International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

Next-Gen PDF Question and Answer Platform

Dr. Nupoor M. Yawale[1], Prasad S. Kapatkar[2], Devanshu M. Pachpohar[3], Pratham R. Nimsatkar[4], Rohit R. Gurjar[5]

[1]*Assistant Professor, Department of Computer Science and Engineering, Prof. Ram Meghe Institute of Technology & Research, Amravati, India*

[2,3,4,5]*Undergraduate Student, Department of Computer Science and Engineering, Prof. Ram Meghe Institute of Technology & Research, Amravati, India*

***Abstract*— The increasing use of digital documents such as research papers, reports, and technical manuals has created a need for efficient information retrieval systems. Traditional keyword-based search methods often fail to understand the context of user queries, resulting in irrelevant or incomplete results. This paper presents a Next-Gen PDF Question & Answer Platform that allows users to interact with PDF documents using natural language queries. The system is based on Retrieval-Augmented Generation (RAG) and integrates the LangChain framework to connect language models with external document data. The platform processes uploaded PDF files, converts them into semantic embeddings, and stores them in a vector database for efficient retrieval. When a user submits a query, the system retrieves the most relevant document segments and generates context-aware answers. The proposed system also supports features such as voice interaction, real-time response generation, and a user-friendly interface. Experimental results demonstrate that the system improves accuracy, reduces search time, and enhances overall user experience in document interaction.**

***Keywords*- LangChain, NLP, PDF Question Answering, RAG, Semantic Search, Vector Database**

I. Introduction

The rapid growth of digital information has significantly increased the use of electronic documents such as research papers, reports, and manuals. As a result, the need for efficient and intelligent systems to manage and retrieve information from these documents has become essential. Traditional search techniques often rely on keyword matching, which fails to capture the actual meaning and context of user queries [1][2]. Therefore, there is a growing demand for systems that can understand natural language and provide accurate responses based on document content. With the advancement of artificial intelligence and natural language processing, it is now possible to design systems that can interpret user queries and deliver meaningful results efficiently [3].

The proposed PDF Question & Answer Platform is designed to provide a modern and effective solution for document interaction through a structured and intelligent framework.

The system consists of multiple modules, including the Document Processing Module, which handles PDF extraction and text conversion, the Retrieval Module, responsible for identifying relevant information using semantic search, and the Response Generation Module, which produces accurate answers using language models. Additionally, the system includes a User Interface that allows users to upload documents, ask questions, and receive responses in real time. The integration of these modules ensures efficient data processing, retrieval, and response generation, thereby improving overall system performance [4][5].

Another important aspect of such intelligent systems is the interaction between users and the platform. A well-designed system should provide a smooth and user-friendly experience that encourages users to interact more frequently and efficiently. Effective interaction improves user satisfaction, builds trust in the system, and enhances decision-making capabilities. Furthermore, advanced features such as voice-based interaction and real-time response generation contribute to better accessibility and usability. A system that continuously adapts to user needs and provides accurate responses will naturally gain user confidence and reliability over time [6][7].

Moreover, maintaining the quality and accuracy of responses is a critical factor in the success of such systems. By using techniques such as Retrieval-Augmented Generation (RAG) and vector-based search, the system ensures that responses are context-aware and relevant to the user's query. Proper management of document data and continuous improvement of the model can further enhance system performance. As user expectations continue to evolve, the system must also adapt to new challenges and provide scalable solutions. This will ensure long-term usability and effectiveness in real-world applications [8][9].

II. Literature Review

With the rapid advancement of Natural Language Processing (NLP) and machine learning techniques, several approaches have been developed for document-based question answering systems.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

Earlier systems mainly relied on keyword-based search mechanisms, which were limited in understanding the context of user queries and often resulted in irrelevant or incomplete answers [10][11]. These limitations highlighted the need for more intelligent systems capable of processing semantic meaning rather than just matching keywords.

Lewis et al. introduced the Retrieval-Augmented Generation (RAG) model, which combines information retrieval with generative language models to produce accurate and context-aware responses. This approach ensures that the generated answers are grounded in relevant document content, thereby reducing the chances of misinformation and improving reliability [12][13]. The integration of retrieval and generation has significantly enhanced the performance of question answering systems.

Chase proposed the LangChain framework, which provides an efficient way to connect large language models with external data sources such as documents, APIs, and databases. It offers functionalities including document loading, text splitting, embedding generation, and query handling, making it a powerful tool for developing AI-based applications. The flexibility and modularity of LangChain enable developers to build scalable and efficient systems [14][15].

Furthermore, transformer-based models such as BERT have greatly improved the ability of systems to understand contextual relationships between words in a sentence. These models use deep learning techniques to capture semantic meaning, resulting in more accurate information retrieval compared to traditional approaches. However, challenges such as processing large volumes of data, computational cost, and maintaining response accuracy still persist in modern systems [16][17].

The proposed system addresses these challenges by integrating RAG, vector databases, and optimized text chunking techniques. This combination improves retrieval efficiency, enhances response accuracy, and ensures scalability for handling large document collections. As a result, the system provides a more reliable and efficient solution for document-based question answering applications [18][19].

III. Methodology

The proposed PDF Question & Answer System is designed using a modular and pipeline-based architecture that ensures efficient document processing, real-time response generation, and scalability. The system integrates advanced Natural Language Processing techniques with vector-based retrieval mechanisms to provide accurate and context-aware answers.

The architecture consists of multiple layers, including document processing, embedding generation, retrieval, and response generation, as illustrated in the system architecture diagram [20][21]

A. Overview of System Architecture

The system follows a multi-layered architecture consisting of the presentation layer (user interface), application layer (processing and retrieval services), and data layer (vector database and document storage). This layered approach improves system maintainability, scalability, and performance. The frontend provides an interactive interface where users can upload PDF documents and submit queries, while the backend processes requests and communicates with the database. The separation of layers ensures efficient handling of tasks and secure data processing [22][23].

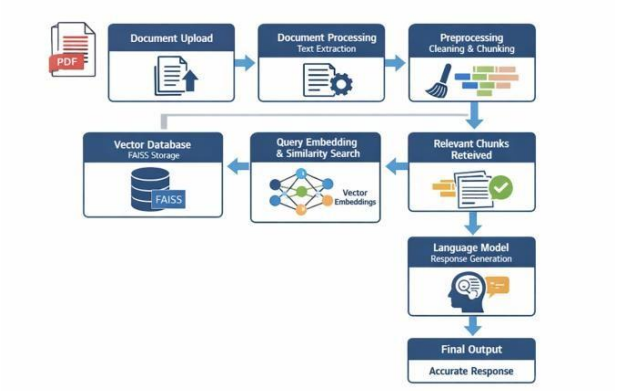

Fig. 1. The Workflow of the Next-Gen PDF Q&A Platform

B. Document Processing and Embedding Layer

The document processing layer is responsible for extracting and preparing text data from uploaded PDF files. The extracted text undergoes preprocessing, including removal of noise, formatting corrections, and segmentation into smaller chunks. These chunks are then converted into vector embeddings using advanced embedding models, which capture the semantic meaning of the content. This process enables the system to understand context rather than relying on simple keyword matching [24][25].

To improve efficiency, optimized chunking and indexing techniques are applied to the processed text. Each chunk preserves contextual relationships, ensuring that important information is retained during segmentation. The embeddings are stored with relevant metadata such as document source and position, enabling accurate retrieval.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

C. Vector Database and Retrieval Mechanism

The generated embeddings are stored in a vector database such as FAISS, which allows efficient similarity-based search operations. When a user submits a query, the system converts the query into an embedding and compares it with stored vectors to retrieve the most relevant document segments. This retrieval mechanism ensures that only contextually appropriate information is selected, thereby improving the accuracy of the responses [26][27].

D. Response Generation Layer

The Response Generation Layer is one of the most important parts of the system, as it is responsible for producing the final answer to the user's query. After the relevant information is retrieved from the processed PDF documents, this layer uses a Large Language Model to generate a meaningful and context-aware response. The system combines the user's question with the retrieved document content and sends it to the selected AI model. Based on this input, the model understands the context and generates a clear and accurate answer. This ensures that the response is not only correct but also easy to understand.

E. User Interface and Interaction Module

The frontend layer provides a user-friendly interface that allows users to interact with the system efficiently. Users can upload documents, ask questions, and receive answers in real time. The system also supports additional features such as chat-based interaction and voice input/output for enhanced accessibility. The interface is designed to ensure ease of use while maintaining responsiveness and performance, thereby improving overall user experience and engagement [28][29].

Table I
System Module Description

No	Next-Gen PDF Question and Answer Platform		
	Module	Function	Tools/Tech Used
1	Document Processing	Extract text from PDF	PyPDF, NLP
2	Text Chunking	Divide text into parts	LangChain
3	Embedding Generation	Convert text to vectors	OpenAI / Sentence Transformers
4	Vector Database	Store embeddings	FAISS
5	Retrieval Module	Find relevant chunks	Similarity Search
6	Response Generation	Generate answers	LLM (GPT / Mistral)

IV. Result & Analysis

This section presents the execution and testing of the proposed system, highlighting its real-time performance and behavior during implementation. The following screenshots demonstrate different functionalities of the system.

A. System Interface Overview

The system interface represents the initial screen of the "Chat with PDF" application running on a local environment. The layout is divided into three main sections: the history panel, central interaction area, and settings panel. At this stage, no document is uploaded, so the system prompts the user to add a PDF file. The interface is designed to be simple and user-friendly, allowing easy navigation for all users. The dark theme improves readability and reduces eye strain. This screen acts as the starting point for all further operations.

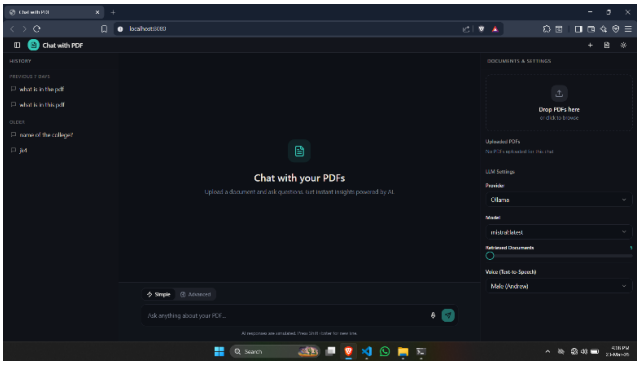

Fig. 2. System Interface Overview

B. PDF Upload and Settings Panel

This screen illustrates the successful upload of a PDF document along with various system configuration options that enable efficient processing. Once the document is uploaded, its status is displayed as "Ready," indicating that preprocessing steps such as text extraction, cleaning, and indexing have been completed successfully, ensuring that the system is prepared to handle user queries without delay. For instance, users can select different language models based on performance requirements or adjust retrieval settings to control the number of document chunks used during response generation.

In addition to basic processing, the system converts the uploaded PDF into structured textual data and divides it into smaller meaningful chunks, which improves the efficiency of semantic search by retrieving only the most relevant content. Each chunk is associated with metadata such as its position in the document, enabling more accurate and context-aware responses.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

Furthermore, the settings panel enhances user control by allowing dynamic adjustments without requiring system restarts, making the system flexible for different use cases such as academic research, technical document analysis, and report summarization. The interface is designed to be intuitive and user-friendly, ensuring that even non-technical users can easily upload documents and configure system settings, while real-time feedback mechanisms such as status indicators and progress updates improve usability and provide a smooth interaction experience. Overall, this stage confirms that the system is fully initialized and optimized for fast, accurate, and context-aware document-based question answering.based question answering.

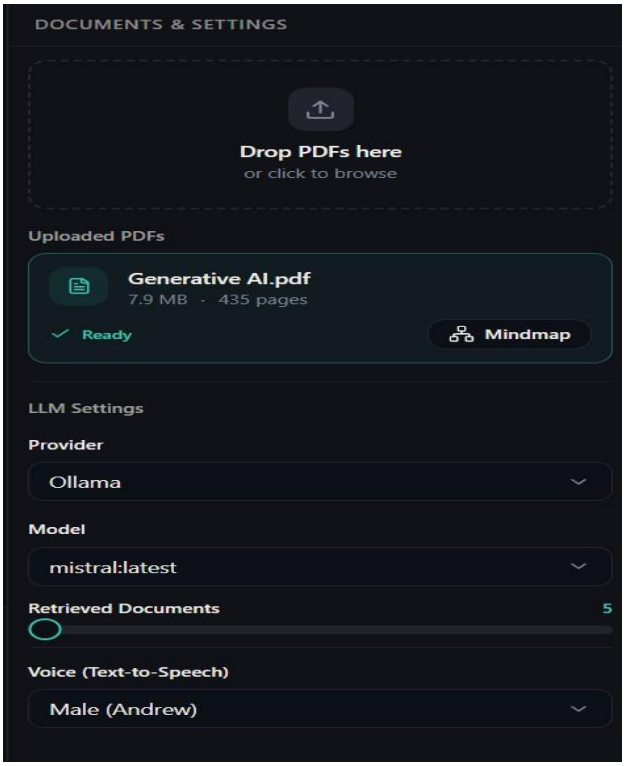

Fig. 3. PDF Upload and Settings Panel

C. LLM Provider Selection

This screenshot illustrates the selection of the AI provider used by the system. Users can choose between local and cloud-based providers. In this case, a local provider is selected, ensuring better data privacy and offline functionality.

The dropdown-based selection makes it easy to switch between providers. This feature enhances flexibility and allows the system to adapt to different user requirements.

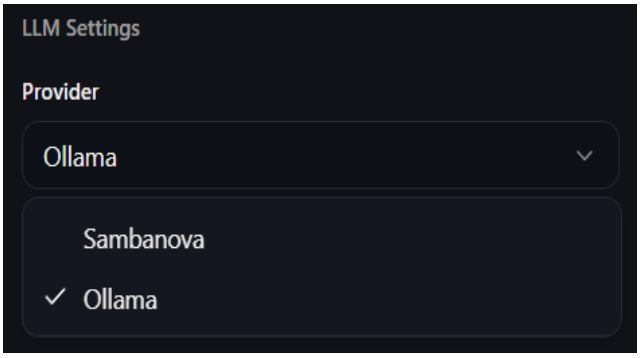

Fig. 4. LLM Provider Selection

D. Model Selection

The model selection feature allows users to choose different language models based on performance needs. Each model offers a trade-off between speed and accuracy.

This flexibility enables users to optimize the system according to their hardware capabilities. The simple interface ensures that model selection can be done easily without technical complexity.

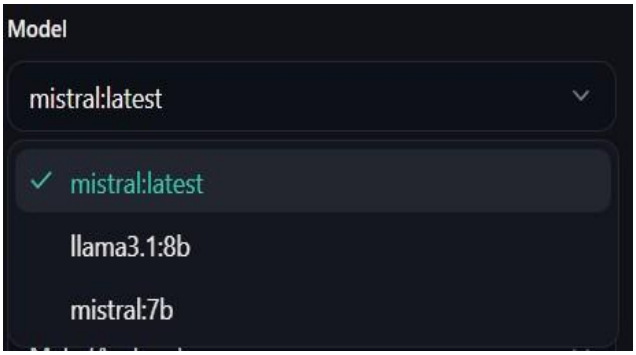

Fig. 5. Model Selection

E. Voice Selection Feature

This screenshot shows the text-to-speech functionality of the system. Users can select different voice options to convert text responses into audio output. This feature improves accessibility and enhances user interaction. It is particularly useful in scenarios where users prefer audio-based responses instead of reading text.

Furthermore, the voice selection feature allows users to choose preferred accents or tones, making the interaction more personalized. It also enhances usability for visually impaired users and in situations where listening is more convenient than reading.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

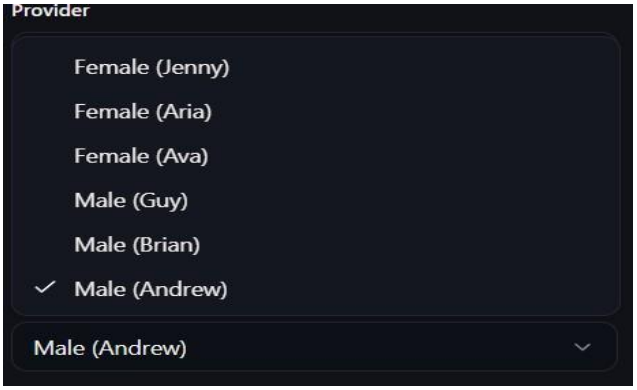

Fig. 6. Voice Selection Feature

F. Chat Response After PDF Upload

This screen demonstrates the actual working of the system after document processing. The user query is answered based on the content of the uploaded PDF.The response generated is relevant and context-aware, showing that the system successfully retrieves and processes document information. This confirms the effectiveness of the retrieval based approach.

The system efficiently analyzes the uploaded document and extracts meaningful information to generate accurate responses. It minimizes irrelevant outputs by focusing only on the context of the given PDF. This improves user experience by providing quick and precise answers. Overall, the approach ensures reliability and scalability for handling large document-based queries.

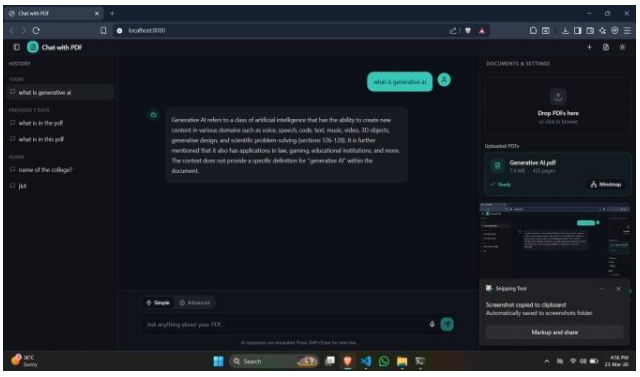

Fig .7. Chat Response After PDF Upload

G. Mind Map Generation Feature

The mind map feature provides a visual representation of the document's structure. It organizes key concepts into a hierarchical format, making it easier to understand complex information.This functionality enhances learning and quick analysis of large documents. It also improves user experience by presenting information in a structured and interactive manner.

Additionally, the system automatically extracts key topics and subtopics from the document to generate a clear conceptual map. The visual layout helps users quickly understand relationships between different sections, reducing the need to read the entire document in detail. This feature is especially useful for revision, presentations, and knowledge summarization, significantly improving information accessibility and understanding.

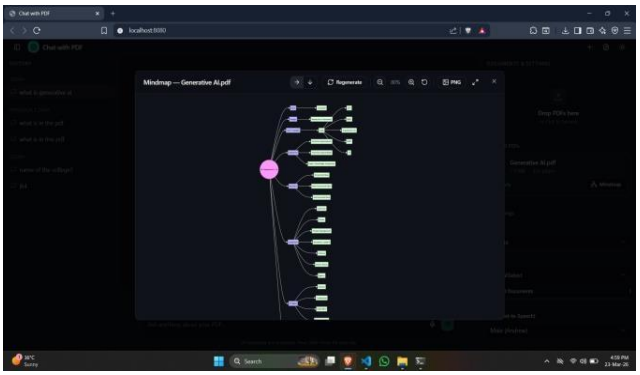

Fig. 8. Mind Map Generation Feature

V. Conclusion

The Next-Gen PDF Q&A Platform successfully demonstrates the integration of modern artificial intelligence techniques with document processing to create an intelligent and user-friendly system. By using the Retrieval-Augmented Generation (RAG) approach, the system generates accurate and context-aware answers directly from uploaded PDF documents. The use of semantic search enables better understanding of user queries beyond simple keyword matching. Vector databases help in efficient storage and fast retrieval of document embeddings. The integration of language models ensures that responses are meaningful and human-like. This approach overcomes the limitations of traditional search systems and improves overall efficiency. The system provides reliable and precise information retrieval from large documents.

The system also delivers a smooth and interactive user experience through its well-designed frontend and efficient backend processing. Features such as chat-based interaction, document upload, and real-time responses make it easy to use. Voice input and output improve accessibility and user engagement. The backend efficiently manages document processing, embedding generation, and query handling. The modular architecture ensures scalability and allows easy future enhancements. The system can handle large datasets and multiple users effectively. Overall, the project achieves its goal of building a smart and efficient document question answering platform for real-world applications.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Khandelwal, and S. Riedel, "RetrievalAugmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2] H. Chase, "LangChain: Building LLM-Powered Applications," 2022. [Online]. Available: https://www.langchain.com/

[3] S. Abacha and D. Demner-Fushman, "A question answering system for clinical literature," *Journal of Biomedical Informatics*, vol. 66, pp. 1–9, 2017.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL*, 2019.

[5] T. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[6] D. Chen and R. Socher, "Reading Wikipedia to Answer Open-Domain Questions," in *Proceedings of ACL*, 2017.

[7] N. Joshi and S. Patel, "BERT-Based Document QA System for Academic PDFs," *International Journal of Computer Applications*, vol. 182, no. 29, pp. 1–6, 2021.

[8] K. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in *Proc. EMNLP*, 2020, pp. 6769–6781.

[9] OpenAI, "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," 2022. [Online]. Available: https://openai.com/research/whisper

[10] J. Kiseleva et al., "Improving QA Systems through User Feedback," in *SIGIR*, 2016.

[11] J. Amershi et al., "Guidelines for Human-AI Interaction," in *CHI Conference on Human Factors in Computing Systems*, 2019.

[12] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *EMNLP-IJCNLP*, 2019.

[13] J. Guu et al., "REALM: Retrieval-Augmented Language Model Pre-Training," in *Proc. ICML*, 2020.

[14] A. Ahmed, M. Khan, and R. Mehmood, "Automated Legal Document Reader using RAG and LangChain," in *International Conference on AI and Law*, 2023.

[15] V. Kumar and A. Rathi, "Question Answering System for Corporate Reports using RAG and FAISS," *IEEE Access*, vol. 10, pp. 10756–10764, 2022.