

MultiPDF RAG: An Intelligent Retrieval-Augmented Generation Framework for Multi-Document Question Answering

Ashay Abasaheb Vairat¹, Atharv Sarang Jadhav¹, Shantanu Ganesh Deshmukh¹, Ashwini Sawant²

¹Department of Computer Science & Engineering,

Karmaveer Bhaurao Patil College of Engineering, Satara, Maharashtra, India

²Project Guide, Department of Computer Science & Engineering, Karmaveer Bhaurao Patil College of Engineering, Satara, Maharashtra, India

Abstract

The rapid growth of digital documents in academic, legal, corporate, and technical domains has created significant challenges in efficient information retrieval and knowledge extraction. Traditional keyword-based search systems often fail to capture contextual meaning across multiple documents, leading to incomplete or irrelevant results. Recent advancements in Large Language Models (LLMs) have enabled more natural interaction with documents; however, these models suffer from hallucinations and limited context windows. This paper proposes a **MultiPDF Retrieval-Augmented Generation (MultiPDF RAG)** framework that enables accurate, context-aware question answering across multiple PDF documents. The system integrates document parsing, semantic chunking, vector embeddings, and similarity-based retrieval with a generative language model to produce grounded and reliable responses. By combining dense vector search with LLM-based generation, the proposed approach ensures factual accuracy, scalability, and efficient handling of large document collections. The framework is designed to be modular, cloud-deployable, and suitable for academic research, enterprise knowledge bases, and decision-support systems.

Keywords: Retrieval-Augmented Generation, MultiPDF RAG, Large Language Models, Semantic Search, Vector Databases, Document Question Answering, Natural Language Processing

1. Introduction and Motivation

The exponential increase in digital documentation has transformed how information is stored and accessed across industries. Research papers, legal contracts, technical manuals, and policy documents are increasingly distributed as PDF files, often spanning hundreds or thousands of pages. Extracting precise information from such large and heterogeneous document collections remains a significant challenge.

Traditional document retrieval systems rely primarily on keyword matching and rule-based indexing. While effective for simple queries, these systems fail to capture semantic relationships, contextual relevance, and cross-document dependencies. As a result, users must manually search, read, and synthesize information, which is time-consuming and error-prone.

Large Language Models (LLMs) have demonstrated impressive capabilities in natural language understanding and generation. However, when used in isolation, LLMs lack direct access to external documents and are prone to hallucinating incorrect information. Additionally, their limited context windows restrict the number of documents they can process simultaneously.

Retrieval-Augmented Generation (RAG) has emerged as a promising solution by combining information retrieval with generative modeling. In this paradigm, relevant document segments are retrieved from an external knowledge base and provided as context to the LLM during response generation. This approach grounds the model's output in factual data while maintaining linguistic fluency.

This research focuses on **MultiPDF RAG**, a specialized RAG framework designed to handle multiple PDF documents simultaneously. The proposed system enables users to upload multiple PDFs, index their contents semantically, and ask natural language questions that require information synthesis across documents. The system aims to improve accuracy, scalability, and usability for real-world document intelligence applications.

II. Literature Review

2.1 Document Retrieval Systems

Early document retrieval systems were based on keyword matching and statistical models such as TF-IDF and BM25. While computationally efficient, these methods lack semantic understanding and perform poorly for complex or ambiguous queries.

2.2 Semantic Search and Embeddings

With advancements in deep learning, dense vector representations using transformer-based models have significantly improved semantic search. Sentence embeddings allow documents and queries to be compared based on meaning rather than exact word matches, enabling more relevant retrieval.

2.3 Large Language Models for QA

LLMs such as GPT-style models have been applied to document question answering tasks. While they produce fluent responses, they often generate answers not supported by source documents, raising concerns about reliability in critical domains.

2.4 Retrieval-Augmented Generation (RAG)

RAG architectures integrate retrieval mechanisms with LLMs to ground generated responses in external knowledge. Studies show that RAG systems significantly reduce hallucinations and improve factual accuracy. However, many implementations focus on single-document or small-scale datasets.

2.5 Research Gap

Existing systems lack robust support for **multi-document PDF reasoning**, efficient chunking strategies, and scalable architectures suitable for large document collections. This paper addresses these limitations by proposing a structured MultiPDF RAG framework.

III. Problem Definition and Objectives

3.1 Problem Statement

Users often need to query information spread across multiple PDF documents. Manual searching is inefficient, and traditional search tools fail to provide contextual and synthesized answers.

Standalone LLMs are unreliable due to hallucinations and limited document access. There is a need for a system that can accurately retrieve and reason over multiple PDFs while ensuring factual grounding.

3.2 Objectives

- To design a system capable of processing and indexing multiple PDF documents.
- To implement semantic chunking and vector embedding for efficient retrieval.
- To integrate retrieval mechanisms with a generative language model.
- To enable accurate, context-aware question answering across documents.
- To ensure scalability, modularity, and cloud deployability.

IV. Proposed System Architecture and

Methodology

The proposed **MultiPDF Retrieval-Augmented Generation (MultiPDF RAG)** system is an intelligent document understanding framework designed to enable accurate, scalable, and context-aware question answering across multiple PDF documents. The system combines semantic retrieval techniques with Large Language Models (LLMs) to ensure that generated responses are factually grounded in source documents while maintaining natural language fluency.

The architecture follows a **modular, pipeline-based design**, allowing independent optimization of each stage and easy integration with cloud-based infrastructures. By separating document retrieval from response generation, the system minimizes hallucinations, overcomes context length limitations of LLMs, and supports large-scale document repositories.

4.1 PDF Ingestion and Text Extraction Module

This module is responsible for accepting multiple PDF documents from users and converting them into machine-readable text.

Functions and Features:

- Supports bulk upload of multiple PDF files in a single session.
- Extracts textual content using reliable PDF parsing libraries.

- Handles both digitally generated PDFs and scanned documents.
- Optional Optical Character Recognition (OCR) is applied to scanned or image-based PDFs.
- Retains structural elements such as headings, subheadings, paragraphs, and page numbers.
- Removes irrelevant metadata, formatting artifacts, and noise introduced during extraction.

Role in System:

This module ensures accurate and clean text extraction, which is critical for downstream semantic processing. Poor-quality extraction can propagate errors throughout the system; hence, preprocessing and validation are performed at this stage.

4.2 Document Chunking and Preprocessing Module

After text extraction, documents are segmented into smaller, semantically meaningful units called chunks.

Functions and Features:

- Splits extracted text into fixed-size chunks with overlapping boundaries to preserve context.
- Ensures that each chunk represents a coherent semantic unit rather than arbitrary text.
- Removes stop words, redundant whitespace, and irrelevant symbols.
- Assigns metadata to each chunk, including document ID, page number, and section title.
- Normalizes text to improve embedding consistency.

Role in System:

Chunking enables efficient retrieval by reducing the search space and ensuring that only relevant document segments are processed by the language model. Overlapping chunks help maintain continuity across paragraphs and prevent information loss.

4.3 Embedding Generation and Vector Storage Module

This module converts textual chunks into dense numerical representations that capture semantic meaning.

Functions and Features:

- Uses transformer-based embedding models to generate high-dimensional vector representations.
- Embeddings capture contextual similarity rather than keyword matching.
- Each chunk embedding is stored in a vector database for efficient similarity search.
- Metadata associated with chunks is stored alongside embeddings.
- Supports incremental updates when new PDFs are uploaded.

Role in System:

Semantic embeddings enable accurate retrieval of relevant information even when user queries use different wording than the original document text. This module forms the backbone of semantic search in the RAG pipeline.

4.4 Query Processing and Semantic Retrieval Module

This module processes user queries and retrieves the most relevant document chunks.

Functions and Features:

- Converts user queries into vector embeddings using the same embedding model.
- Performs similarity search (e.g., cosine similarity) against stored document embeddings.
- Retrieves top-k most relevant chunks based on similarity scores.
- Filters results using metadata such as document relevance or recency.
- Ranks retrieved chunks to maximize contextual relevance.

Role in System:

This module ensures that only the most relevant and contextually aligned document segments are passed to the generative model, improving accuracy and reducing unnecessary computation.

4.5 Retrieval-Augmented Response Generation Module

This module generates the final answer using a Large Language Model.

Functions and Features:

- Combines retrieved document chunks with the user query into a structured prompt.
- Ensures the LLM generates responses strictly based on retrieved context.
- Reduces hallucinations by grounding answers in verified document data.
- Supports concise or detailed responses based on user preferences.
- Can optionally include citations or source references.

Role in System:

This module bridges semantic retrieval and natural language generation, delivering accurate, readable, and context-aware responses.

V. Expected Performance

The performance of the proposed **MultiPDF Retrieval-Augmented Generation (MultiPDF RAG)** system is evaluated based on response time, accuracy, scalability, and reliability. The expected performance is derived from the system design and analysis of similar retrieval-based AI systems.

5.1 System Responsiveness

The system is designed to deliver fast and efficient responses to user queries.

User queries are converted into embeddings and matched against stored document vectors using optimized similarity search techniques. Since only the most relevant document chunks are processed by the language model, the overall

response time is reduced. The system is expected to return answers within a few seconds for standard multi-document queries.

5.2 Retrieval Accuracy and Answer Quality

The accuracy of the system depends on the effectiveness of semantic retrieval and contextual grounding.

Relevant document segments are selected based on similarity scores, ensuring that responses are generated using authentic document content. This approach improves contextual understanding and significantly reduces incorrect or hallucinated answers. The system is expected to provide precise and meaningful responses across multiple PDF documents.

5.3 Scalability and Load Handling

The proposed architecture supports scalable performance as the number of documents and users increases.

The use of vector databases allows efficient storage and retrieval of embeddings for large document collections. Cloud-based deployment enables horizontal scaling, allowing the system to handle multiple users and large volumes of PDFs without performance degradation.

5.4 Reliability and Availability

Reliability is an important performance criterion for document intelligence systems.

The modular architecture ensures that individual components operate independently, reducing the impact of failures. Cloud infrastructure provides high availability and stable system operation, ensuring consistent access for users.

5.5 Resource Efficiency

The system is optimized to use computational resources efficiently.

Only relevant document chunks are passed to the language model, reducing memory usage and processing cost. Embeddings are reused for repeated queries, further improving efficiency and minimizing redundant computation.

VI. Advantages and Limitations

The proposed **MultiPDF Retrieval-Augmented Generation (MultiPDF RAG)** system offers several benefits for intelligent document analysis while also having certain limitations that must be considered for real-world deployment.

6.1 Advantages

1. The MultiPDF RAG system provides accurate and context-aware question answering by combining semantic retrieval with language model generation.
2. The system supports multiple PDF documents simultaneously, allowing users to query information spread across large and diverse document collections. Semantic embeddings enable retrieval based on meaning rather than exact keywords, improving answer relevance and quality.
3. By grounding responses in retrieved document chunks, the system significantly reduces hallucinations commonly observed in standalone language models. The modular architecture allows easy scalability, maintenance, and future enhancements. Cloud-based deployment ensures accessibility, reliability, and support for concurrent users.

4. Additionally, the system is domain-independent and can be applied to academic research, legal analysis, enterprise knowledge management, and technical documentation. Efficient resource usage reduces computational cost by processing only relevant document segments.

6.2 Limitations

Despite its advantages, the MultiPDF RAG system has certain limitations.

1. The system requires substantial computational resources during the embedding and indexing phase, especially when processing large document collections. Initial document ingestion and vector generation may be time-consuming for very large PDFs.
2. The quality of generated responses depends heavily on the accuracy of text extraction and the effectiveness of the embedding model. Poorly scanned PDFs or noisy text can affect retrieval performance. Additionally, the system relies on stable internet connectivity for cloud-based operation.
3. The framework does not inherently verify the correctness of source documents, meaning that inaccurate or outdated content within PDFs may still be reflected in responses. Fine-tuning and domain adaptation may be required for highly specialized applications.

VII. Applications

The proposed **MultiPDF Retrieval-Augmented Generation (MultiPDF RAG)** system can be applied across multiple domains where large collections of documents need to be analyzed, queried, and understood efficiently.

7.1 Academic and Research Assistance

The system can be used by students, researchers, and educators to analyze multiple research papers, theses, and technical reports simultaneously. Users can ask conceptual or comparative questions across documents, enabling faster literature reviews and improved knowledge synthesis.

7.2 Enterprise Knowledge Management

Organizations can deploy the system to manage internal documents such as policies, manuals, reports, and standard operating procedures. Employees can retrieve precise information quickly without manually searching through large document repositories, improving productivity and decision-making.

7.3 Legal and Compliance Analysis

The MultiPDF RAG system is well-suited for legal document review, contract analysis, and compliance checking. It allows legal professionals to query multiple case files, agreements, or regulations and obtain context-aware responses grounded in source documents.

7.4 Technical Documentation and Support

The system can be applied to technical manuals, software documentation, and user guides. Engineers and support teams can quickly extract relevant technical information and troubleshoot issues by querying multiple documents at once.

7.5 Healthcare and Medical Documentation

Medical professionals can use the system to analyze clinical guidelines, research papers, and medical reports. The system helps retrieve accurate information across multiple documents, supporting clinical decision-making and research activities.

VIII. Future Work

The proposed **MultiPDF Retrieval-Augmented Generation (MultiPDF RAG)** system provides a strong foundation for intelligent multi-document question answering; however, several enhancements can be explored to further improve its capabilities and applicability.

8.1 Multimodal Document Support

Future versions of the system can be extended to handle multimodal content such as images, tables, charts, and diagrams present in PDF documents. Integrating vision-language models would enable more comprehensive understanding of complex documents.

8.2 Domain-Specific Optimization

The system can be fine-tuned for specific domains such as legal, medical, or financial applications. Domain-adapted embedding models and prompt strategies can improve retrieval accuracy and answer precision for specialized use cases.

8.3 Source Attribution and Citations

Future enhancements may include automatic citation generation, allowing the system to highlight exact document sources, page numbers, or sections used to generate responses. This would improve transparency and user trust.

8.4 Real-Time and Incremental Indexing

The system can be enhanced to support real-time document ingestion and incremental indexing, enabling immediate querying of newly uploaded PDFs without full reprocessing.

8.5 User Feedback and Learning Mechanisms

Incorporating user feedback loops can help refine retrieval relevance and response quality over time. Feedback-based ranking adjustments can improve system performance with continued usage.

8.6 Offline and Edge Deployment

Future work may explore lightweight deployments for offline or edge environments, making the system usable in low-connectivity scenarios or on local infrastructure.

IX. Conclusion

This paper presented a **MultiPDF Retrieval-Augmented Generation (MultiPDF RAG)** system designed to enable accurate, scalable, and context-aware question answering across multiple PDF documents. The proposed approach effectively addresses the limitations of traditional keyword-based search systems and standalone large language models by integrating semantic retrieval with generative AI.

By converting document content into semantic embeddings and retrieving only the most relevant document segments, the system ensures that generated responses are grounded in authentic source data. This significantly reduces hallucinations and improves the reliability of answers. The modular architecture allows efficient handling of large document collections while supporting scalability and cloud-based deployment.

The system demonstrates strong potential in domains such as academic research, enterprise knowledge management, legal analysis, technical documentation, and healthcare information systems. Its ability to synthesize information across multiple documents enhances decision-making, productivity, and knowledge accessibility.

Overall, the MultiPDF RAG framework provides a robust and practical solution for intelligent document understanding. With future enhancements in multimodal processing, domain-specific optimization, and source attribution, the system can further evolve into a comprehensive and trustworthy document intelligence platform suitable for real-world applications.

References

- A. Omar *et al.*, “Multi-Level Information Retrieval Augmented Generation for Visual QA,” *EMNLP*, 2024. [ACL Anthology](#)
- Z. Wang *et al.*, “Document Segmentation Matters for Retrieval-Augmented Generation,” *Findings of ACL*, 2025. [ACL Anthology](#)
- M. Kwon *et al.*, “A Dynamic-Selection-Based Retrieval-Augmented Generation Framework: Enhancing Multi-Document QA,” *Electronics*, 2025. [MDPI+1](#)
- Z. Liu *et al.*, “Benchmarking Retrieval-Augmented Generation in Multi-Document Settings,” *ACM*, 2025. [ACM Digital Library](#)
- T. T. U. Hoang & V. A. Nguyen, “PDF Retrieval Augmented Question Answering,” *arXiv*, 2025. [arXiv](#)
- M. Suri *et al.*, “VisDoM: Multi-Document QA with Visually Rich Elements Using Multimodal RAG,” *arXiv*, 2024. [arXiv](#)
- I. Besrouer *et al.*, “SQuAI: Scientific Question-Answering with Multi-Agent RAG,” *arXiv*, 2025. [arXiv](#)
- “A Systematic Review of Key Retrieval-Augmented Generation Components,” *arXiv*, 2025. [arXiv](#)
- M. Abo El-Enen *et al.*, “A Survey on Retrieval-Augmented Generation (RAG) Models,” *Springer*, 2025. [Springer Link](#)