International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

SmartDocQA An Intelligent Document-Based Question Answering System

Krupa Sagar Reddy.C[1], Brahma Teja Reddy.P[2], Kesava Kalyan.S[3], Mohammad Basha.K[4], Dasta Giri.K[5], Anil.P[6]

[1] Assistant Professor, Department of Computer Science and Engineering (AI&ML), CBIT, Proddatur, YSR, A.P
[2–6] UG Students, Department of Computer Science and Engineering (AI&ML), CBIT, Proddatur, YSR, A.P
Corresponding Author E-mail: brahmateja1910@gmail.com

Abstract

Organizations manage a large volume of documents such as policy files, employee handbooks, technical manuals, and operational guidelines. These documents contain critical information required for daily decision-making. However, retrieving specific information from such large and unstructured documents using manual reading or keyword-based search methods is inefficient and time-consuming. Users often struggle to locate precise answers and frequently depend on administrative staff for clarification.

This paper proposes **SmartDocQA**, an intelligent document-based question answering system that allows users to ask questions in natural language and receive accurate answers directly from official documents. The system extracts text from uploaded documents, divides the content into meaningful sections, and converts them into semantic embeddings stored in a vector database. When a query is submitted, the most relevant document sections are retrieved using semantic similarity techniques, and a pre-trained large language model generates a clear response strictly based on the document content.

By ensuring document-grounded responses**, SmartDocQA** improves reliability, reduces manual effort, and enhances information accessibility across organizational environments within a centralized environment.

Keywords: Artificial Intelligence, Document-Based Question Answering, Semantic Search, Natural Language Processing, Embeddings, Large Language Models, Information Retrieval.

1. Introduction

In present-day organizational settings, critical information is largely maintained in digital documents such as policies, manuals, and procedural guidelines. As the volume and complexity of these documents grow, locating specific information becomes increasingly difficult for both employees and administrators. While manual reading of documents can yield accurate information, it is inefficient and impractical for frequent use. Similarly, traditional keyword-based search methods often fail to retrieve meaningful results because they rely on exact word matching and lack an understanding of user intent.

Although recent conversational AI systems can generate responses in natural language, they generally follow a decentralized usage model and are trained on broad, open-domain data rather than organization-specific documents. Consequently, the answers produced by such systems may not always align with official policies and can be inconsistent or unsuitable for formal decision-making.

To overcome these limitations, this paper introduces SmartDocQA, a centralized document-based question answering system designed to support reliable information access within organizations. The proposed system stores official documents in a shared repository and ensures that all user queries are answered strictly based on verified document content. For example, when an employee queries about leave policies or notice period requirements, SmartDocQA retrieves relevant information directly from the official policy document rather than generating generalized responses. By integrating semantic search techniques with document-grounded response generation, the system enhances accuracy, reduces manual effort, and improves overall organizational efficiency.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

2. Literature Review (Existing System & Proposed System)

2.1 Existing Systems

Manual Document Search:

In manual document search, users must go through documents line by line to locate the required information. Although this method can provide accurate results, it is inefficient, timeconsuming, and impractical when dealing with large volumes of documents.

Keyword-Based Search Systems:

Keyword-based search systems retrieve information by matching exact words or phrases entered by the user. While they reduce search time compared to manual reading, these systems lack semantic understanding and often fail to deliver relevant results when queries are phrased differently or use synonyms.

AI-Based Conversational Assistants:

AI-based conversational assistants can answer user queries in natural language but generally function in a decentralized manner, focusing on individual interactions rather than a shared organizational document source. Since these systems are not strictly restricted to verified organizational documents, the responses generated may not always be officially validated and can sometimes be inconsistent or inaccurate.

2.2 Need for the Proposed System

The increasing reliance on large organizational documents highlights the need for a centralized system that can provide accurate and consistent access to official information. Existing approaches often rely on individual or fragmented access methods, which can lead to inconsistent responses and repeated clarification requests. A centralized document-based solution ensures that all users query the same verified document source, improving reliability and efficiency. This requirement forms the basis for the design of the proposed **SmartDocQA** system.

3. Methodology

3.1 System Architecture

SmartDocQA follows a two-phase architecture: document processing and question answering. During document processing, uploaded PDF documents are analyzed to extract textual content. The extracted text is divided into smaller, meaningful chunks to preserve contextual information. These chunks are then converted into semantic embeddings and stored in a vector database.

During the question answering phase, the user query is transformed into an embedding and compared with stored document embeddings using similarity measures. The most relevant document section is retrieved and passed to a pre-trained large language model, which generates a context-aware response strictly based on the retrieved content.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

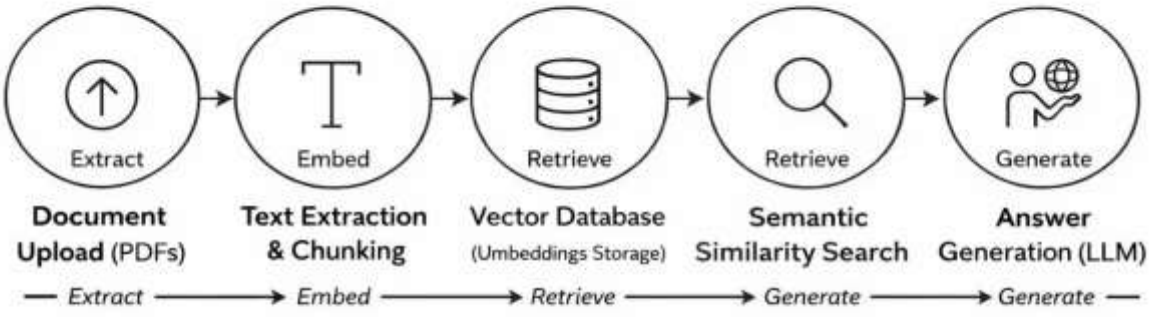

Fig. 1. Architecture of the SmartDocQA System

3.2 System Modules

Modules Used:

- **Document Upload Module:**
  This module enables authorized users to upload organizational documents such as policies and manuals in supported formats. The uploaded documents are then stored in a centralized repository for further processing and access.Handles the upload and management of organizational documents in supported formats.
- **Text Extraction Module:**
  The Text Extraction Module is responsible for extracting readable text from uploaded documents. The module is able to convert the pages of a document into plain text, making the document content amenable to automated analysis and processing.

- **Text Chunking Module:**
  This module breaks down the extracted text into smaller chunks of meaningful text. Chunking helps in increasing efficiency by allowing the system to process large documents in an organized way.
- **Embedding Generation Module:**
  The Embedding Generation Module is responsible for the generation of vector representations of text chunks, which carry their semantic meaning. This allows the system to compare the semantic meaning of user queries with the semantic meaning of document content.
- **Vector Database Module:**
  This module stores the generated embeddings and facilitates similarity-based retrieval. It assists in determining the most relevant document sections for the user queries.
- **Answer Generation Module:**
  The Answer Generation Module generates short and correct answers based solely on the content of the retrieved document. This is to ensure that the generated answers are always correct and consistent with the official information.

4. Results and Discussion

The SmartDocQA system was evaluated in its initial development phase by testing the document ingestion and content preparation pipeline using organizational policy documents. The primary objective of this evaluation was to verify the system's ability to load documents, extract textual content, and organize the extracted information into meaningful segments suitable for further processing.

As shown in **Fig. 2**, the system successfully loads the uploaded PDF document and identifies the total number of pages. The console output confirms that the document is processed without errors, demonstrating the reliability of the document ingestion module. This step ensures that organizational documents of varying length can be handled efficiently by the system.

Following document loading, the text extraction process is performed on each page of the PDF. The extracted text preserves the original content, including headings and policy descriptions, as observed in the sample output. This confirms that the text extraction module can accurately convert document content into machine-readable form, eliminating the need for manual reading.

After text extraction, the system applies sentence-based chunking to divide the document into smaller, meaningful sections. As illustrated in **Fig. 3**, the extracted text is segmented into multiple chunks, each containing a limited number of sentences. A total of seven chunks were generated for the test document, demonstrating effective content segmentation while preserving contextual information.

The chunking process plays a critical role in preparing the document for semantic retrieval by reducing document complexity and improving information organization. Instead of processing the entire document as a single block, the system enables focused retrieval of relevant sections during query handling.

Although semantic similarity computation and answer generation are still under development, the current results validate the feasibility of the proposed SmartDocQA architecture. The successful execution of document loading, text extraction, and chunk generation confirms that the foundational components of the system function as intended. The centralized document processing approach ensures that all extracted information originates from a single verified source, supporting consistency and reliability.

Overall, the results demonstrate that SmartDocQA effectively prepares organizational documents for document-based question answering and establishes a strong foundation for integrating semantic retrieval and answer generation in subsequent development phases.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

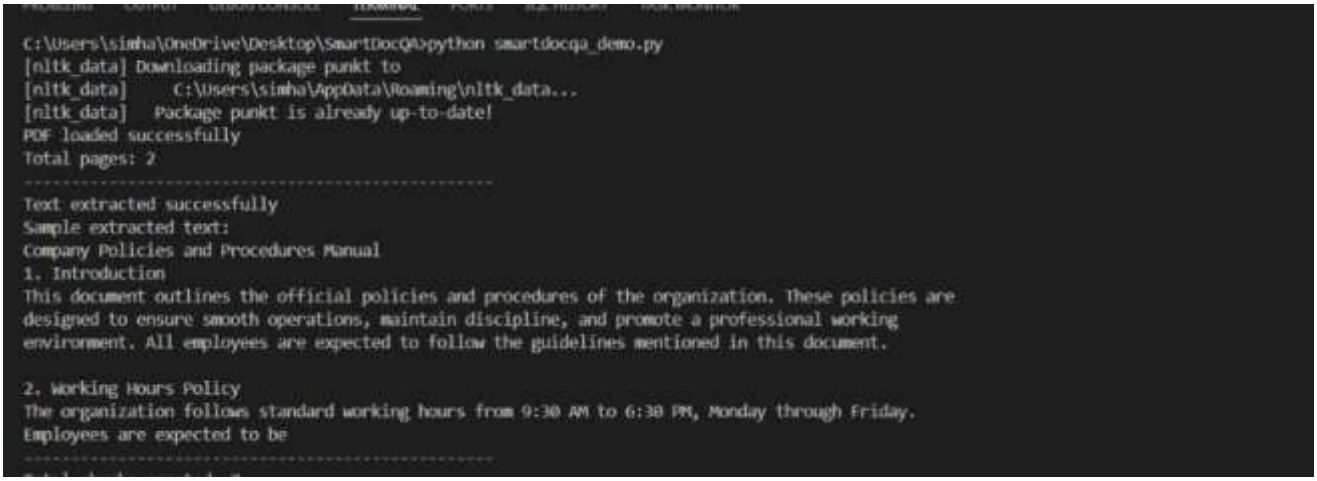

Fig. 2. Sample results of document loading

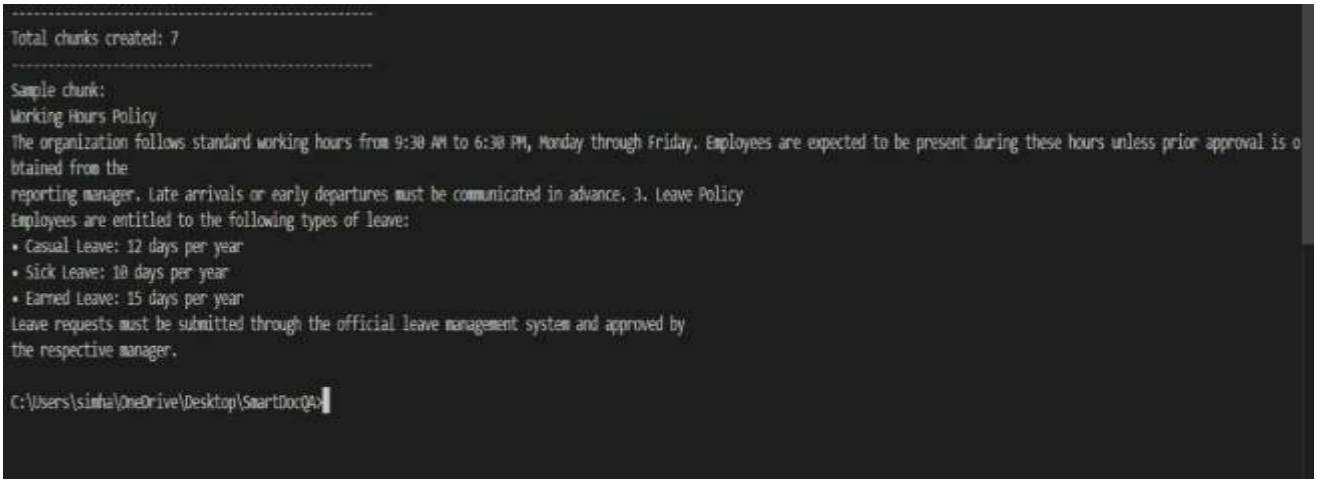

Fig. 3. Sample results of text extraction & chunking

Future Work and Assumed Results

Future work will focus on integrating semantic embedding generation and similarity-based retrieval to match user queries with relevant document chunks. Advanced similarity measures will be employed to improve retrieval accuracy for semantically similar queries. Additionally, a document-grounded answer generation module will be incorporated to produce concise and reliable responses strictly derived from official documents.

Further evaluation will include testing the system with larger document collections and multiple user queries to assess performance, scalability, and response accuracy. These enhancements are expected to further improve the effectiveness of SmartDocQA as a centralized document-based question answering system for organizational environments.

5. Conclusion

This paper presented SmartDocQA, an intelligent document-based question answering system designed to improve information retrieval from large organizational documents. By combining semantic search techniques with document-grounded answer generation, the system provides accurate and reliable responses to user queries based on verified document content. The proposed approach reduces manual effort, saves time, and enhances productivity, making it suitable for use in corporate, academic, and administrative environments through a centralized information system.

The current implementation demonstrates the feasibility of the document processing pipeline, including document ingestion, text extraction, and chunk-based content organization. These components form a strong foundation for semantic retrieval and reliable question answering. Future enhancements will focus on integrating advanced semantic similarity methods, improving answer generation capabilities, and evaluating system performance on larger document collections. Overall, SmartDocQA represents a practical and scalable solution for efficient document-based knowledge access in organizational settings.

References

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Pearson. https://web.stanford.edu/~jurafsky/slp3/

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf