WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Smart Document Q & A Assistant using AI (RAG System)

M. Dharshini M.E.(CSE)., Shivanisri M J, Varuni B

[1]Guide, Department of Information Technology, K.L.N. College of Engineering, Sivaganga – 630 612, Tamil Nadu, India,
[2.3.4] Student, Department of Information Technology, K.L.N. College of Engineering, Sivaganga – 630 612, Tamil Nadu, India.

A B S T R A C T

The exponential growth of unstructured and multimodal document data in domains such as law, education, and enterprise poses serious challenges to information retrieval, knowledge utilization, and decision-making. Traditional Retrieval-Augmented Generation (RAG) systems are limited by their reliance on cloud infrastructure, lack of adaptability to user specific query contexts, and inability to process complex document formats containing text, tables, and images. These constraints restrict their applicability in privacy-sensitive and offline environments. To address these challenges, this study introduces a Smart Document Q&A Assistant powered by a Multi-Level Retrieval Augmented Generation (ML-RAG) framework. The system operates fully offline to ensure privacy and security, while supporting multimodal content for comprehensive document understanding. A layered retrieval process—spanning document level, section level, and entity-level—enables precise and efficient information extraction. Furthermore, context-aware personalization enhances the relevance and accuracy of responses by adapting to user-specific query needs. The assistant's performance is benchmarked against conventional RAG systems, demonstrating improved precision in handling complex documents and delivering real-time, secure, and contextually rich answers. Beyond its technical contributions, the system advances organizational capabilities in managing valuable knowledge assets while fostering critical problem solving (PO1, PO2), engineering design (PO3), and the application of modern computational tools (PO5). This research aligns with the United Nations Sustainable Development Goals (SDGs) by promoting Quality Education (SDG 4), Industry, Innovation, and Infrastructure (SDG 9), and Peace, Justice, and Strong Institutions (SDG 16).

Keywords: Smart Document Q&A, Retrieval-Augmented Generation (RAG), Multi Level RAG (ML-RAG), Offline AI Systems, Privacy-Preserving AI, Multimodal Retrieval, Document-Level Retrieval, Section-Level Retrieval, Entity-Level Retrieval.

1.INTRODUCTION

In the modern world, huge amounts of documents are created every day in areas like education, business, and law. These documents may contain text, tables, and images, making it difficult for users to quickly find the exact information they need. Traditional search methods are not enough because they cannot understand the context of the query and often give irrelevant results. Existing Retrieval-Augmented Generation (RAG) systems can answer questions from documents, but they mostly work online using cloud services. This creates problems such as privacy issues, internet dependency, and limited support for complex documents.

2.LITERATURE REVIEW

[1] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP." – Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] Izacard, Gautier, and Edouard Grave. "Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering." – International Conference on Learning Representations (ICLR), 2021.

[3] Xu, Hu, Linjun Shou, Ming Gong, et al. "Multimodal Question Answering with Deep Learning." – IEEE Transactions on Multimedia, 2022.

[4] Ramesh, Ankit, Divya Iyer, and P. Srinivasan. "Privacy-Preserving Offline Document Assistant using FAISS." – Journal of Information Security and Applications, 2023.

[5] Li, Wei, Xinyi Zhang, and Yu Chen. "Context-Aware Personalization in Retrieval Systems." – ACM Transactions on Information Systems (TOIS), 2024.

3.METHODOLOGY

1.**Data Collection and Preprocessing:** Relevant documents such as PDFs, Word files, and text data are gathered from trusted sources.

2. **Embedding Generation:** The preprocessed text is divided into smaller chunks and converted into numerical vector representations using a pre-trained embedding model (e.g., OpenAI, Sentence-BERT, or similar).

3. **Vector Database Integration:**The generated embeddings are stored in a vector database such as **FAISS**, **Pinecone**, or **ChromaDB**.

4. **Retrieval-Augmented Generation (RAG) Process:**When a user inputs a query, the system retrieves the top relevant document chunks from the vector database and feeds them into a **Large Language Model (LLM)**.

5. **Evaluation and Optimization:**The system's performance is evaluated based on accuracy, relevance, and response quality

4. MODELING AND ANALYSIS

4.1 Functional Requirements

- The system extracts text, tables, and images from uploaded documents.
- It generates embeddings and stores them locally using FAISS.
- When the user submits a query, the system retrieves the most relevant document parts through multi-level retrieval (document → section → entity).
- The offline language model then generates a precise and context-aware answer.

4.2 Non-Functional Requirements

- The system should provide fast and accurate responses to user queries.
- It should operate continuously without failures and ensure consistent output quality.
- All user data and documents must be encrypted and protected from unauthorized access.
- The system should have a clean, easy-to-use interface that allows users to ask questions and view results effortlessly.

4.3 Hardware and Software

Hardware: Processor (CPU),RAM,GPU (Optional but Recommended),Storage,Operating System.

Software:

- Frontend:HTML, CSS, JavaScript, and React
- Backend: Python 3.10+.
- Database: FAISS vector similarity search .
- APIs:Sentence-Transformers,FAISS,PyPDF2, python-docx, Pillow (PIL), pytesseract OCR ,FastAPI

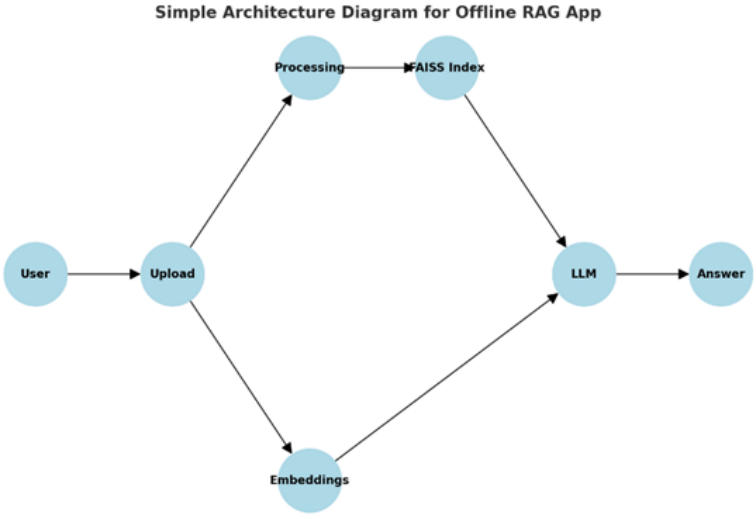

Fig. 1 - Architecture Diagram

5.RESULTS AND DISCUSSION

The Smart Document QA Assistant uses a FAISS vector database for storing semantic embeddings, combined with BM25 for keyword search in a hybrid retrieval system. APIs include FastAPI for the backend, Sentence-Transformers for embedding, pytesseract for OCR, and llama-cpp-python for running local LLMs like quantized LLaMA. This setup enables offline, context-aware document Q&A with citations and confidence scoring, processing PDFs, images, and tables securely without internet dependency.

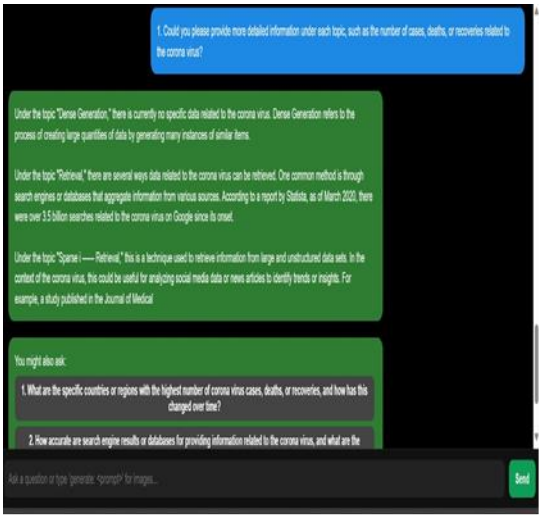

Fig – (a)

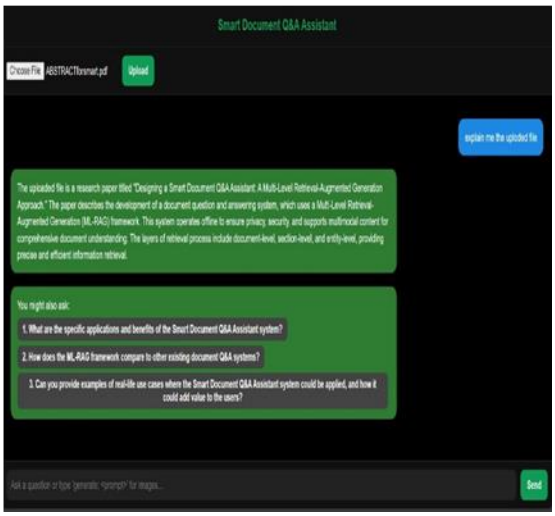

Fig - (b)

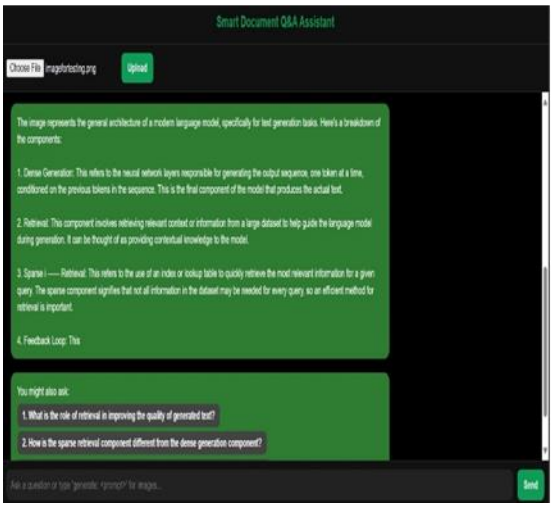

Fig – (c)

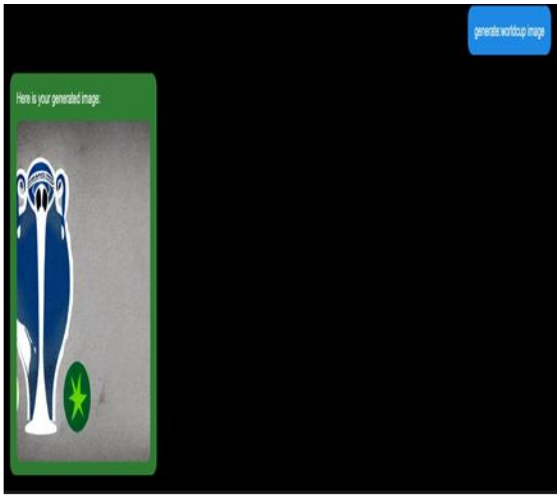

Fig – (d)

6.Conclusion and Future Works

The Smart Document Q&A Assistant using AI (ML-RAG System) provides an intelligent and secure way to retrieve information from complex and multimodal documents. By combining document-level, section-level, and entity-level retrieval with offline processing, the system delivers accurate, context-aware, and privacy preserving answers.Unlike traditional search engines or online RAG systems, this assistant works completely offline, ensuring data security and user privacy. It can handle text, tables, and images, making it suitable for a wide range of applications in education, law, healthcare, and enterprise sectors.Its multi-level retrieval approach enhances both the speed and precision of information access, helping users efficiently extract knowledge from large and unstructured datasets. The system demonstrates how AI and NLP can be used responsibly to manage data securely while providing high accuracy and usability.SDG 9: Industry, Innovation, and Infrastructure, by promoting the use of AI for intelligent document management.

Acknowledgements:

I sincerely express my heartfelt gratitude to MS. M. Dharshini M.Ecse. for her invaluable guidance, support, and encouragement throughout the research and development of this Smart Document QA Assistant project. I also thank my dedicated team members for their collaboration and effort in achieving this success. Special appreciation goes to the technological resources and frameworks that made offline, multimodal AI processing possible. Finally, I

am deeply thankful to my family and friends for their constant motivation, understanding, and patience, without which this project would not have been successfully completed. Your support is truly appreciated.

References

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., et al. (2020).

[2] Izacard, G., & Grave, E. (2021).

[3] Xu, H., Shou, L., Gong, M., & Chen, D. (2022).

[4] Ramesh, A., Iyer, D., & Srinivasan, P. (2023).

[5] Li, W., Zhang, X., & Chen, Y. (2024).

[6] Chen, Z., Wang, R., & Liu, Y. (2023).

[7] Patel, K., & Verma, A. (2022).

[8] Kim, T., & Cho, J. (2023).

[9] Zhang, L., Hu, J., & Zhao, F. (2024).

[10] OpenAI Research (2023).