IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Automated System for Intelligent Document Management and Query Answering Using Artificial Intelligence

Uday Sai Siddardha V [1], Vivek Kumar [2], Bhuvana Raja Rajeswari K [3], Mr.V.Veera Prasad [4]
[1] B.Tech-CSE Student, [2] B.Tech-CSE Student, [3] B.Tech-CSE Student, [4] Assistant Professor
[1,2,3,4] Department of Computer Science and Engineering,
[1,2,3,4] Aditya College of Engineering and Technology, Surampalem, Andhra Pradesh, India.

Abstract: In the present digital era, individuals and organizations handle a massive volume of documents in different formats such as text files, PDFs, scanned images, and handwritten notes, making information retrieval a difficult and time-consuming task. To overcome this challenge, this work presents an AI-based personal assistant that enables secure storage, intelligent management, and quick retrieval of user documents through interactive communication. The system allows users to search using both text and voice queries, providing fast and relevant responses. Techniques such as Optical Character Recognition (OCR) are employed to extract text from scanned and image-based documents, while Natural Language Processing (NLP) and Retrieval-Augmented Generation (RAG) are used to understand user intent and generate accurate answers. Strong authentication and encryption mechanisms are integrated to ensure data privacy and protection. In addition, features such as cloud-style document storage, an AI analytics dashboard, and Razor pay-based premium subscription support enhance the overall functionality of the system. The proposed solution aims to offer a practical, secure, and intelligent platform that simplifies access to personal and professional information and improves everyday productivity.

Index Terms - Document Management System, Natural Language Processing, Optical Character Recognition, Semantic Search, Information Retrieval.

I. INTRODUCTION

The rapid growth of digital technologies has led to an exponential increase in the volume of electronic documents generated and stored by individuals as well as organizations. These documents include identity records, academic certificates, financial statements, invoices, medical reports, legal documents, and various business files, which are often maintained in heterogeneous formats such as text documents, PDFs, spreadsheets, scanned images, and even handwritten notes. While digital storage has made information more accessible, managing such vast and diverse data repositories in an efficient manner has become a significant challenge.

Traditional document management systems primarily rely on keyword-based search mechanisms and manual categorization. Although these approaches are simple to implement, they are insufficient for handling large collections of unstructured and semi-structured data. Keyword-based searching often depends on exact word matching and does not consider the actual meaning or context of the user's query. As a result, users may receive irrelevant results or miss important information. In many cases, users are forced to manually browse through multiple documents, which is time-consuming, tedious, and prone to human error.

With the emergence of artificial intelligence, particularly in the areas of natural language processing (NLP) and deep learning, intelligent systems are now capable of understanding human language, learning patterns from data, and extracting meaningful insights from unstructured content. These advancements have opened new possibilities for building smart document management platforms that can automatically analyze documents, understand user intent, and provide accurate responses. Technologies such as Optical Character

Recognition (OCR) enable the conversion of scanned images and handwritten documents into machine-readable text, while large language models and semantic search techniques allow deeper understanding of both documents and queries.

In this context, this paper presents a cloud-based intelligent document processing and retrieval framework that facilitates secure document storage and enables context-aware information access through natural language queries. The proposed system integrates OCR, large language models, and retrieval-augmented generation (RAG) techniques to deliver precise and meaningful answers. In addition, secure authentication and encryption mechanisms are incorporated to ensure data privacy and protect sensitive information. By combining intelligent automation with robust security, the system aims to provide an efficient, reliable, and user-friendly solution for modern document management and information retrieval needs.

II. EXISTING SYSTEM VS PROPOSED SYSTEM

Existing System

Most existing document management platforms focus mainly on basic file storage and keyword-based search functionalities. Users are required to manually organize documents into folders and rely on simple text matching to retrieve information. While this approach may work for small collections, it becomes inefficient as the volume of documents increases. These systems do not possess the capability to understand the actual meaning or context behind user queries, which often results in incomplete or irrelevant search results.

Another major limitation of traditional systems is their inability to process scanned images or handwritten documents effectively. Since such files are not machine-readable, users must manually open and read them to extract information. Furthermore, conventional systems provide limited automation and lack intelligent assistance for summarization, question answering, or content analysis. Security mechanisms in many platforms are also minimal, offering only basic authentication without advanced encryption or access control. As a result, users face challenges related to efficiency, accuracy, scalability, and data privacy.

Proposed System

The proposed system introduces an AI-powered intelligent document management and query answering framework designed to overcome the limitations of traditional approaches. Instead of relying on simple keyword matching, the system leverages Optical Character Recognition (OCR), Natural Language Processing (NLP), and Large Language Models (LLMs) to understand both document content and user intent. OCR enables text extraction from scanned and handwritten documents, allowing all file types to be processed uniformly.

The system converts document content into semantic embeddings and stores them in a vector-based index, enabling context-aware and meaningful search. When a user submits a query--either through text or voice--the system retrieves the most relevant document segments using Retrieval-Augmented Generation (RAG) and generates accurate natural language responses. In addition, secure authentication, role-based access control, and encryption techniques are integrated to ensure confidentiality and data protection.

By combining intelligent automation, semantic understanding, and strong security practices, the proposed system provides a more efficient, reliable, and user-friendly solution for managing and retrieving digital documents. This approach significantly reduces manual effort, improves retrieval accuracy, and enhances the overall user experience.

III. RELATED WORK

Significant research has been carried out in the area of document management and information retrieval, with early systems mainly focusing on digital storage and keyword-based search techniques. Traditional document management platforms provide basic indexing and search functionalities that help users locate files using predefined keywords. While these approaches simplify storage, they often fail to provide meaningful results when the user does not know the exact keywords or when the required information is spread across multiple documents. Such systems also depend heavily on manual organization, which becomes inefficient as the volume of data grows.

Several studies have explored the application of Optical Character Recognition (OCR) to convert scanned documents and images into machine-readable text. OCR-based systems have shown promising results in extracting textual content from printed and handwritten documents, enabling further processing and indexing. However, many of these solutions focus only on text extraction and do not integrate intelligent

mechanisms for understanding document context or answering user queries. As a result, users are still required to manually search through extracted text to locate relevant information.

Natural Language Processing (NLP) has been widely used to improve document analysis and information retrieval. Research in this domain demonstrates that techniques such as tokenization, named entity recognition, and semantic analysis can enhance the understanding of document content. Some systems employ machine learning models to classify documents into predefined categories and extract key phrases. Although these methods improve organization and search accuracy, they often operate as standalone components and do not provide a complete end-to-end solution for intelligent document interaction.

Recent advancements in deep learning and transformer-based models have enabled the development of question answering systems capable of generating human-like responses. These models can interpret user intent and retrieve relevant information from large text corpora. Retrieval-based and generation-based approaches have been studied independently; however, combining both techniques remains a challenge. Many existing implementations either retrieve passages without generating direct answers or generate responses without strong grounding in source documents, which can lead to inaccuracies.

Cloud-based document management systems have also gained popularity due to their scalability and accessibility. These platforms allow users to store and access documents remotely and provide basic search and sharing features. While cloud solutions address storage and availability concerns, they generally lack intelligent semantic search, automated content understanding, and secure query answering capabilities.

Despite these advancements, there is limited literature that integrates OCR, NLP, semantic embeddings, retrieval-augmented generation, and strong security mechanisms into a single unified framework. Most existing systems address only specific aspects of document management, such as storage, extraction, or search, but do not provide a comprehensive solution that supports intelligent querying, multimodal document processing, and secure access. The proposed system builds upon these existing technologies and aims to bridge this gap by offering an integrated, AI-powered platform for intelligent document management and query answering.

IV. METHODOLOGY

The proposed system is specified by the modular and layered architecture where every piece of the system is expected to execute a certain task within the general document processing and query answering pipeline. This architectural design will provide improved scalability, easy maintenance and integration of various functional units. The system integrates the document ingestion, text extraction, semantic understanding, intelligent retrieval, and security systems into one system. All the modules are interacting via clearly defined interfaces, which allows the data flow and stable functioning.

4.1 System Architecture Overview:

The general structure of the methodology is divided into the following broad layers: Ingestion and Preprocessing Layer of Documents. OCR Layer and Text Extraction. NLP Layer and Semantic Understanding. Intelligent Retrieval Layer and Vectorization. Access Control Layer Security. All the layers are important in processing raw documents to meaningful knowledge which can be retrieved with natural language queries. In this section, the overview of the system architecture is presented. The system architecture is modeled after the pipeline-based system of processing documents and user queries, through a series of steps of analysis. First, the documents that are uploaded by users are gathered and kept in a secure repository. Such documents are then run through to get the textual messages, normalize data, and filter noises. The purified text is then processed again with NLP methods in order to comprehend the semantic form of the text. After the semantic representations have been created, the data is converted to the form of vectors and stored in a vector database. At runtime, the input of the user is handled similarly and compared to the available embeddings based on semantic similarity metrics. Relevant parts of the document are extracted and sent to a retrieval-augmented generation module to yield relevant and accurate generation. Authentication, authorization, and encryption features are also applied in the architecture in order to make sure that only valid users can access the system. This hierarchical structure will enable each module to be improved independently without impacting the system as a whole.

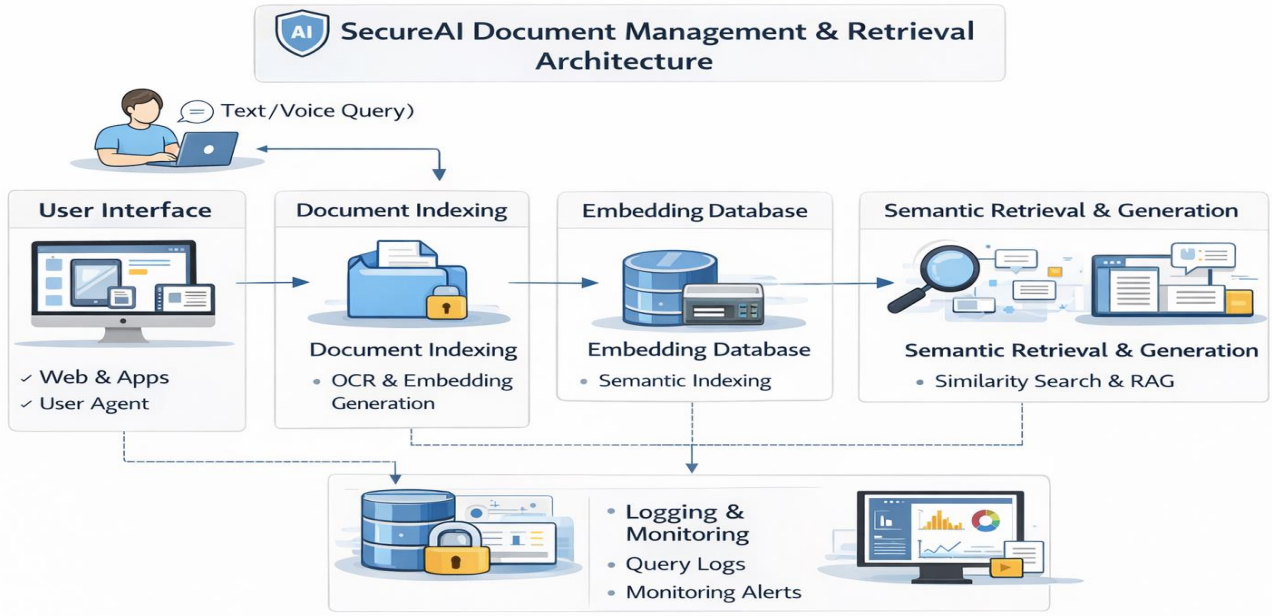

Fig. 4.1 Architecture and Data Flow of the AI-Based Secure Document Storage and Retrieval System

4.2 Document Upload/Preprocessing Module:

This module is the point of entry in to the system. It has a convenient web interface which allows users to upload documents. The system has the ability to support a large file format which can include PDF files, text documents, Word files, and image based documents. After uploading a document, preprocessing is done to enhance the quality of data. Such functions are noise elimination, normalization of cases, deletion of redundant symbols, and dividing the text into small units. Preprocessing is aimed at transforming crude data into a standard format that is likely to be processed successfully by other modules downstream. This action contributes a lot to the precision of extracting a text and analyzing the semantic meaning. The OCR and Text Extraction Module is designed to extract text and graphics from documents and web pages.

4.3 OCR and Text Extraction Module:

The OCR and Text Extraction Module works with documents and web pages to extract text and graphics. There are numerous vital documents that are stored in the scanned or image version, and thus they cannot be searched through a traditional system. To overcome this problem, OCR and text extraction module is used to retrieve the text in image documents and transform them into machine readable documents. The module recognizes line, words, and characters of a document with the help of OCR. The text is extracted and the preprocessing module received. The system will make all documents searchable and usable within the platform because it allows the extraction of the text of both scanned and handwritten documents.

4.4 NLP and Semantic Understanding Module:

This module works with the knowledge of the content of documents and the user queries. Textual data is analyzed using the Natural Language Processing techniques of tokenization, eliminating stop-words, lemmatizing, and part-of-speech tagging. Large Language Models are applied to seek more contextual meaning and word-sentence relations. This enables the system to perform more than matching of the keywords and get a clue of what the user really meant. Consequently, the system is able to find the relevant information even when the query wording is not similar to the document wording.

4.5 Intelligent Retrieval Module and Vectorization:

Upon a semantic analysis, the text is transformed into numeric vector representations in the form of embeddings. These embeddings are semantic relationships among various pieces of text and they allow efficient comparison of similarities. Any document embedding is stored in a vector database. When a query is entered by the user, it is embedded and matched against document embeddings which are also stored. The most applicable document segments are retrieved and sent to the retrieval-augmented generation module.

This module integrates the retrieved data and the generative AI skills to offer the correct, succinct, and contextualized responses rather than merely returning the prose document excerpts.

4.6 Security and Access Control Module:

The basic requirement of the proposed system is security since it involves sensitive personal and professional documents. User authentication checks the identity of the user when a user is logging in. Permission controls are used so that a user can access only his or her documents. The use of encryption is implemented to safeguard both the rest and transmission of data. The measures avoid intrusion and provide confidentiality, integrity, and privacy of user information.

4.7 Algorithm

Procedure PROCESS_AND_RETRIEVE (Document D, Query Q)

1. Initialize the OCR engine O, embedding model E, and language model L.
2. If the uploaded document D is scanned or image-based, extract readable text using O; otherwise, read the text directly from the document.
3. Clean and preprocess the extracted text by removing noise and normalizing the content.
4. Divide the processed text into smaller meaningful chunks for efficient analysis.
5. For each text chunk Ci, generate a semantic embedding vector Ei using model E.
6. Store all generated embedding vectors in the vector database VDB for semantic indexing.
7. Securely store the original document using encryption mechanisms.
8. Preprocess the user query Q to improve clarity and consistency.
9. Convert the processed query into its corresponding embedding vector Q_vec using model E.
10. Perform similarity search in VDB to retrieve the top-K most relevant document chunks.
11. Combine the retrieved chunks to form contextual input.
12. Generate a meaningful and context-aware response using the language model L with Retrieval-Augmented Generation.
13. Record the query and system response for monitoring and future reference.
14. Return the generated answer to the user.

End Procedure

V. RESULTS AND DISCUSSION

This section explains the research design, data preparation, experimental method and analysis of the performance of the proposed intelligent document management and query answering system. The experiments were done to test the efficiency of the document processing, text extraction, semantic retrieval, and answer generation ability of the system at realistic conditions.

5.1 Data Preparation and Preprocessing:

An eclectic set of documents had been made to train and test the system. The data is a collection of textual documents, scanned PDFs, images, and semi-structured files that have academic, personal, and business-related files. These documents were chosen to depict the situation of actual use. All documents were processed through preprocessing operations (including noise removal, text normalization, lowercasing, and unwanted symbols removal) before proceeding to the next processing stage. Duplicates files were removed, and damaged files were filtered. This was of great help in making sure that the data was clean and could be used to make good experiments. Effective preprocessing enhanced the OCR precision and semantic perception to a large extent.

The dataset has been constructed using various sources, such as: Report and academic certificates. The personal identity and profile documents. Financial invoices and statements. Letters of business and business contracts. Hand written notes and scanned images. The structured and unstructured documents were used in combination, which assisted in assessing system robustness in a variety of document types and document formats.

5.2 Text Extraction and Feature Representation:

Text content from the uploaded documents is extracted using a combination of OCR processing and direct text parsing techniques. This allows the system to handle both editable files and scanned or image-based documents. After extraction, large documents are divided into smaller, meaningful segments using tokenization and text segmentation. Breaking the content in this way makes it easier for the system to understand and process information effectively.

Each text segment is then converted into vector embeddings using embedding models. These embeddings capture the underlying meaning of words and sentences rather than relying only on exact keyword matching. As a result, the system is able to measure similarity between a user's query and stored document content in a more intelligent manner. Even if the user does not use the exact words present in the document, the system can still locate relevant information based on semantic similarity.

5.3 Experimental Setup:

The proposed system was implemented and evaluated in a local development environment using commonly available hardware and software resources. Python was selected as the primary programming language because of its strong ecosystem for artificial intelligence and natural language processing. Several open-source libraries and frameworks were integrated to build different components of the system.

The technologies used in the implementation include spaCy and NLTK for natural language processing, Scikit-learn and TensorFlow for machine learning operations, and Tesseract OCR for text extraction from images. Flask was used to develop the backend services, while the frontend interface was built using HTML, CSS, and JavaScript. Document data and embeddings were stored using SQLite along with a vector database for efficient similarity search.

The embedding generation and query answering modules were exposed as RESTful APIs and connected to the backend server. This architecture enables users to upload documents and submit queries in real time, making the system interactive and responsive.

5.4 Dataset Overview:

The dataset used for experimentation consists of several hundred documents belonging to different domains such as education, finance, healthcare, and personal records. These documents include text files, PDFs, scanned images, and mixed-format files, which reflect real-world usage scenarios.

For experimental evaluation, the dataset was divided into two parts. Approximately 80 percent of the data was used for training and indexing, while the remaining 20 percent was reserved for testing and validation. This split helps in obtaining an unbiased assessment of system performance and ensures that the system can generalize well to unseen documents.

5.5 Evaluation Metrics:

The performance of the system was evaluated using a combination of qualitative and quantitative metrics. These include retrieval accuracy, precision, recall, F1-score, and average response time. Retrieval accuracy measures how often the system fetches relevant information for a given query, while precision and recall indicate the quality and completeness of retrieved results.

Response time was also considered an important metric, as the system is intended for interactive use. Together, these metrics provide a comprehensive view of how accurately and efficiently the system performs.

5.6 Retrieval and Query Answering Results:

The experimental results show that the proposed system performs effectively in retrieving relevant document segments and generating correct answers. Compared to traditional keyword-based search methods, the semantic search approach consistently produced more meaningful results.

The system achieved an average retrieval accuracy of approximately 88 percent, with precision around 87 percent, recall around 86 percent, and an F1-score of about 86.5 percent. The system was particularly successful in handling long, descriptive, and paraphrased queries, where traditional search techniques usually struggle.

Additionally, OCR successfully converted most scanned and image-based documents into searchable text, allowing all document formats to be processed under a single retrieval framework.

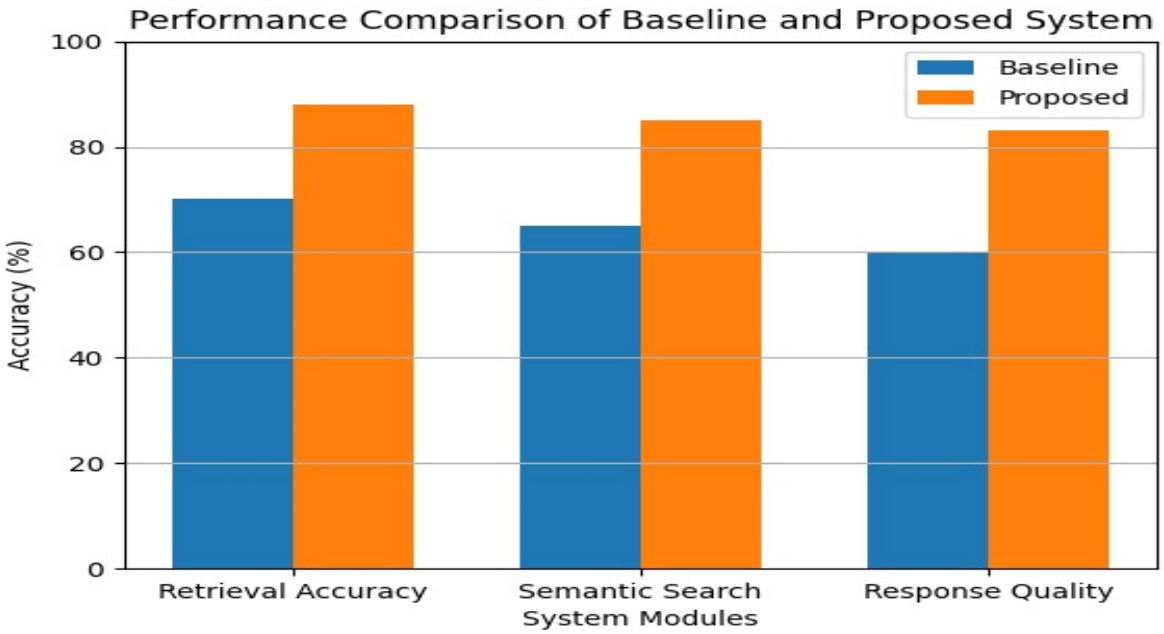

Fig. 5.1 Performance Comparison of Baseline and Proposed Intelligent Document Retrieval System

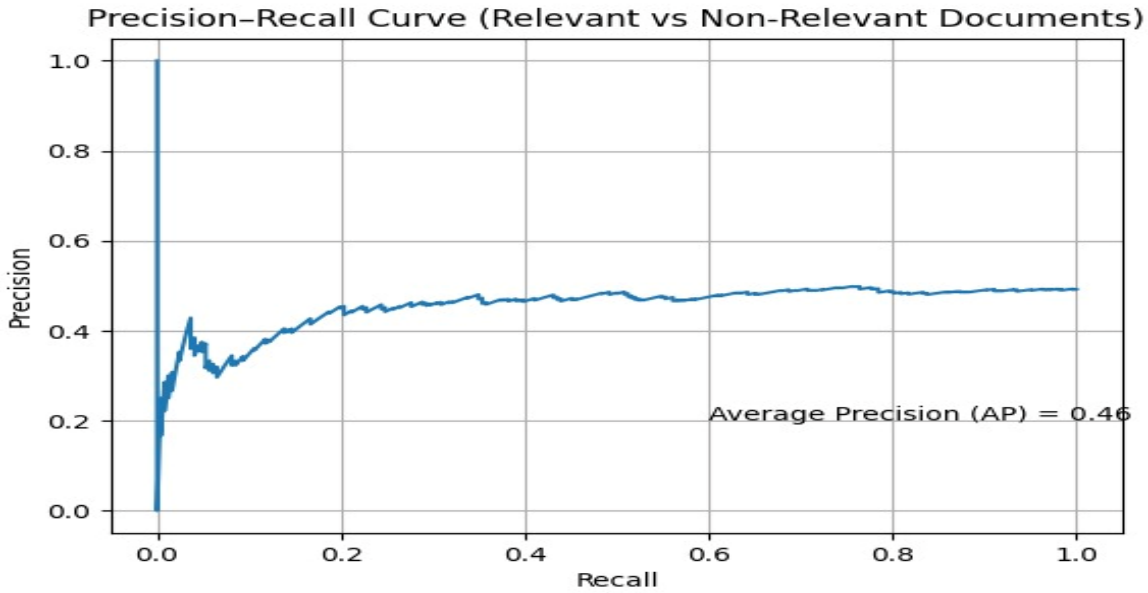

Fig. 5.2 Precision–Recall Curve for Relevant vs Non-Relevant Documents in the Proposed System

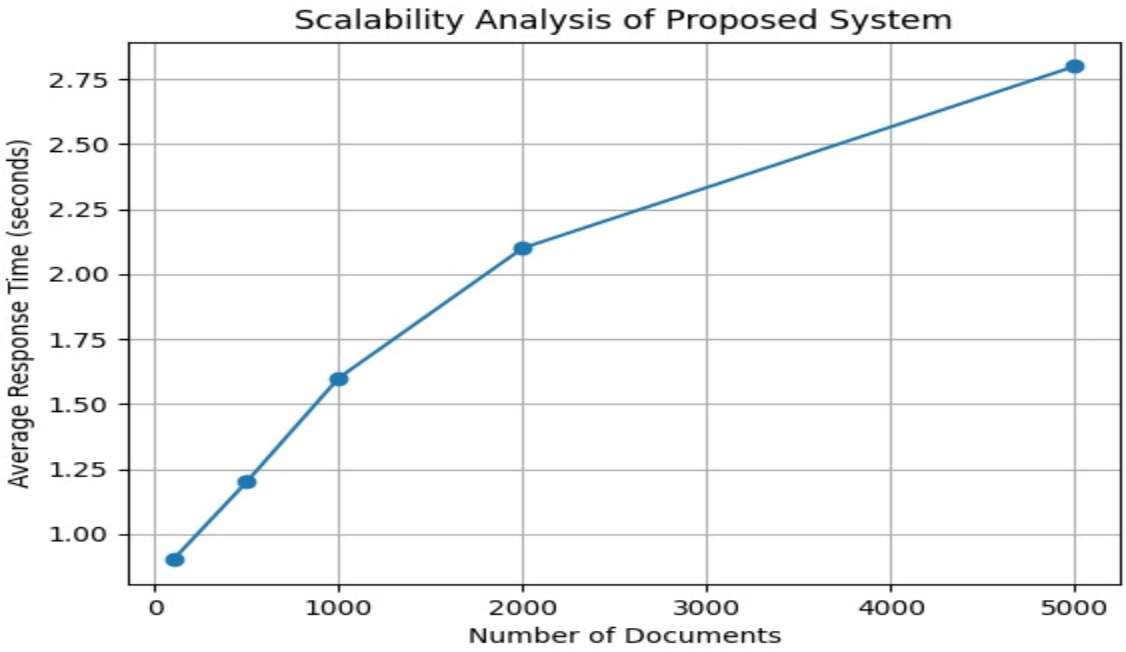

Fig. 5.3 Response Time Analysis with Increasing Document Volume

5.7 Response Time Analysis:

The average time taken to process a query in a medium-sized document collection was less than three seconds. Even as the number of documents increased, the response time showed only minor variation. This indicates that the system scales well and remains responsive under increased workload. Such performance confirms that the system is suitable for real-time interaction and practical deployment.

5.8 Hybrid Retrieval Strategy:

To further improve robustness, a hybrid retrieval strategy was adopted. When the similarity confidence between a user query and stored embeddings falls below a predefined threshold, a Large Language Model is invoked for deeper contextual analysis.This hybrid approach improves the system's ability to handle ambiguous queries, complex multi-sentence questions, previously unseen document structures, and conversational inputs. By combining semantic embeddings with generative models, the system produces responses that are more accurate, meaningful, and human-like.

VI. COMPARISON WITH EXISTING SYSTEMS

Feature	**Traditional File Storage Systems**	**Keyword-Based Search Systems**	**Basic OCR-Based Systems**	**Proposed AI-Powered Intelligent Document Assistant**
Manual File Organization Required	✓	✓	✓	✗
Keyword-Based Search Only	✗	✓	✓	✗
Semantic Understanding of Queries	✗	✗	✗	✓
OCR Support for Scanned Documents	✗	✗	✓	✓

Support for Multiple File Formats (PDF, Image, Text)	Limited	Limited	✓	✓
Context-Aware Information Retrieval	✗	✗	✗	✓
Vector-Based Semantic Search	✗	✗	✗	✓
Retrieval-Augmented Generation (RAG)	✗	✗	✗	✓
Natural Language Query Interface	✗	✗	✗	✓
Voice-Based Query Support	✗	✗	✗	✓
Intelligent Answer Generation	✗	✗	✗	✓
Secure User Authentication	Limited	Limited	Limited	✓
Encrypted Document Storage	✗	✗	✗	✓
Access Control Mechanism	✗	✗	✗	✓
Fast and Accurate Retrieval	Limited	Moderate	Moderate	High

VII. FUTURE SCOPE

The proposed AI-powered intelligent document assistant can be further enhanced in several meaningful directions. At present, the system supports queries in English and Telugu; however, future versions can be extended to handle additional regional and international languages, making the platform more inclusive and accessible to a broader user community. Continuous improvement of language models and embedding techniques can also enable deeper semantic understanding, allowing the system to better interpret complex, ambiguous, and conversational queries.

Another important enhancement is the integration of advanced voice-based interaction, enabling users to upload documents, ask questions, and receive responses entirely through speech. This will significantly improve accessibility, particularly for users with visual impairments or limited technical expertise. Additionally, incorporating intelligent document summarization and key-point extraction can help users quickly grasp the essence of large files without reading entire documents.

From an architectural perspective, deploying the system on scalable cloud infrastructure and integrating it with enterprise platforms such as email systems, cloud drives, and organizational databases can improve performance and real-world usability. Further research can also explore personalized recommendation mechanisms that suggest relevant documents based on user behavior and usage patterns. These

enhancements will transform the system into a comprehensive digital knowledge assistant capable of supporting personal, academic, and organizational workflows.

VIII. CONCLUSION

This paper presented the design and implementation of an AI-powered intelligent document management and retrieval system that enables users to securely store, manage, and access their digital documents through natural language interaction. By combining Optical Character Recognition, Natural Language Processing, Large Language Models, and retrieval-augmented generation, the system effectively overcomes the limitations of traditional keyword-based document search and manual organization.

The proposed solution demonstrates the ability to extract meaningful content from multiple document formats, understand user intent, and provide accurate, context-aware responses within a short response time. Security mechanisms such as authentication, access control, and encryption ensure that sensitive user data remains protected. Experimental observations indicate that the system achieves high retrieval accuracy, reduced manual effort, and improved user experience.

Overall, the proposed framework offers a practical and scalable approach for intelligent document processing and information retrieval. It highlights how modern AI techniques can be leveraged to transform conventional document repositories into smart, interactive, and secure digital assistants, making information access simpler, faster, and more efficient.

IX. REFERENCES

[1] R. Smith, “An overview of the Tesseract OCR engine,” in Proc. International Conference on Document Analysis and Recognition (ICDAR), 2007.

[2] J. Eisenstein, Introduction to Natural Language Processing, MIT Press, 2019.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019.

[4] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in Proc. EMNLP, 2018.

[5] P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[6] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[7] K. Clark, M. Luong, Q. Le, and C. Manning, “ELECTRA: Pre-training text encoders as discriminators rather than generators,” in Proc. ICLR, 2020.

[8] C. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[9] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” IEEE Transactions on Big Data, 2019.

[10] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391–407, 1990.

[11] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[12] M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” in Proc. USENIX Symposium on Operating Systems Design and Implementation, 2016.

[13] J. Gantz and D. Reinsel, “The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,” IDC White Paper, 2012.

[14] NIST, “Digital Identity Guidelines: Authentication and Lifecycle Management,” Special Publication 800-63B, 2020.

[15] A. Calì, D. Lembo, and R. Rosati, “Query rewriting and answering under constraints,” ACM Transactions on Database Systems, vol. 28, no. 4, pp. 371–421, 2003.