Chat with PDF: Your Go-to Website for Smarter Exam Prep with PDF Chat Support

Madhav Thigale[1]; Aditya Kumar[2]; Chetna Girme[3]; Apurva Gargote[4]

[2;3;4]Students
[1]Dr. D.Y. Patil Institute of Management and Research, Pune, India
[2;3;4]Department of Electronic and Telecommunication Engineering,
Dr. D.Y. Patil Institute of Management and Research, Pune, India

Publication Date: 2025/04/29

Abstract: This project builds an interactive application where users can upload multiple PDF documents and ask questions about them, offering a dynamic way to explore and retrieve information from large texts. The system processes the PDFs by extracting their text, chunking it into smaller sections, and converting these sections into numerical embeddings using advanced language models. These embeddings are stored in a FAISS vector database, enabling efficient similarity search and fast retrieval of relevant information based on user queries. The project uses Stream lit as the frontend framework to create a user-friendly web app, enabling users to interact with the system, upload PDFs, and receive chatbot responses. Accessed via API, powers the conversational AI, generating responses by creating text embeddings for similarity search, which are stored in FAISS for efficient retrieval. Lang Chain orchestrates the interactions between the AI model, memory, and retrieval systems, while utilities like PyPDF2 extract text from PDFs, and dotenv manages environment variables. The chatbot uses Open AI embeddings for text conversion and Conversation Buffer Memory to maintain context throughout user interactions.

Keywords: *PDF Interaction, Conversational AI, NLP, Text Extraction, Semantic Search, Intelligent Search, PyPDF2, User-Friendly Interface, Document Analysis, Information Retrieval.*

How to Cite: Madhav Thigale; Aditya Kumar; Chetna Girme; Apurva Gargote (2025). Chat with PDF: Your Go-to Website for Smarter Exam Prep with PDF Chat Support. *International Journal of Innovative Science and Research Technology*, 10(4), 1711-1716. https://doi.org/10.38124/ijisrt/25apr956

I. INTRODUCTION

In today's fast-paced world, efficiently extracting and understanding information from lengthy PDF documents can be a daunting and time-consuming task. The *Chat with PDF* application addresses this challenge by transforming the way we interact with PDFs. Instead of manually searching through pages of text, this innovative tool leverages Python and natural language processing (NLP) technologies to enable users to communicate with their PDF documents in a conversational style. The application provides an intuitive and interactive experience where users can ask questions or seek specific content from a PDF file, and the system responds intelligently by retrieving relevant information.

This approach simplifies the process of extracting insights from complex documents, making it particularly valuable for professionals, students, and researchers who work extensively with digital reports, articles, or legal documents.By combining advanced text extraction methods, semantic search algorithms, and a user-friendly interface, the *Chat with PDF* application redefines document analysis, making it faster, more accurate, and highly user-centric. This modern solution not only saves time but also enhances productivity, providing a practical tool for various fields and industries. Expanding on the concept of the *Chat with PDF* application, this innovative tool is designed to transform document management into an efficient and seamless experience. It aims to reduce the traditional frustrations of searching through PDFs for particular phrases, keywords, or ideas by making the interaction as simple as having a conversation.

Built in Python, the application harnesses the power of libraries like PyPDF2 for text extraction, combined with NLP frameworks such as spaCy or transformers to interpret and understand user queries. This allows the tool to deliver precise answers and meaningful summaries from the document's content. Built in Python, the application harnesses the power of libraries like PyPDF2 for text extraction, combined with NLP frameworks such as spaCy or transformers to interpret and understand user queries. It integrates FAISS for efficient similarity search, ensuring precise and context-aware responses. Additionally, the system maintains conversational context, allowing users to ask follow-up questions seamlessly.

II. LITERATURE REVIEW

[1]“Sharly AI: Conversational AI for Document Interaction” by Sharly AI. This study explores how AI-driven document interaction enhances comprehension by enabling users to chat with PDFs, receive summaries, and extract citations. It addresses the growing need for efficient document analysis, emphasizing AI’s ability to improve research productivity. The study highlights benefits such as quick information retrieval but acknowledges limitations, including challenges in handling complex document structures and occasional inaccuracies in AI-generated content. The authors suggest refining contextual accuracy and expanding integration with different file formats for better usability.

[2] “docAnalyzer.ai: AI-Powered Document Analysis” by docAnalyzer.ai. It examined AI-powered document analysis tools, allowing users to extract key information and engage in conversational queries with uploaded PDFs. Users benefited from automated data extraction and summarization while also facing issues with processing highly technical or unstructured documents. The study addresses the need to improve AI adaptability and incorporate domain-specific knowledge to enhance functionality, but it is limited by its focus on specific datasets and lacks empirical validation across diverse document types.

[3] “ChatDOC: AI-Powered PDF Assistant” by ChatDOC. It provides an overview of AI-powered PDF assistants to help users extract information, generate citations, and summarize content. The study finds the tool beneficial for research and academic use but identifies challenges in interpreting nuanced text and processing non-standard document layouts. While useful for improving document accessibility, the research lacks an in-depth assessment of AI comprehension accuracy. The study suggests future advancements in contextual understanding and AI refinement for better accuracy.

[4] “ChatPDF: AI for Research Paper Analysis” by ChatPDF. They explored AI’s role in facilitating research by allowing users to interact conversationally with PDFs. The study highlights its usefulness in summarizing academic papers and retrieving key insights while pointing out difficulties in handling complex theoretical concepts and mathematical expressions. The findings emphasize AI’s potential for academic support but acknowledge its reliance on text-based models, limiting its effectiveness in subjects requiring diagram or formula interpretation. The authors recommend improving AI’s comprehension of specialized academic content to enhance its research applications.

[5] “Semantic Search in AI-Powered Document Processing” by Smith et al. This study examines how semantic search improves AI-driven document interactions by enabling context-aware query responses. It highlights the efficiency of vector-based search techniques, such as FAISS, in retrieving relevant document sections but notes challenges in handling ambiguous queries and context misinterpretation. The authors suggest enhancing AI models with improved language understanding to refine search precision.

[6] “LangChain for Document Processing: AI-Driven Knowledge Retrieval” by Brown et al. It explores the integration of LangChain in document processing, focusing on its ability to orchestrate AI interactions, memory retention, and dynamic query responses. The study finds that LangChain improves user experience but faces issues with handling long and complex documents effectively. The research suggests refining retrieval strategies and expanding model capabilities to improve response accuracy.

➢ *Objective*

The Chat with PDF application aims to enhance document interaction by enabling users to query PDFs using natural language. It employs text extraction tools like PyPDF2, NLP techniques, and AI models such as OpenAI embeddings with FAISS for efficient retrieval. A Streamlit-based interface ensures seamless PDF uploads and user-friendly interactions. Context retention via Conversation Buffer Memory enables coherent multi-turn queries. The tool is designed for legal, academic, and business use, improving information retrieval speed and accuracy. The application simplifies document analysis by reducing the need for manual searching, making information access faster and more intuitive. It leverages LangChain to manage AI interactions, ensuring smooth query processing and response generation. Future improvements include expanding support for more document formats and enhancing retrieval accuracy with advanced AI models.

➢ *Background*

Traditional methods of document analysis require extensive manual effort, making information retrieval slow and inefficient. With the rise of AI-driven solutions, there is an increasing demand for tools that simplify interaction with large volumes of text. While search engines and document management systems provide basic retrieval, they often lack contextual understanding and conversational accessibility. The Chat with PDF application addresses this gap by enabling users to interact with documents conversationally, allowing for intuitive question-answering and rapid content retrieval. Built on advancements in natural language processing and vector search, the application transforms document analysis into a seamless experience. By leveraging AI models, embedding techniques, and efficient search algorithms, it ensures accurate and context-aware responses. The tool is designed to enhance productivity across various domains, from research and legal analysis to business and education, making complex information more accessible and actionable.

➢ *Problem Definition*

Many professionals, researchers, and students struggle with efficiently extracting relevant information from lengthy and complex PDF documents. Traditional keyword-based search methods often fail to provide context-aware results, requiring users to manually sift through extensive text. This process is time-consuming and reduces productivity, especially in fields that rely on rapid data retrieval, such as legal analysis, academic research, and business intelligence. The Chat with PDF application addresses this issue by

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr956

enabling users to interact with documents conversationally, allowing for precise and efficient information retrieval through AI-driven natural language processing.

III. METHODOLOGY AND DESIGN

➢ *User Onboarding Process*

The user onboarding process in the Chat with PDF application is designed to be intuitive and efficient, ensuring users can quickly interact with their documents. Upon signing up, users are prompted to upload their PDF files, which are processed using text extraction techniques like OCR and NLP-based parsing. To personalize the experience, users can set preferences such as selecting query styles or defining key topics of interest. The system provides an AI-driven tutorial to introduce core features, including intelligent search, summarization, and document segmentation. To ensure data security, users have the option to enable authentication methods, preventing unauthorized access to confidential files. By simplifying onboarding, the application enables users to efficiently retrieve insights from documents while ensuring a seamless and secure experience.

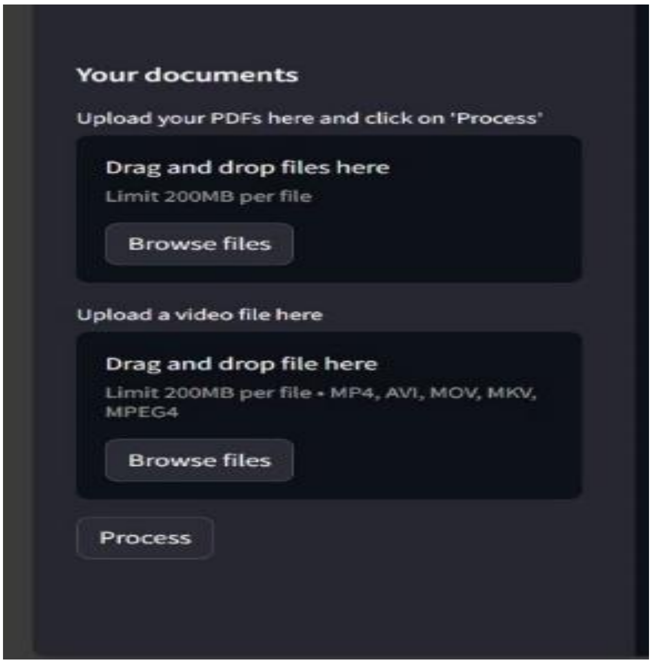


Fig 1 PDF Upload

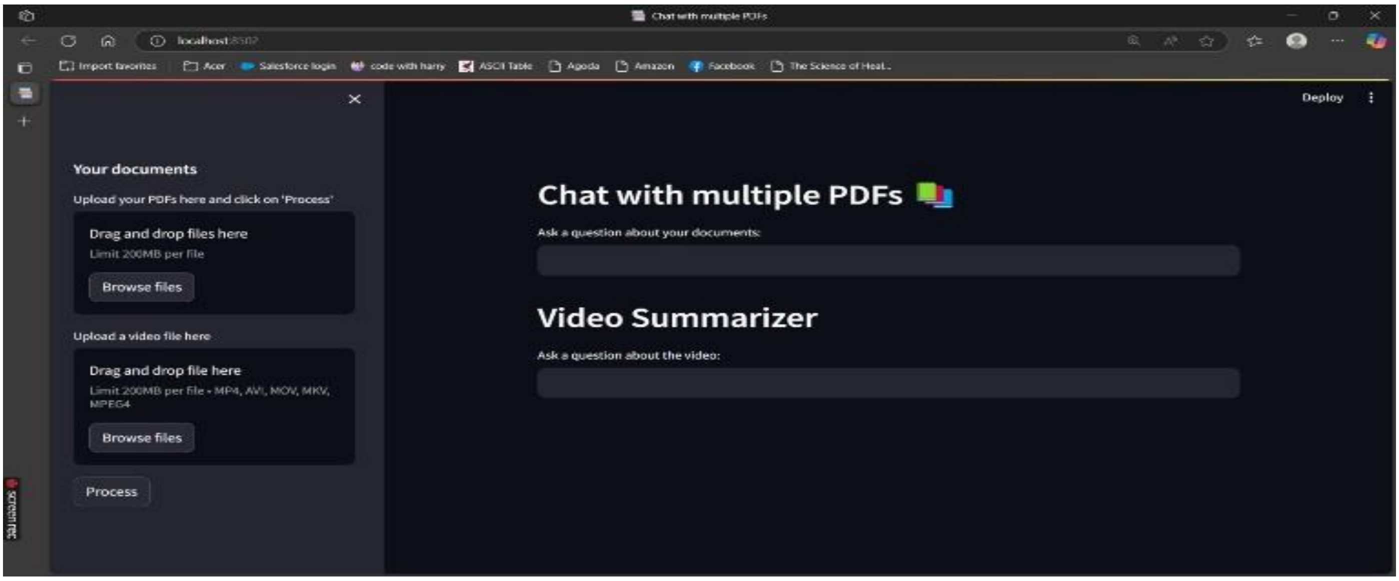


Fig 2 Chat Interface

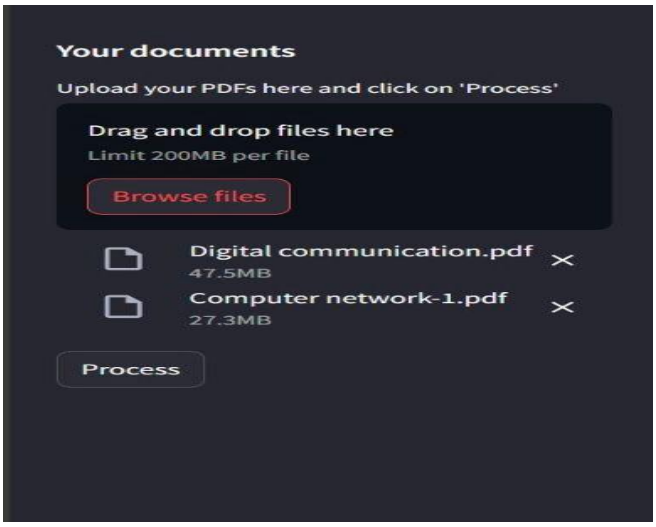


Fig 3 Multiple PDF upload

➢ *Productivity Features*

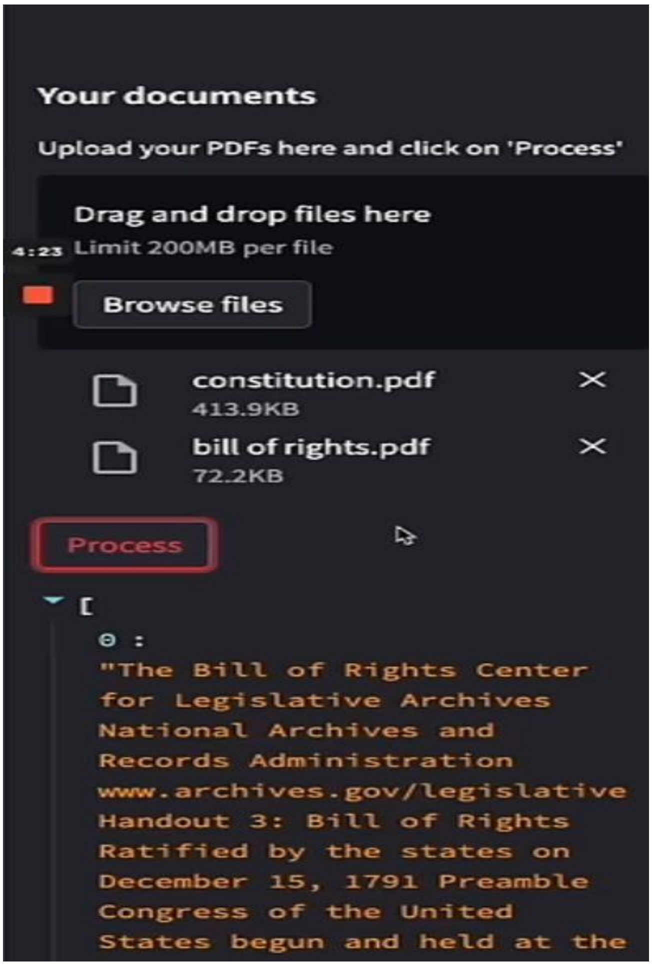
The Chat with PDF application enhances productivity by providing intelligent search for quick content retrieval and AI-powered summarization for key insights. Annotation and bookmarking features allow users to highlight important text and add notes. Keyword extraction helps identify essential themes, streamlining document navigation.

- *Task Manager:*

The Chat with PDF application includes a Task Manager to help users efficiently manage multiple PDFs by leveraging PDF text extraction, text chunking, embeddings, vector storage, and conversational retrieval chains. Users can create task lists, set deadlines, and prioritize sections for review based on extracted content. The progress tracking feature enables users to monitor completion rates while interacting with PDFs, ensuring structured document analysis. Additionally, smart reminders alert users about important sections and deadlines, enhancing workflow efficiency in retrieving and analyzing PDF content.

- *Text Chunking:*

One key component of your Chat with PDF project is Text Chunking. This process involves breaking down extracted text from PDFs into smaller, meaningful chunks to improve retrieval and conversational interactions. By dividing long documents into manageable sections, the system ensures more precise and context-aware responses when users query specific information. This step is crucial for maintaining coherence and relevance in conversations with multiple PDFs.



Your documents

Upload your PDFs here and click on 'Process'

Drag and drop files here

4:23 Limit 200MB per file

constitution.pdf
413.9KB

bill of rights.pdf
72.2KB

```
[
  0 :
  "The Bill of Rights Center
  for Legislative Archives
  National Archives and
  Records Administration
  www.archives.gov/legislative
  Handout 3: Bill of Rights
  Ratified by the states on
  December 15, 1791 Preamble
  Congress of the United
  States begun and held at the
```

Fig 4 Text Chunking

➢ *Vector Storage for Efficient Retrieval*

In your Chat with PDF application, Vector Storage plays a crucial role in retrieving relevant document content efficiently. After extracting text and breaking it into smaller chunks, the system converts these chunks into numerical representations called embeddings using a pre-trained language model (e.g., OpenAI's embeddings, SBERT, or FastText). These embeddings capture the semantic meaning of the text rather than just the keywords.

All the embeddings are stored in a vector database (such as FAISS, Pinecone, or ChromaDB), allowing for fast similarity searches. When a user asks a question, the system doesn't just look for exact keyword matches but instead retrieves the most contextually similar chunks based on their vector representations.

➢ *Natural Language Processing for Intelligent Interaction*

Advancements in Natural Language Processing (NLP) have revolutionized digital communication by enabling seamless human-computer interaction. In the context of this project, NLP plays a crucial role in understanding user queries and delivering relevant responses based on document content. Rather than relying on predefined keyword searches, NLP allows for contextual understanding, ensuring users receive meaningful and precise information.

This project leverages NLP techniques such as Named Entity Recognition (NER), text summarization, and semantic search to enhance user interactions. By analyzing sentence structures and extracting key insights, the system can provide concise and context-aware responses. Unlike traditional search algorithms, which rely on direct keyword matching, NLP enables more intuitive and conversational engagement with digital content.

Various approaches were evaluated, including rule-based models and traditional keyword-based search mechanisms. However, these methods often failed to capture contextual nuances and complex queries. As a result, advanced NLP techniques were chosen to provide a more dynamic and intelligent way of processing information, improving user experience and accessibility.

IV. THEORETICAL BACKGROUND

➢ *Matching Algorithm: Cosine Similarity*

Cosine Similarity is a widely used technique in information retrieval and natural language processing, particularly for measuring the similarity between two non-zero vectors. In the context of your Chat with PDF application, it is used to match user queries with the most relevant document sections by calculating the cosine of the angle between their vector representations. The smaller the angle, the higher the similarity, ensuring precise retrieval of relevant information.

This algorithm converts extracted text chunks into a multi-dimensional space, where each dimension represents a specific semantic feature of the content. For example, document sections with similar meanings—regardless of exact wording—are mapped closer together in this space.

By leveraging Cosine Similarity, the Chat with PDF application enhances the accuracy of document retrieval, providing users with highly relevant answers based on context rather than just keyword matches. This ensures a more efficient and intelligent document interaction experience, allowing users to quickly access the most important information without manually searching through PDFs.

➢ *Technology Stack*

The technology stack for the Chat with PDF application has been carefully selected to ensure efficient document retrieval, scalability, and seamless user interaction. The architecture includes:

- Frontend: The application interface is built using Streamlit, a Python-based framework that enables interactive and user-friendly UI development. Streamlit allows users to upload PDFs, query documents, and view responses in a simple and efficient manner.

- Backend: The backend is powered by Flask, which handles text processing, embeddings generation, and conversational retrieval chains. Text chunking is implemented to split extracted content into smaller sections, making it easier to retrieve relevant information. The system uses OpenAI embeddings or SBERT to convert text into vector representations, ensuring semantic search capabilities rather than relying on simple keyword matching.

- Vector Storage & Database: FAISS or ChromaDB is used for vector storage, allowing fast and accurate retrieval of relevant document sections based on user queries. Additionally, MongoDB Atlas stores user interactions and document metadata, ensuring structured and efficient data management.

- Hosting & Deployment: The application is hosted on Streamlit Cloud or Render, providing scalable and reliable deployment. This ensures real-time querying, fast document retrieval, and smooth performance, even with large PDF files.

V. FLOWCHART

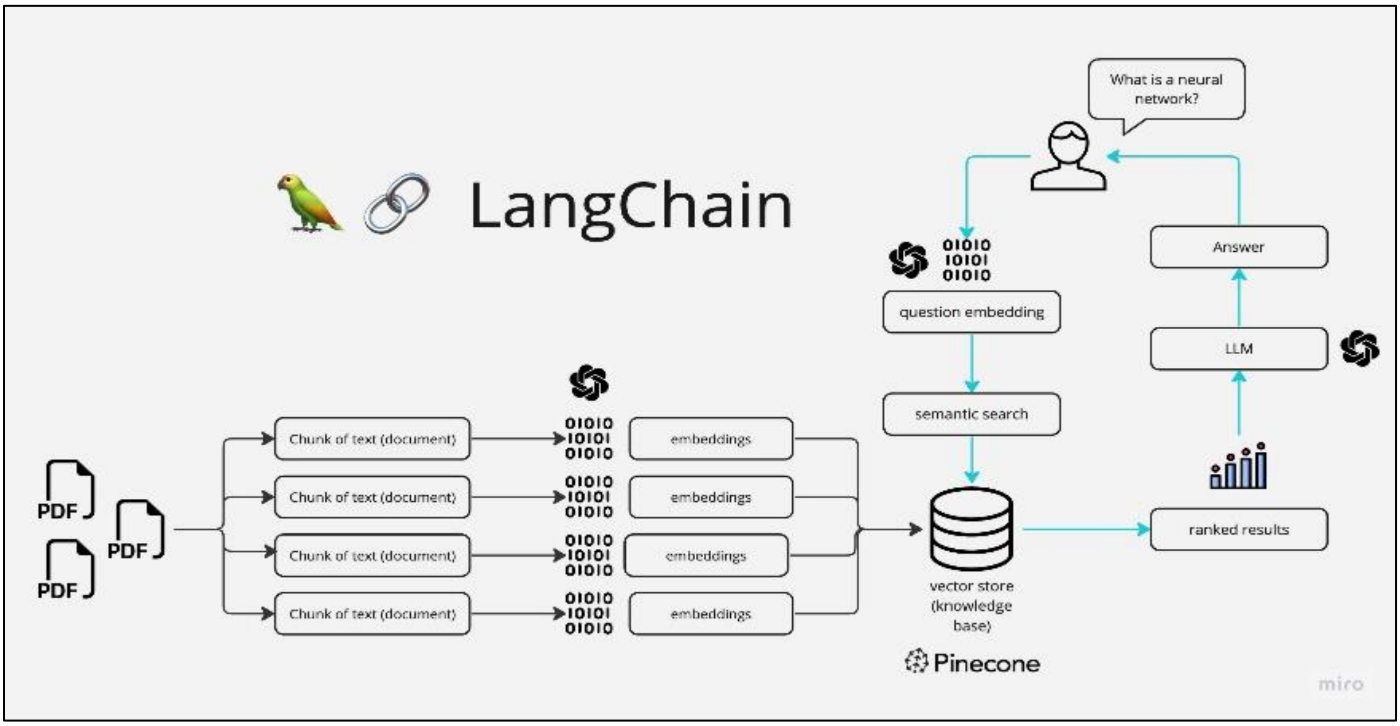


Fig 5 Long Chain

This image illustrates a **retrieval-augmented generation (RAG) system** using **LangChain** and **Pinecone** for question-answering over documents. The process begins with loading PDF documents, which are then split into smaller chunks of text. These chunks are converted into embeddings using an AI model and stored in a vector database (Pinecone), creating a knowledge base.

When a user asks a question, it is also converted into an embedding and compared with stored embeddings through semantic search. The most relevant document chunks are retrieved and passed to a large language model (LLM), which generates an answer based on the retrieved content.

The results are ranked to improve accuracy, ensuring that the user receives the most relevant information. This system enhances LLM responses by providing external context, making it a powerful tool for AI-driven knowledge retrieval from documents.

VI. FUTURE SCOPE

Future enhancements for this project focus on integrating advanced NLP techniques to improve user interaction with digital documents. Planned developments include AI-driven contextual understanding to refine responses, enabling deeper engagement with PDF content. Additionally, interactive learning features such as summarization tools and automated question-answering models will be introduced to streamline information retrieval. To enhance accessibility, voice-based interaction and multilingual support are also being considered. Looking ahead, incorporating generative AI for real-time content explanations and personalized study assistance will further elevate the user experience, ensuring seamless and intelligent engagement with digital documents.

VII. ACKNOWLEDGEMENT

We extend our sincere gratitude to Dr. D. Y. Patil Institute of Management and Research for providing the essential resources and support for this research. We sincerely thank our mentor for their invaluable guidance and our peers for their support and insightful feedback.

VIII. CONCLUSION

The Chat with PDF application represents a significant advancement in enhancing user interaction with digital documents. By integrating natural language processing (NLP) with an intuitive conversational interface, the platform enables users to efficiently extract, comprehend, and retrieve information from PDFs. This eliminates the need for manual searching and enhances accessibility, making document analysis more seamless and efficient.

With features such as context-aware question answering, intelligent summarization, and interactive document navigation, the application empowers users across academic, professional, and research domains. Its focus on user-friendly interaction and AI-driven assistance positions it uniquely in the digital workspace, making it a valuable tool for individuals handling large volumes of textual information. As we move into the next phase of development, our commitment remains strong in refining these features, improving accuracy, and expanding functionality to better serve diverse user needs.

REFERENCES

[1]. “Massive Open Online Course Study Group: Interaction Patterns in Face- to-Face and Online (Facebook) Discussions” by Pin-Ju Chen and Yang-Hsueh Chen https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.670533/full

[2]. “AN EVALUATION OF STUDENTS EXPERIENCES OF USING VIRTUAL STUDY SPACES” by UCL LIBRARY SERVICES with INFORMATION SERVICES DIVISION, FACULTIES and DEPARTMENTS https://discovery.ucl.ac.uk/id/eprint/10132327/1/An%20Evaluation%20of%20UCL%20Virtual%20Learning%20Spaces%20-%20Final%20Report%20July%202021.pdf

[3]. “Web-based Collaborative Learning” by Fan Qing, Lin Li https://www.sciencedirect.com/science/article/pii/S1878029611008528?ref=pdf_download&fr=RR-2&rr=8d7c43483bde3b4f

[4]. “Exploring the role of social media in collaborative learning the new domain of learning” by Jamal Abdul Nasir Ansari and Nawab Ali Kha. https://slejournal.springeropen.com/articles/10.1186/s40561-020-00118-7