International Journal of Scientific Research & Engineering Trends
Volume 11, Issue5 , Sep-Oct-2025, ISSN (Online): 2395-566X

Retrieval-Augmented Generation for Intelligent Question Answering from OCR-Processed PDFs

Ms.Usha Dhankar[1], Ms. Preeti Kalra[2], Ms.Agrima Samanotra [3], Mr.Aaditya Shriv Astava[4]
Department of Computer Science & Engineering, HMR Institute of Technology and Management, New Delhi, India

Abstract -This research explores the application of Retrieval-Augmented Generation (RAG) for enhancing information extraction and question-answering tasks from scanned PDF documents using Optical Character Recognition (OCR). By integrating a retrieval mechanism with a generative language model, we present a novel framework that intelligently interprets noisy, unstructured OCR outputs and enables contextual interaction via natural language queries[1][2]. The approach bridges the gap between image-based document archives and intelligent systems, facilitating improved document accessibility in fields like legal, academic, and archival research.

Keywords– Retrieval-Augmented Generation (RAG), Optical Character Recognition (OCR),PDFs

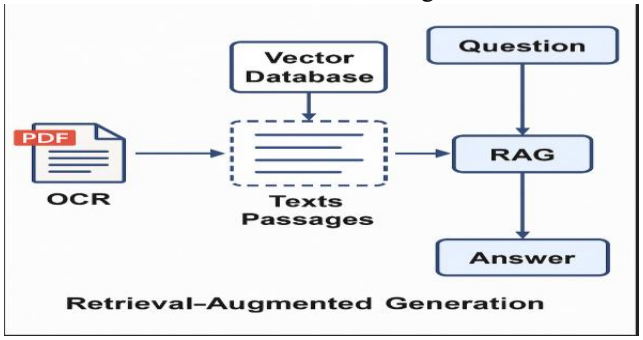
I. INTRODUCTION

PDFs are widely used for storing digital documents. However, scanned PDFs store content as images, making it difficult for traditional systems to extract and interpret the text. OCR systems like Tesseract provide a way to convert these images to machine-readable text[2]. Still, without semantic understanding, such systems fall short in supporting tasks like question answering. This gap is filled by Retrieval-Augmented Generation (RAG), which combines document retrieval with generative modeling to enable in-depth understanding[1].

RAG introduces a hybrid model that retrieves relevant text fragments and generates coherent, context-aware responses using transformers.

It is intended to enhance the LLM's performance by adding more details about a particular subject. By providing additional context and lowering ambiguity, this data aids the model in providing better answers to the queries.

Data Indexing and Data Retrieval and Generation are the two primary topics to concentrate on while developing a basic Retrieval-Augmented Generation (RAG) system. The system can store and/or search for documents as needed thanks to data indexing. The process of querying these indexed documents, extracting the necessary data, and using that data to generate answers is known as data retrieval and generation



PDFs are widely adopted for digital document storage due to their ability to preserve the original layout and formatting across devices. However, many PDFs, especially older or scanned ones, store content as images, making it challenging for traditional text processing systems to extract meaningful information. Such image-based PDFs require Optical Character Recognition (OCR) techniques to convert visual data into machine-readable text.OCR engines like Tesseract have advanced significantly, enabling the conversion of scanned images into editable and searchable text formats. However, OCR output often contains noise, errors, and lacks deeper semantic structure, which limits its usefulness for complex tasks such as question answering or knowledge extraction.

To overcome these limitatins, Retrieval-Augmented Generation (RAG) has emerged as a powerful solution. RAG frameworks combine a retrieval module — responsible for finding relevant text fragments — with a generative module that synthesizes coherent, context-aware responses. By integrating document retrieval with generative modeling, RAG systems enable deeper understanding and interaction with previously unstructured or noisy data.

In this work, we propose a RAG-based system that processes scanned PDFs using Tesseract OCR, semantically organizes the extracted text with Sentence-BERT embeddings, retrieves relevant chunks through FAISS vector search, and generates intelligent responses using large language models like GPT-3. This approach aims to make previously inaccessible or unstructured scanned documents interactive, searchable, and useful for fields such as academic research, archival studies, and legal document analysis.

II. PROBLEM STATEMENT

The proposed system uses OCR (Tesseract) to process scanned PDFs [2], Sentence-BERT to segment and embed the content [3], FAISS to obtain pertinent data [4], and a language model (GPT) to generate answers [1]. Academic research, archival digitization, and legal document assessment are some examples of applications. The main goal of this project is to create an intelligent pipeline that can analyze scanned PDF documents,

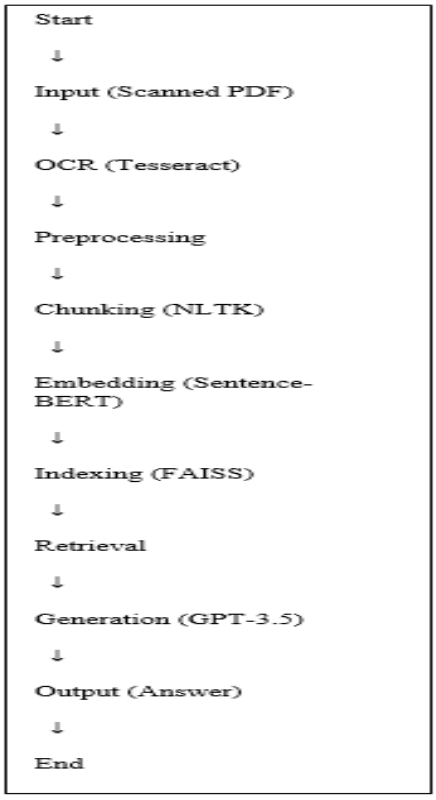
extract text using optical character recognition (OCR), arrange it semantically, and allow interactivity through natural language queries. The scope includes the following: Tesseract for OCR processing; Sentence-BERT for semantic chunking and embedding; FAISS vector search for retrieval; and big language models such as GPT-3 for natural language creation.

By making unstructured scanned text interactive and accessible, the technology seeks to serve use cases like corporate report evaluation, legal document analysis, and academic research.

III. IDEA AND PROPOSED SOLUTION

Five steps make up the RAG-OCR system: OCR text extraction [2], chunking, Sentence-BERT embedding [3], FAISS retrieval [4], and GPT answer generating [1]. This mimics a "open-book" architecture in which data is retrieved from memory prior to producing a response.

IV. METHODOLOGY



Data Preparation: Text is extracted from scanned PDFs using Tesseract OCR[2]. Denoising, line segmentation, and page alignment are examples of preprocessing.

Chunking: To maintain context, text is divided into a sliding window of three to five sentences with overlaps.

- Semantic continuity is guaranteed via heuristics and sentence tokenization (e.g., NLTK).

Embedding: Sentence-BERT preserves semantic similarity by transforming chunks into dense vector embeddings.

Retrieval: The vectors are indexed by FAISS [4], a quick similarity search library.

- Cosine similarity is calculated during a query in order to retrieve the top-N pertinent pieces.

Generation: To provide context-aware responses, GPT [1] (such as GPT-3.5-turbo) uses the query and the content that was retrieved as prompt input.

Even with unstructured OCR-derived text, our multi-stage pipeline allows for intelligent and effective document Q&A.

V. IMPLEMENTATION

Equipment Used Tesseract, an open-source OCR engine with multilingual support Chunking: a sentence tokenizer based on NLTK Sentence-transformers/all is embedded. For quick and tiny embeddings, use MiniLM-L6-v2. Facebook AI Similarity Search (FAISS) is the Vector Store. - Language Model: OpenAI's GPT-3.5 API Interface: RESTful service based on FastAPI Flexibility and scalability for adding multilingual OCR capabilities or changing out components are guaranteed by the modular design.

VI. RESULT AND EVALUATION

The system was assessed using a variety of metrics:

OCR Accuracy: Character-level accuracy on high-quality scans is approximately 94%; on older documents, it is lower.

Retrieval Precision/Recall: 85% of pertinent query fragments were retrieved with top-5 precision.

- Generative Quality: ROUGE-L of 0.63 and BLEU-4 score of 0.51 in comparison to reference solutions created by humans.

In terms of memory and semantic richness, the RAG-OCR system performed noticeably better than conventional keyword-based QA pipelines.

VII. CONCLUSION

This study shows how Retrieval-Augmented Generation and OCR together provide a strong foundation for analyzing and working with data from scanned documents. The solution bridges the gap between intelligent access and unstructured scanned content by using generative models for Q&A and converting noisy OCR outputs into semantically indexed vectors. Future additions might look into: - Support for several languages - Model optimization for particular domains (e.g., legal or medical texts)

For improved structure preservation, incorporate layout-aware models such as LayoutLM.

VIII. DECLARATIONS

Conflict of Interest: With regard to the publishing of this study, the authors disclose no conflicts of interest.

Funding: No specific grant from a public, private, or nonprofit organization was obtained for this research.

Author Contributions: Each author made an equal contribution to the study's conception, execution, composition, and evaluation.

Statement on Data Availability: All datasets utilized are either openly accessible or produced with open-source software (such as Tesseract and HuggingFace datasets).

We thank the maintainers of FAISS, OpenAI, and SentenceTransformers for their open-source contributions, as well as the creators of Tesseract OCR.

XI. ETHICAL CONSIDERATIONS

The Association For Computing Machinery (Acm) Code Of Ethics And Professional Conduct [10] And The Ieee Code Of Ethics [9] Both Provide Ethical Guidelines That Were Followed In The Conduct Of This Work. No Sensitive Or Personal Data Was Processed; All Of The Data Used Was Either Synthetically Generated Or Publically Available.

Important Ethical Factors Include:

Privacy: There Was No Use Of Personally Identifying Information.

Bias: By Testing On A Variety Of Document Types, Attempts Were Made To Reduce Model Bias.

Transparency: The Methods And Resources Are Freely Available And Replicable.

Responsible Ai Use: To Avoid False Information Or Hallucinations, Generative Answers Were Assessed.

X. STUDY LIMITATIONS

Notwithstanding Encouraging Findings, This Study Has a Number Of Shortcomings That Should Be Addressed In Subsequent Research:

Ocr Quality Variance: Performance Downstream Is Greatly Impacted By The Precision Of Ocr Output. Handwritten Or Low-Resolution Pdfs Generate Louder Text, Which Reduces The Efficiency Of Generation And Retrieval. Ocr Quality Is Still Quite Document-Dependent, Despite Tesseract's Advancements [2].

Limitations Of Generalization: Unless Refined With Customized Datasets, The Rag Model, Which Was Trained Mostly On General Knowledge Corpora, May Find It Difficult To Deliver Precise Answers For Extremely Domain-Specific Queries [1].

Despite Promising Results, This Study Contains Several Flaws That Should Be Fixed In Future Investigations:

Ocr Quality Variance: The Accuracy Of Ocr Output Has A Significant Influence On Downstream Performance. Low-Resolution Or Handwritten Pdfs Produce Louder Text, Which Decreases Generation And Retrieval Performance. Despite Tesseract's Improvements, Ocr Quality Is Still Very Document-Dependent [2].

Limitations Of Generalization: The Rag Model, Which Was Trained Mostly On General Knowledge Corpora, May Struggle To Provide Accurate Responses For Very Domain-Specific Questions Unless It Is Improved By Specialized Datasets [1].

Computational Constraints: Embedding Generation, Vector Indexing, And Inference From Large Language Models Require Substantial Compute Resources, Which May Limit Scalability Or Real-Time Applications [8].

Bias In Language Models: The Underlying Generative Model May Reflect Biases Present In Its Training Data. This Can Lead To Inappropriate Or Misleading Responses In Sensitive Or Ambiguous Cases [10].

XI. Acknowledgement

We Gratefully Acknowledge The Contributions Of Open-Source Communities And Tools That Made This Research Possible:

The Tesseract Ocr Team For Providing A Robust And Extensible Ocr Engine Used In Pre-Processing Scanned Documents [2].

The Sentencetransformers Team For Sentence-Bert, Enabling Efficient And Meaningful Sentence Embeddings [3].

The Developers Of Faiss For Making Large-Scale Similarity Search Practical And Efficient [4].

The Community Maintaining The Transformers Library And Huggingface, Which Simplified The Integration Of Advanced Retrieval And Generative Models [1].

We Also Thank The Ieee And Acm Organizations For Providing Accessible Codes Of Ethics [9][10] Which Guided Our Ethical Framework.

XII. COMPETING INTEREST

The Authors Declare That There Are No Competing Interests Influencing The Results Or Interpretations In This Study.

XIII. FUNDING SOURCE

No Specific Grant From A Public, Private, Or Nonprofit Organization Was Obtained For This Research.

XIV. WARNING FOR HAZARDS

Although This Research Involves No Physical Or Chemical Risks, We Identify The Following Computational And Ethical Hazards:

Misinformation Risk: Generative Models May Produce Confident But Inaccurate Responses, Especially When Working With Noisy Ocr Output [8].

Privacy Concerns: Use Of Ocr And Nlp Tools On Sensitive Documents Must Comply With Data Protection Regulations (E.G., Gdpr). We Used Only Public Or Anonymized Data.

Bias Amplification: Language Models May Inadvertently Propagate Biases Present In Their Training Data [10].

Overreliance On Ai: Users Should Not Use Ai-Generated Answers For Legal, Medical, Or Critical Decision-Making Without Human Oversight.

REFERENCES

1. Lewis, P., Et Al. (2020). Retrieval-Augmented Generation For Knowledge-Intensive Nlp Tasks. Arxiv:2005.11401
2. Smith, R. (2007). An Overview Of The Tesseract Ocr Engine. Icdar 2007
3. Reimers, N., & Gurevych, I. (2019). Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks. Arxiv:1908.10084
4. Johnson, J., Et Al. (2021). Open-Domain Question Answering With Bert And Faiss. Acl Workshop On Machine Reading
5. Vaswani, A., Et Al. (2017). Attention Is All You Need. Neurips 2017
6. Xu, Y., Et Al. (2020). Layoutlm: Pre-Training Of Text And Layout For Document Image Understanding. Kdd 2020
7. Radford, A., Et Al. (2019). Language Models Are Unsupervised Multitask Learners. Openai Blog
8. Radford, A., Et Al. (2019). Language Models Are Unsupervised Multitask Learners. Openai Technical Report.
9. IEEE (2020). IEEE Code Of Ethics. Retrieved From Https://Www.Ieee.Org/About/Corporate/Governance/P7-8.Html
10. ACM (2018). ACM Code Of Ethics And Professional Conduct. Retrieved From Https://Www.Acm.Org/Code-Of-Ethics