As per UGC guidelines an electronic bar code is provided to secure your paper

International Journal for Modern Trends in Science and Technology
Volume 12, Issue 06, pages 354-361.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue06.html
DOI: https://doi.org/10.5281/zenodo.20739569

Interactive Pdf Assistant: Leveraging LangChain and Streamlit for Seamless Document Engagement

K. Bhanu Chand, Velagala Jaya Naga Santhosh Reddy, Kolla Venkata Rama Harsha Vardhan, Adidela Rahul Bunny

Department of Information Technology Swarnandhra College of Engineering & Technology, Seetharamapuram, Narsapur, Andhra Pradesh, INDIA.

To Cite this Article
K. Bhanu Chand, Velagala Jaya Naga Santhosh Reddy, Kolla Venkata Rama Harsha Vardhan & Adidela Rahul Bunny (2026). Interactive Pdf Assistant: Leveraging LangChain and Streamlit for Seamless Document Engagement. International Journal for Modern Trends in Science and Technology, 12(06), 354-361. https://doi.org/10.5281/zenodo.20739569

Article Info
Received: 16 May 2026; Revised: 07 May 2026; Accepted: 12 June 2026.

Copyright © The Authors ; This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

KEYWORDS

PDF Documents, Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Vector Databases, LangChain, Streamlit

ABSTRACT

The Interactive PDF Assistant is an intelligent document analysis system designed to enhance the way users interact with PDF files using natural language processing. This project leverages a Retrieval-Augmented Generation (RAG) architecture to enable accurate, context-aware question answering from uploaded documents. The system extracts text from PDFs using PyMuPDF, processes and segments the content into manageable chunks, and generates vector embeddings stored in a ChromaDB database for efficient retrieval. A key feature of the system is its integration with local Large Language Models (LLMs) through Ollama, ensuring data privacy and offline capability. The application supports multilingual interaction, allowing users to ask questions and receive responses in multiple languages such as English, Telugu, Hindi, and others. Additionally, it incorporates text-to-speech functionality, enabling users to listen to responses, thereby improving accessibility. The platform is built using Streamlit for an interactive user interface and LangChain for orchestrating the retrieval and generation pipeline. It also includes intelligent question suggestion capabilities to guide user exploration of documents. Overall, the system improves document comprehension, reduces manual reading effort, and provides an efficient, user-friendly solution for extracting meaningful insights from large PDF files.

1. INTRODUCTION

In the contemporary era of the "Information Explosion," the global digital landscape is defined by an unprecedented accumulation of unstructured data. Among the various formats used to store this data, the Portable Document Format (PDF) remains the most ubiquitous standard for document exchange across every major sector, including academia, legal industries, healthcare, and corporate governance. Statistics indicate that trillions of PDF files exist globally, housing critical

knowledge ranging from complex scientific breakthroughs to intricate financial audits. However, as the volume of these documents grows, a significant technical paradox has emerged: while we have managed to digitize vast amounts of human knowledge, the ability to interact with and extract specific insights from these static files remains remarkably primitive. Most users still rely on manual scrolling, skimming, and basic keyword matching—methods that were designed for an era of much smaller data scales. The "Information-Dense" nature of modern PDFs often leads to a state of "Information Overload," where the sheer depth and length of documents act as a barrier to productivity rather than a facilitator. For a researcher or a professional, navigating a 500-page technical manual or a multi-volume legal contract is "interaction-poor." The user is required to have prior knowledge of the terminology used by the author to find specific facts, and they lack the ability to ask high-level conceptual questions. This inefficiency represents a major bottleneck in the modern knowledge economy. There is a desperate need for a paradigm shift—moving from "reading" a document to "conversing" with it. To bridge this gap, this project proposes the development of a sophisticated Interactive PDF Assistant powered by Retrieval-Augmented Generation (RAG). This architecture represents the cutting edge of Artificial Intelligence, combining the precision of traditional information retrieval with the creative reasoning of Large Language Models (LLMs). By transforming static text into a searchable vector space, the system allows users to engage in a natural language dialogue with their documents. Instead of searching for words, users search for meaning. This framework is designed to prioritize three core pillars: accuracy, by grounding responses in the source text; privacy, by utilizing local inference through the Ollama framework; and accessibility, by integrating multilingual support and Text-to-Speech (TTS) capabilities. Ultimately, this system provides a scalable, secure, and highly efficient solution for transforming massive unstructured PDF datasets into immediate, actionable intelligence.

1.1 OVERVIEW OF DOCUMENT INTERACTION AND ANALYTICAL CHALLENGES

The evolution of digital documentation has led to a situation where the utility of information is often restricted by the formats in which it is stored. In the modern information landscape, the ability to quickly extract accurate information from vast PDF libraries is a significant global requirement for researchers, students, and professionals alike. Traditionally, document interaction has been a linear and manual process. Users are forced to spend hours skimming through pages or rely on primitive search functions that depend on exact character matching. This process is inherently limited because it assumes the user already knows the specific vocabulary used by the author, creating a barrier for those seeking to understand complex or unfamiliar subjects. A major analytical challenge in document interaction is the unstructured nature of PDF data. Information is often trapped in complex multi-column layouts, nested tables, and diverse font styles which make automated extraction highly difficult. Most standard scripts fail to maintain the logical flow of such documents, leading to fragmented data clusters that lose their original context. Beyond structural issues, the "Information Silo" effect presents a significant challenge where valuable insights are buried under layers of technical jargon, making it difficult for users to cross-reference data points across different chapters. When dealing with hundreds of pages, human readers struggle with a heavy cognitive load that leads to fatigue and overlooked details. These analytical challenges necessitate a shift toward intelligent, automated systems that can accurately interpret document structures and facilitate a natural, human-like interaction with digital text, effectively turning a static file into a responsive knowledge base.

1.2 NEED FOR AI IN DOCUMENT INTERACTION

The exponential increase in the volume and complexity of digital records has rendered manual reading and traditional indexing methods inefficient and prone to human error. Complex reports, academic papers, and multi-volume manuals require advanced computational methods to enable effective synthesis and data retrieval. Artificial Intelligence offers the necessary tools to process and interpret such data at high speeds, enabling faster and more accurate knowledge extraction than any conventional software could provide. As the volume of data grows, the "human-only" approach to document review has reached its breaking point, making AI-assisted interaction a fundamental requirement for

modern information management. AI is required to move beyond the surface level of text analysis to identify underlying concepts rather than just matching characters. This capability is vital for recognizing synonyms, identifying that different words share the same underlying meaning, and providing answers that are specific to the user's intent. Furthermore, AI is uniquely capable of global synthesis, identifying connections between a statement on page one and a related fact found hundreds of pages later, a task that is nearly impossible for a human reader to perform consistently. However, general AI models often struggle with hallucinations when they lack a specific reference, which drives the development of specialized architectures that ground the model's intelligence within the specific pages of a provided PDF. By integrating AI into the document workflow, the guesswork involved in information retrieval is eliminated, providing users with a reliable and data driven assistant.

1.3 ROLE OF NLP AND LLMs IN TEXT PROCESSING

Natural Language Processing (NLP) and Large Language Models (LLMs) serve as the core technological pillars for modern document analysis. NLP is the branch of artificial intelligence that enables computers to understand, interpret, and generate human language. In document analysis, NLP has evolved from simple rule-based grammar checking to sophisticated Transformer-based architectures. These modern architectures allow for the processing of entire paragraphs simultaneously, capturing the subtle nuances of language such as tone, context, and complex logical relationships. By employing NLP, systems can read a document with a level of comprehension that mirrors human cognitive processes, involving complex steps such as tokenization, text splitting, and embedding generation to turn raw text into structured intelligence.

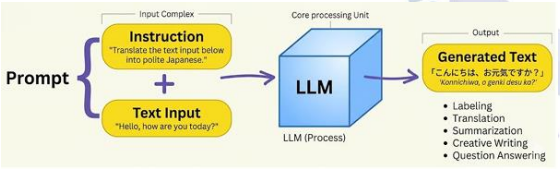


Fig 1: Working of Large Language Model

The integration of Large Language Models provides the project with a high-level reasoning engine that goes beyond following a fixed script. An LLM can synthesize information from multiple sections of a document to provide a coherent summary or answer while maintaining contextual logic. This ensures the system understands the difference in meaning of the same word based on surrounding sentences. Furthermore, the use of local LLMs ensures that this processing occurs entirely on the user's machine, providing a crucial layer of security for confidential data. By combining advanced NLP techniques with powerful local LLMs, a balance is achieved between high-performance intelligence and strict data privacy, allowing the system to follow complex instructions such as translating summaries or identifying specific data points without requiring prior specific training for those tasks.

1.4 DOCUMENT INTERACTION USING RAG ARCHITECTURE

The implementation of a Retrieval-Augmented Generation (RAG) framework is essential for achieving accurate and grounded document interaction. RAG is an architecture that enhances the capabilities of an LLM by providing it with external, verifiable data specifically from the content of an uploaded PDF. This approach solves the fundamental limitation of standard AI, which is the tendency to generate factually incorrect information when it lacks knowledge. By using RAG, the system first retrieves the specific chunks of text that are relevant to the user's question and then directs the LLM to generate an answer based exclusively on that evidence. This ensures that every response is directly supported by the source material, providing a level of reliability that is necessary for professional and academic use. This process involves a sophisticated technological pipeline starting with text extraction using libraries to convert PDF formatting into raw text. The text is then split into smaller, overlapping chunks for granular analysis and converted into vector embeddings using specialized models. These embeddings are stored in a high-performance vector database like ChromaDB, which allows for rapid similarity searches based on a user's natural language query. The RAG process is further enhanced by an orchestration layer and a user-friendly interface that supports multi-modal features like Text-to-Speech and multilingual support. These features represent a significant advancement in information retrieval, providing a real-time decision support tool that reduces manual labor while ensuring

high factual accuracy and deeper engagement with complex datasets.

2 LITERATURE SURVEY

The evolution of document processing has transitioned from simple keyword matching to sophisticated cognitive systems capable of human-like reasoning and contextual understanding. This chapter explores the historical and technological progression of automated document analysis, focusing on the shift from traditional Information Retrieval (IR) to the modern Retrieval-Augmented Generation (RAG) framework. By examining previous research in Natural Language Processing (NLP) and the emergence of Large Language Models (LLMs), we establish the foundation upon which this project is built. The literature review highlights the persistent challenges in handling unstructured data and the innovative solutions developed to ensure factual grounding, data privacy, and multilingual support in interactive systems.

2.1 STATIC PDF READERS AND TRADITIONAL SYSTEMS

For decades, the standard for document interaction was defined by software that treated the Portable Document Format (PDF) essentially as a "digital printout." These Static PDF Readers were designed primarily for visual fidelity—ensuring that a document looked the same on every device—rather than for data accessibility or structural intelligence. In these environments, the document is a dead end for data; users are forced to manually navigate through hundreds of pages, relying on their own memory to connect disparate ideas. Because these readers do not "understand" the underlying text or its logical flow, they cannot assist with summarization, cross-referencing, or answering complex conceptual questions. This creates a "Static Information Barrier," where valuable data remains trapped in a non-responsive format, making the extraction of specific insights a labor-intensive and error-prone process.

Beyond the interface of the readers themselves, Traditional Information Retrieval Systems relied on basic statistical techniques such as the Term Frequency–Inverse Document Frequency (TF-IDF) model. These systems represent documents as numerical vectors based purely on the frequency of words within the text, assigning higher importance to terms that appear frequently in a specific file but rarely across a broader collection. While TF-IDF improved the speed of digital searching compared to physical filing, it suffered from severe functional constraints. Primarily, it lacked the ability to capture the contextual meaning of words and treated similar concepts with different spellings—such as "AI" and "Artificial Intelligence"—as entirely unrelated terms. These traditional systems were effectively "keyword-bound," creating an "Information Silo" where data remained unreachable if the user could not phrase a query in the exact language used by the author.

2.2 NEURAL NETWORK-BASED NLP MODELS

The introduction of neural networks marked a definitive turning point in Natural Language Processing (NLP) by enabling models to learn dense vector representations of text. A major breakthrough was Word2Vec, introduced by Mikolov et al., which generated word embeddings by analyzing the context in which words appeared. While this allowed words with similar meanings to be placed closer together in a mathematical space, the embeddings remained "static." This meant that a word like "bank" would receive the same representation regardless of whether it referred to a financial institution or a riverbank. This technical stage led researchers to explore contextual embeddings, which adapt a word's representation based on its immediate surroundings, providing a more nuanced understanding of language that began to mirror human cognitive processing. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks further advanced this by allowing models to remember information from earlier in a sequence. However, these models struggled with "long-range dependencies," meaning they often forgot the beginning of a long paragraph by the time they reached the end. This gap paved the way for the development of parallel-processing architectures that could handle entire documents with greater consistency.

2.3 ADVANCEMENTS IN TRANSFORMERS ARCHITECTURE AND LLMs

The development of the Transformer architecture, introduced in the research paper “Attention is All You

Need" by Vaswani et al. in 2017, revolutionized natural language processing. At the core of the Transformer is the Self-Attention Mechanism, which allows the model to analyze relationships between all words in a sentence simultaneously. Unlike previous sequential models, Transformers can "attend" to distant words that are grammatically or semantically linked, regardless of their position in the document. This allows for an unprecedented grasp of context, enabling the model to understand complex sentence structures and long-form narrative arcs within technical documents.

This breakthrough led to the rise of Large Language Models (LLMs) such as BERT, GPT, and LLaMA, which are trained on trillions of tokens to perform tasks including text generation, summarization, and translation. One of the most influential models, BERT (Bidirectional Encoder Representations from Transformers), captures deep semantic relationships by processing text in both directions. Despite their generative prowess, standalone LLMs face a critical bottleneck known as "hallucination." Since these models rely on internal weights learned during training rather than a real-time retrieval database, they may produce incorrect or outdated answers when asked about specific details they haven't "seen" recently. This necessitates a framework that can provide the model with a specific, external factual context during the response phase.

3 PROPOSED METHODOLOGY

The proposed system is an Interactive PDF Assistant designed to provide an intelligent and efficient way of interacting with PDF documents. It is based on a Retrieval Augmented Generation (RAG) architecture, which combines document retrieval with AI-based response generation. The system allows users to upload PDF files and ask questions in natural language, receiving accurate and context-aware responses. The system works by extracting text from the uploaded PDF and dividing it into smaller chunks for better processing. These chunks are converted into vector embeddings and stored in a vector database, enabling fast and relevant information retrieval. When a user submits a query, the system retrieves the most relevant content from the document and provides it as context to a local Large Language Model. The model then generates a response based strictly on this context, ensuring accuracy and reliability.

A key feature of the proposed system is the use of a local LLM through Ollama, which ensures that all processing is performed on the user's system. This eliminates data privacy concerns and allows the system to function even without an internet connection. Additionally, the system supports multilingual responses and includes text-to-speech functionality, enhancing accessibility and user experience.

Advantages of Proposed System

- **Context-Aware Responses:** The system uses a RAG framework to generate answers based on the actual content of the document, ensuring high accuracy and relevance.
- **Reduced Manual Effort:** Users can quickly obtain required information without reading the entire document, saving time and effort.
- **Enhanced Data Privacy:** By using a local LLM, the system ensures that user data remains on the local machine and is not shared with external servers.
- **Offline Capability:** The system can operate without continuous internet connectivity, making it more reliable and accessible.
- **Multilingual Support:** Users can interact with the system in multiple languages, improving usability for a diverse audience.
- **Improved Accessibility:** The integration of text-to-speech functionality allows users to listen to responses, making the system accessible to visually impaired users.
- **Intelligent Question Suggestions:** The system automatically generates relevant questions based on the document, helping users explore content more effectively.
- **Efficient Handling of Large Documents:** The use of vector databases and similarity search enables efficient processing and retrieval from large PDF files.

SYSTEM ARCHITECTURE

The system architecture is designed as a modular pipeline that integrates multiple components to enable seamless document interaction. The user interacts with the system through a Streamlit-based interface, where a PDF document can be uploaded and queries can be entered. Once a document is uploaded, the system extracts text using a PDF processing module and divides it into smaller chunks for efficient handling. These

chunks are then converted into vector embeddings using an embedding model and stored in a vector database. When a user submits a query, the system converts the query into an embedding and performs a similarity search to retrieve the most relevant chunks from the database.

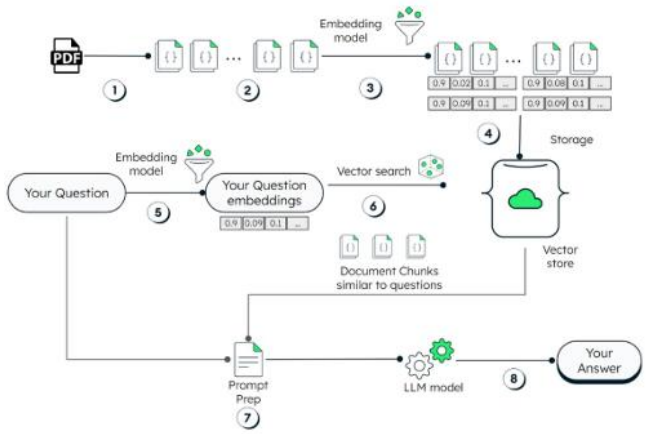


Fig 2: System Architecture

The retrieved content is passed as context to a local Large Language Model, which generates a response based on the provided information. The response is displayed to the user and can also be converted into audio using a text-to-speech module. This architecture ensures accurate, fast, and secure interaction with PDF documents.

1.1 Use case Diagram

The Use Case Diagram represents the interaction between the user and the system. The primary actor is the user, who uploads PDF documents and interacts with the system by asking queries. The system processes the document and generates responses based on user input. The main functionalities include uploading documents, asking questions, retrieving relevant information, generating responses, and converting output into speech. This diagram provides a high-level understanding of how users interact with the system.

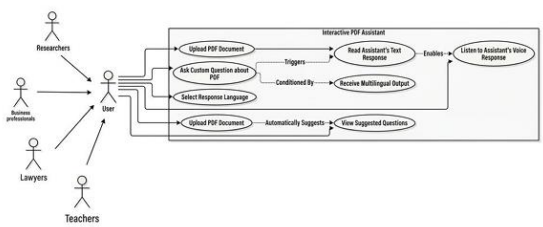


Fig 3: Use Case Diagram

1.2 Class Diagram

The Sequence Diagram shows the step-by-step interaction between system components over time. The process begins when the user uploads a PDF file. The system extracts text, splits it into chunks, and generates embeddings stored in the vector database. When a query is submitted, the system retrieves relevant content and sends it to the language model. The language model generates a response, which is then displayed to the user. Optionally, the response is converted into audio format. This diagram helps in understanding the dynamic workflow of the system.

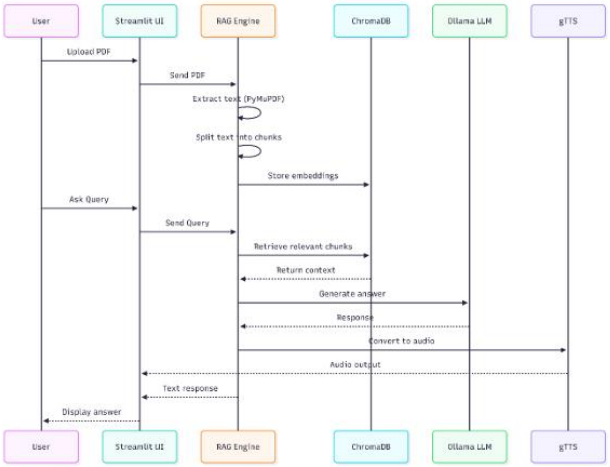
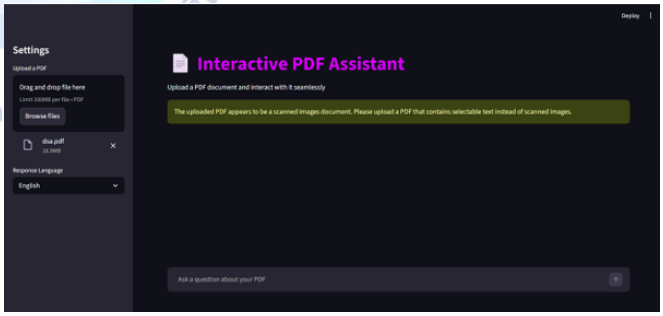


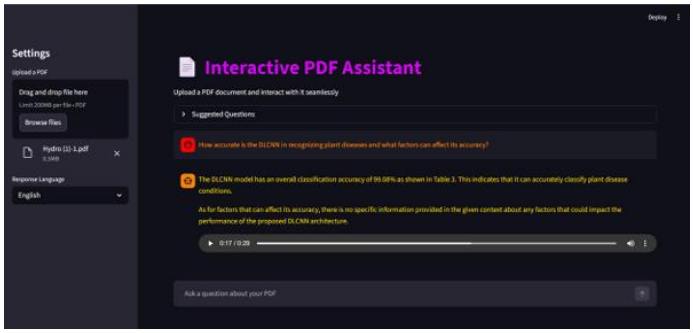
Fig 4: Sequence Diagram

4 RESULTS

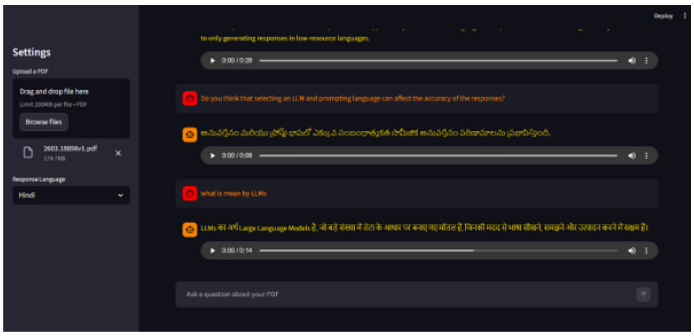
Uploaded Scanned Image PDF



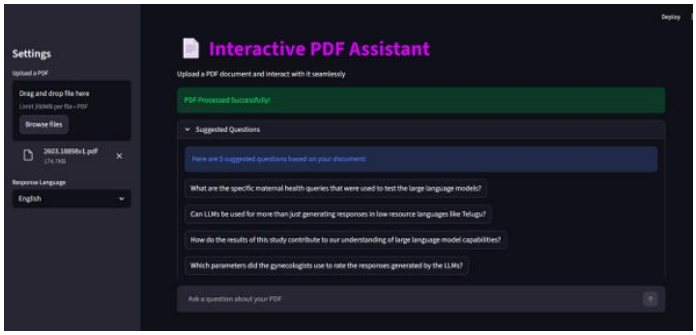
Response in User Preferred Language



Response in both Text and Audio



Suggested Questions Generation



5 . CONCLUSION

The proposed Interactive PDF Assistant presents an efficient and intelligent framework for transforming static PDF documents into an interactive conversational system. By integrating document processing techniques with a Retrieval-Augmented Generation (RAG) pipeline, the system enables users to extract meaningful information and receive context-aware responses. The use of text extraction, chunking, embedding generation, and similarity-based retrieval ensures that the responses are accurate and relevant to the document content. The system effectively addresses the limitations of traditional document reading methods by providing an automated and user-friendly approach to information retrieval. The integration of a vector database allows fast and efficient storage and retrieval of document embeddings, while the use of local Large Language Models through Ollama ensures data privacy and eliminates dependency on external cloud-based services. This makes the system more secure and suitable for handling sensitive documents. The implementation using Streamlit provides a simple and interactive interface, enabling users to upload documents, ask questions, and receive responses in real time. Additional features such as text-to-speech and suggested questions enhance usability and improve the overall user experience. The system demonstrates reliable performance across different test scenarios, including large documents and multiple queries, proving its scalability and effectiveness. Overall, the project delivers a scalable, privacy-focused, and intelligent solution for document interaction, making it useful for students, researchers, and professionals who need efficient access to information from large documents.

6 FUTURE ENHANCEMENT

The future scope of the Interactive PDF Assistant offers several exciting opportunities to enhance its capabilities and make it more intelligent, interactive, and impactful. One of the most promising directions is the integration of real-time collaborative features, where multiple users can interact with the same document simultaneously, enabling team-based analysis and discussions. This would transform the system into a collaborative knowledge platform rather than just an individual assistant.

Another interesting enhancement is the addition of advanced conversational memory, allowing the system to remember previous interactions and provide more personalized and context-aware responses. This would make conversations feel more natural and continuous, similar to interacting with a human assistant. Along with this, integrating multilingual understanding and translation capabilities would allow users to upload documents in one language and query them in another, significantly expanding the system's usability across different regions. The system can also be improved by incorporating visual understanding features, enabling it to interpret images, charts, and tables within PDFs. This would allow users to ask questions about graphical data and receive meaningful explanations. Additionally, integrating voice-based interaction, where users can ask questions through speech and receive spoken responses, would make the system more accessible and user-friendly.

Another advanced enhancement is the use of hybrid deployment models, combining local processing with optional cloud support to balance privacy and scalability. This would allow the system to handle larger workloads while still maintaining data security.

Finally, the system can evolve into a personalized knowledge assistant by learning user preferences, suggesting relevant questions automatically, and providing intelligent summaries of documents. These enhancements would make the system not just a tool for

document querying, but a powerful assistant for learning, research, and decision making.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Lewis, P., Perez, E., Piktus, A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020. https://arxiv.org/abs/2005.11401

[2] Brown, T., Mann, B., Ryder, N., et al. Language Models are Few-Shot Learners. NeurIPS, 2020. https://arxiv.org/abs/2005.14165

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL, 2019. https://arxiv.org/abs/1810.04805

[4] Reimers, N., & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP, 2019. https://arxiv.org/abs/1908.10084

[5] Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is All You Need. NeurIPS, 2017. https://arxiv.org/abs/1706.03762

[6] Touvron, H., Lavril, T., Izacard, G., et al. LLaMA: Open and Efficient Foundation Language Models. Meta AI, 2023. https://arxiv.org/abs/2302.13971

[7] OpenAI GPT-4 Technical Report. 2023. https://arxiv.org/abs/2303.08774

[8] Guu, K., Lee, K., Tung, Z., et al. REALM: Retrieval-Augmented Language Model Pre-Training. ICML, 2020. https://arxiv.org/abs/2002.08909

[9] Robertson, S., & Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 2009. https://dl.acm.org/doi/10.1561/1500000019

[10] Facebook AI Research FAISS: A Library for Efficient Similarity Search and Clustering of Dense Vectors. https://github.com/facebookresearch/faiss

International Journal for Modern Trends in Science and Technology IJMTST