chatbot Response

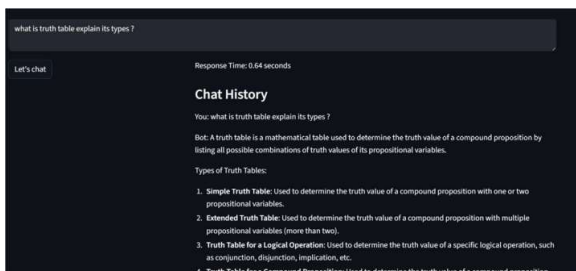


Fig. 4. Chatbot Response

Fig. 4 shows a chatbot interaction within a PDF-based chatbot interface where the user asks, "What is truth table explain its types?" The chatbot responds contextually by defining the truth table and its types. The response is quick, with a recorded response time of 0.64 seconds, demonstrating the chatbot's ability to retrieve relevant definitions from the uploaded document content efficiently.

V. DISCUSSION

The results indicate that integrating semantic search with LLMs significantly improves document retrieval.

A. Advantages

- Context-aware answers
- Conversational interface
- Reduced manual effort

B. Limitations

- Cannot process images or tables in PDFs
- Limited to text-based documents
- Depends on external API

The system performs well for academic and professional use but requires enhancements for real-world scalability.

VI. CONCLUSION

This research presents an AI-powered document retrieval system that transforms static PDF files into interactive knowledge sources. By combining semantic search and large language models, the system improves efficiency, accuracy, and user experience. The solution is particularly useful for students, researchers, and professionals who work with large documents. Future improvements can include OCR support, multilingual capabilities, and enhanced context handling.

REFERENCES

- [1] X. Li, Y. Wang, and Z. Zhang, "Intelligent Recruitment System Based on Deep Learning," 2022.
- [2] J. Chen and Y. Huang, "A Data-Driven Approach for Talent Acquisition Using Machine Learning," 2022.
- [3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," 2021.
- [4] A. Jain and R. Gupta, "Automated Resume Classification Using Machine Learning Approaches," 2021.
- [5] P. Kumar and A. Singh, "AI-Based Resume Ranking System Using NLP," 2021.
- [6] H. Zhang, J. Zhang, and Y. Liu, "Resume Screening with Natural Language Processing and Machine Learning Techniques," 2020.
- [7] N. Goyal and K. Singh, "Predictive Analytics in Recruitment Using Machine Learning," 2020.
- [8] S. Sajjadi et al., "Using Machine Learning to Translate Applicant Work History into Predictors of Performance," 2019.
- [9] P. van Esch, J. S. Black, and J. Ferolie, "Marketing AI Recruitment: The Next Phase in Job Application and Selection," 2019.
- [10] M. Alghamdi and K. Alfalqi, "A Survey of Topic Modeling in Text Mining," 2019.
- [11] C. C. S. Liem et al., "Psychometrics and Fairness in Automated Recruitment Systems," 2018.
- [12] M. Bogen and A. Rieke, "Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias," 2018.
- [13] J. Paparrizos et al., "Machine Learning Techniques for Resume Classification," 2011.
- [14] R. Fernandez and I. Fernandez-Mateo, "Networks, Race, and Hiring," 2006.
- [15] J. Malinowski, T. Keim, and T. Weitzel, "Matching People and Jobs: A Bilateral Recommendation Approach," 2006.