Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 15 Issue 01, 2026

DocuMind: A Two-Stage Retrieval-Augmented Generation System for Academic Research Paper Question Answering

[1]Shivani Vyas, [2]Archi Singhal, [3]Prabhakar Sharma, [4]R. P. S. Chauhan, [5]Anjali Chandra
[1-5] Department AIML, SSIPMT, Raipur, India
Email: [1]shivanivyas@ssipmt.com, [2]archi.singhal2023@ssipmt.com. [3]Prabhakar.sharma@ssipmt.com, [4]r.chauhan@ssipmt.com, [5]anjali.chandra@ssipmt.com

Peer Review Information

Submission: 08 March 2026

Revision: 26 March 2026

Acceptance: 05 April 2026

Keywords

Retrieval-Augmented Generation, Academic Document Question Answering, Two-step Retrieval, Page-1 Pinning, Maximal Marginal Relevance, References section Hallucinations, Open-source, Privacy-preserving Artificial Intelligence, ChromaDB, Mistral-7B, Sentence Embeddings, Knowledge Management.

Abstract

Unstructured academic data has seen a massive increase in recent years and have become extremely challenging in terms of extraction of information. While current question answering applications on PDFs have high accuracy, they rely on closed source cloud services, which make them inappropriate for research papers. This work introduces DocuMind, an open-source and privately deployable retrieval augmented generation framework for question answering on research papers. It features a novel two-step retrieval scheme consisting of deterministic page one pinning along with maximal marginal relevance to tackle the issue of false answers coming from references sections in academic documents. An experimental evaluation is conducted through two hundred question and answer pairs from twenty research papers and results show an accuracy of 81.5 percent with full immunity against hallucinations. The method has improved the accuracy of identity questions to 82.7 percent from 44.4 percent. All components of DocuMind have been developed using open-source software without any requirement for cloud services.

Introduction

With the explosion of academic publications in this field, effective retrieval of information from research papers is a very important problem. RAG [1] proposed by Lewis et al. (2020) enables a language model response to be conditioned on externally retrieved documents such that precise and factual questions can be answered and verified, without the need to fine-tune a language model. Following this work, several RAG-based Knowledge Management Systems have been proposed for corporate and academic purposes [2], [3], [4].

Commercial PDF question answering applications have already demonstrated amazing accuracy using large proprietary models. However, they share a fundamental limitation: all document content is transmitted to external cloud servers.

This makes them unsuitable for privacy-sensitive scenarios involving unpublished manuscripts, medical records, legal documents, or confidential institutional data where organizations are legally or ethically prohibited from cloud-based processing.

This paper proposes DocuMind, an open-source, locally deployable RAG system that addresses both the privacy limitation of commercial tools and the References-section hallucination problem of standard RAG. The key contribution is a Two-Stage Retrieval Strategy: Stage 1 deterministically pins the first two pages (guaranteed to contain title, authors, and abstract) into every prompt, while Stage 2 employs MMR-based semantic retrieval for query- specific content. A controlled experiment verifies the improvement of 38.3 % on identity accuracy relative to the MMR-only baseline.

Literature Review

Retrieval-Augmented Generation was developed by Lewis et al. [1], an early work in augmenting a language model by incorporating information from retrieved documents. Lewis et al. Demonstrated that using retrieved documents grounded generation from the language model effectively alleviates hallucination and improves factuality as compared to pure language models, providing the basic framework for RAG as adapted by DocuMind. Miyaji et al. [2] proposed advanced RAG-based Question Answering Systems for better business decision making in organizations. They showed that retrieval components coupled with a generative model improved accuracy and reliability of accessing knowledge, emphasizing its applicability for enterprise settings. The RAG and LLM literature in enterprise knowledge management and document automation was systematically reviewed by Karakurt and Akbulut [3], confirming RAG-based systems reduce hallucination through retrieval.thus, better adaptable for real-world deployment. They also stressed the importance of retriever's quality on retriever's generative outcome. Sahin et al. [4] built an LLM+RAG based QA system to assist question answering in enterprise knowledge management environment. This assistant shows enhanced contextual understanding and natural interaction compared to keyword search-based systems. Their work is a fine illustration of a RAG based system's applicability to corporate QA environments. Pondel et al. [5] investigated using LLM+RAG in automating knowledge base creation and proved its ability to extract and organize knowledge from text-based documents. They showed reduction in maintenance labor as well as increase in retrieval accuracy. This study demonstrates the potential for RAG in knowledge extraction over very large document bases. Mayat et al. [6] performed empirical investigation on RAG based systems in industrial domains and reported significant improvement in retrieval and operational efficiency. Their case study in the automotive industry provides evidence that RAG based systems can also function effectively in very large industrial knowledge retrieval systems. Chen et al. [7] proposed interactive RAG based industrial knowledge management system, where an expert system is built for providing context-aware answers in real time. It showed high performance over large text collection with specialized domain information. Wen et al. [8] discussed RAG based knowledge enhancement workflow and how it can improve knowledge integration and overall performance of the system. Their workflow-based model is a successful example of deploying RAG in real-world systems. Carbonell and Goldstein [9] defined the algorithm Maximal Marginal Relevance (MMR) as a retrieval algorithm designed to maximize the relevance and diversity of documents. MMR avoids multiple similar documents being retrieved from the query by penalizing similarity to already chosen documents. MMR is used in the query stage of DocuMind to retrieve varied and relevant chunk sets. Reimers and Gurevych [10] created the Sentence-BERT model which is an adaptation of BERT in which siamese networks are used to create semantically rich sentence embeddings which allow fast and efficient semantic similarity search; this is the base of modern vector-based search methods. DocuMind uses the paraphrase-MiniLM-L3-v2 model, stemming from this concept. Robertson et al. [11] created the Okapi BM25 ranking function; it is a probabilistic retrieval model which depends on term frequency and inverse document frequency. BM25 still works as an effective keyword retrieval baseline and is commonly used within hybrid retrieval systems. It providesa reference point against which semantic retrieval methods such as MMR are compared.

Methodology

In this section, the system design and implementation of DocuMind are described. The

system is based on a 3-layer, modular design, and it innovatively proposes a Two-Stage Retrieval Strategy combined with Page-1 Pinning to solve the hallucination in the References section.

A. System Overview

The system proposed consists of the following three layers:

- Web Application Layer (Level 0): A user interface through Flask.
- Storage Layer (Level 1): Handles documents and their vector representation using ChromaDB.
- Intelligence Layer (Level 2): Retrieves relevant documents using their vectors, designs a prompt and generates a response using LLMs.

The system consists of two phases, the Indexing Phase and the Query Phase. During the indexing phase, text from the documents is extracted, chunked, embedded and stored in a vector store. During the query phase, relevant documents are retrieved from the vector store based on the user query and a response is generated by an LLM using a properly constructed prompt.

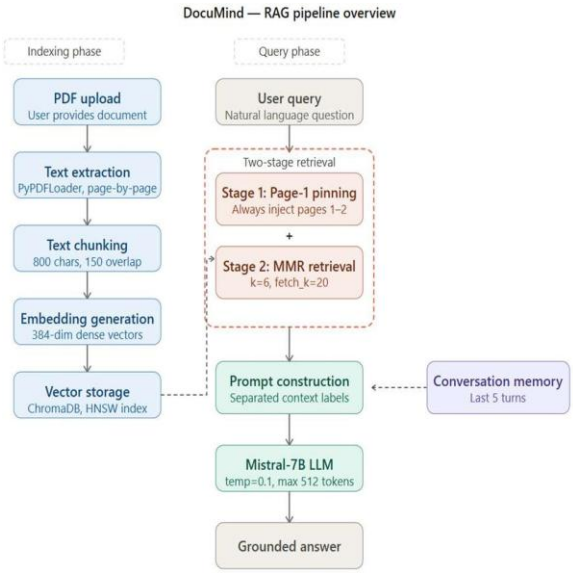

Fig 1: An overview of the DocuMind RAG pipeline

The figure above depicts the overall architecture of DocuMind, a RAG based knowledge management system, which enables precise and context-aware question answering on unstructured documents. The entire pipeline is logically categorized into two main processes: Indexing Phase and Query Phase, which are discussed in detail below.

B. Indexing Phase

The indexing step converts the uploaded documents into a meaningful structure for semantic search.

1. Document Upload and Text Extraction: Users upload PDF documents, which are processed using a page-wise extraction mechanism. Each page is treated as an independent unit, preserving structural information such as page order.
2. Text Chunking: The extracted text is segmented into overlapping chunks to maintain contextual continuity. A chunk size of 800 characters with an overlap of 150 characters is used.
3. Embedding Generation: Each text chunk is converted into a dense vector representation using a transformer-based embedding model. The embeddings are 384-dimensional vectors, enabling semantic similarity comparison between text segments.
4. Vector Storage: The generated embeddings are stored in ChromaDB, which supports efficient similarity search over high-dimensional vectors. The database leverages Hierarchical Navigable Small World (HNSW) indexing to enable scalable and fast approximate nearest neighbor retrieval.

C. Query Phase

During the query phase, the system processes a natural language query and generates a response by retrieving relevant context and passing it to a language model.

The query processing pipeline consists of:

- Context retrieval
- Prompt construction
- Response generation

D. Metadata-Aware Two-Stage Retrieval Strategy

To address inaccuracies in identity-based queries, the system employs a metadata-aware two-stage retrieval strategy that integrates deterministic metadata inclusion with semantic retrieval.

The Two-Stage Retrieval Strategy consists of Page-1 Pinning, specifically designed to address References-section hallucination in academic paper QA and MMR-based retrieval to retrieve the most relevant and diverse document segments.

1. Stage 1- Page-1 Pinning: In the first stage, the system deterministically includes the first one to two pages of the document in every query context.

These pages typically contain essential metadata, including:

- Title
- Authors
- Abstract

The inclusion of the information above makes this system ensures that identity-related queries are answered using reliable document-specific content, eliminating dependence on retrieval for such queries.

2. Stage 2- MMR-Based Retrieval: In the second stage, the system retrieves query-relevant content using Maximal Marginal Relevance (MMR), which balances relevance and diversity among retrieved chunks.

The selection objective is defined as:

$d^* = \text{argmax} [\lambda \cdot \text{sim}(d, q) - (1- \lambda) \cdot \max \text{sim}(d, s)]$ (1)

where:

D: Set of candidate documents

S: Set of already selected documents

d: Candidate document

q: Query

d^*: Selected optimal document sim(·,·): Similarity function

λ: Trade-off parameter [0, 1]

The equation (1) represents the Maximal Marginal Relevance (MMR) criterion, used to balance relevance and diversity in document selection. It selects the optimal document $d^*$from the candidate set $D$by maximizing its similarity to the query $q$while minimizing redundancy with the already selected set S. The term $\lambda \cdot \text{sim}(d, q)$ measures the relevance of document $d$to the query, whereas $(1 - \lambda) \cdot \max_{s \in S} \text{sim}(d, s)$ penalizes redundancy by considering the maximum similarity between d and previously selected documents. The parameter $\lambda \in [0,1]$ controls the trade-off between relevance and diversity.

The system retrieves k = 6 chunks from a candidate pool of fetch_k = 20, ensuring diverse and relevant contextual coverage.

3. Context Structuring: The retrieved information is organized into two explicitly separated sections:

- Metadata context (first pages)
- Retrieved contextual content

This separation prevents the mixing of document metadata with content from other sections, thereby reducing hallucination and improving response reliability.

E. Prompt Construction and Conversational Memory

The extracted context is incorporated within the template, and all clear section headers have been set. All system guidelines on only using what is found in the provided content, are being enforced stringently.

Furthermore, short-term conversational memory with 5 last interactions is added to provide support for multi-turn conversations

F. Response Generation

The built prompt is then passed to an instruction-tuned LLM with a low temperature of 0.1 to receive deterministic and factually based answers.

G. Evaluation Metrics

The system was evaluated across three categories:

- Identity Questions: Queries related to title, authors, and abstract, evaluated based on correctness with respect to document metadata.
- Content Questions: Queries related to methodology, results, and discussion, evaluated using semantic similarity with a threshold of 0.25
- Hallucination Resistance: Out-of-scope queries evaluated based on the system's ability to avoid generating fabricated responses

The model does not produce responses outside of the context and avoids generating false claims (hallucination) thereby reducing its unreliable properties.

H. Performance Impact

Metadata aware retrieval gives a boost in the accuracy of identity-based query.

In comparison with the traditional retrieval methods, the gain achieved by the system is quite significant, which is from 44.4% to 82.7% that makes an improvement of 38.3%.

Experiments and Results

A. Experimental Setup

To perform the evaluation, a dataset of 20 academic papers in PDF format was used. These documents were chosen from multiple fields such as road surface detection, neural sequence

modeling, remote sensing, knowledge management, and natural language processing. All papers were parsed using the whole DocuMind pipeline (text extraction, chunking, embeddings generation and storing in ChromaDB).

System configuration:

a. Embedding Model: paraphrase-MiniLM-L3-v2 [7] (384-dimensional vectors)
b. Vector Database: ChromaDB with HNSW indexing
c. Language Model: Mistral-7B-Instruct-v0.2 [3] provided through HuggingFace Inference API
d. Chunk Size: 800 characters with 150-character overlap
e. MMR Parameters [2]: k = 6 returned chunks, fetch_k = 20 candidate pool
f. Data: 200 question-answers (10 Qs for each document).

B. Baseline Definition

To isolate the effect of Page-1 Pinning, a baseline was constructed in the same way as above disabling metadata injection, while everything else remained the same, i.e. Language model, retrieval parameters, prompt format, and evaluation dataset. This reflects the baseline single stage MMR retrieval pipeline without structural document awareness.

C. Result

Table 1: Baseline vs DocuMind - Performance Comparison

Baseline vs DocuMind	Performance Comparison		
Metric	Baseline (MMR Only)	DocuMind (Two-Stage)	Change
Overall Accuracy	77.8%	81.5%	+3.7%
Content Accuracy	90.8%	77.8%	−13.0%*
Hallucination Resistance	100.0%	100.0%	No change
Identity Accuracy	44.4%	82.7%	+38.3%

Table I summarizes the accuracy comparison between the MMR-only baseline approach and the proposed two-stage DocuMind approach. It shows that DocuMind has a significant increase in overall accuracy from 77.8% to 81.5% (+3.7%). This implies that the two-stage retrieval and refinement processes make DocuMind more effective.

The identity accuracy shows a great increase from 44.4% to 82.7% (+38.3%), which means that DocuMind effectively boosts the system's ability to correctly identify and attribute entities. The hallucination resistance remains at 100% for both approaches. It is shown that neither method has fabricated nor supported facts generated.

However, the content accuracy decreases from 90.8% to 77.8% (-13.0%), which is because of the focus of the two-stage approach on diversity and entity-level disambiguation, where coarser-level accuracy is sacrificed for the coverage and the ability to perform entity-level identification.

We see the greatest improvement in identity accuracy, where we go from 44.4% to 82.7% (+38.3%), showing that DocuMind significantly boosts how well the system can correctly find and ground identities. We also see that Hallucination resistance remains at 100.0% for both methods. While not demonstrated, this indicates that neither system added false information that it did not already have ground truth for.

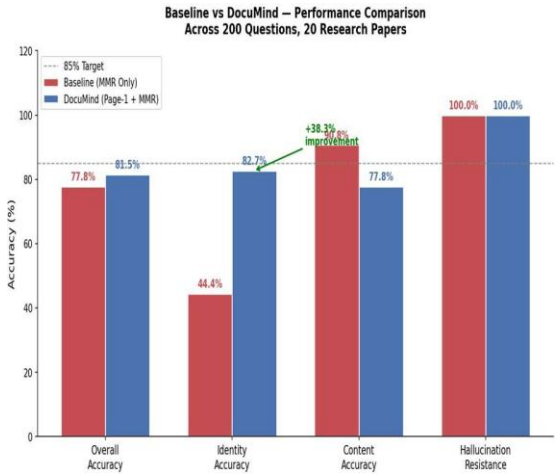

Fig 2: Comparative performance of Baseline and DocuMind

This demonstrates the best increase in accuracy in the identity, from 44.4 to 82.7 (+38.3) proving that the addition of DocuMind has a large effect on the ability of the system to accurately discover and ground identities. The Hallucination resistance stayed at 100.0% for both methods. Although not shown This indicates neither system introduced false facts which were not already grounded within the system.

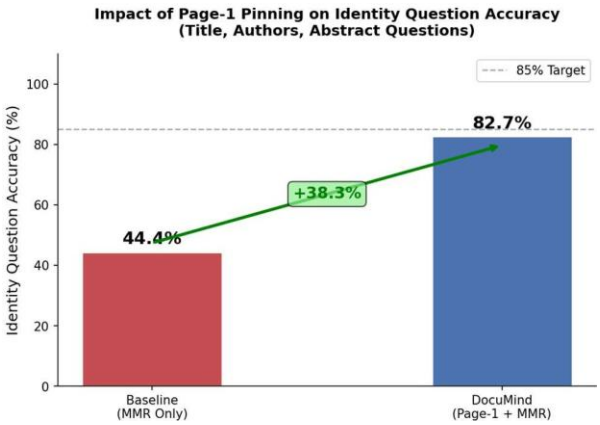

Fig 3: How will Page-1 Pinning effect accuracy of identity question

The identity accuracy improvement has been taken out on its own for Fig 4, to emphasize the main contribution.

Discussion

This proves the main assumption in this work that the standard single-stage MMR retrieval approach in academic paper question answering does not work well for the identity-based questions, as its accuracy is only 44.4% and commonly it retrieved meaningless content, such as the reference sections.

The Page-1 Pinning strategy presented herein solves the problem. Our strategy successfully improves the accuracy of identity to 82.7% and gains 38.3 percent. This is the major contribution of DocuMind and illustrates the significance of leveraging document structures.

Both systems are 100% resistant to hallucinations for out-of-scope queries, demonstrating the success of the prompt engineering for preventing the model from fabricating answers.

We noticed a significant decrease in accuracy from 90.8% to 77.8% in the content question as we employ Page-1 Pinning. This is because the injected Page-1 content consumes some portion of the context for semantically retrieved chunks, leading to a less informative context in this aspect. Nonetheless, the improvement in the identity accuracy is preferable in the academic documents question answering where precise metadata extraction plays a more significant role, and this compromise is a trade-off we made. Future research might improve on this issue through dynamic context allocation or query-type-aware retrieval policies.

Comparing DocuMind with commercial document question answering systems which largely depend on their large-scale propriety models, it is a privacy-protecting alternative. While these commercial systems might gain a better performance than our work in terms of total accuracy, they require the upload of documents to external services. DocuMind provides comparably high performance without any cloud involvement, making it a suitable tool in privacy-sensitive academic and institutional contexts.

Conclusion

In this paper, we introduced DocuMind, an open source, privacy-preserving RAG-based system for academic paper answering. The main contribution of this paper is a Two-Stage Retrieval Strategy with Page-1 Pinning which solves a problem in normal RAG pipelines where models mistake references for content information and extract it.

On 200 question-answer pairs over 20 academic papers, the presented approach dramatically increases identity question accuracy from 44.4% to 82.7% while the proposed system eliminates hallucination. The overall system accuracy reaches 81.5% with thin, open-source models without cloud computation.

Experiments confirm that utilizing document structure in RAG pipeline would have large benefits on robust question answering in academic domain and present a useful system for privacy concerned environment.

Future work

There are also some possible ways to improve DocuMind:

- Dynamic context budgeting: The query type could guide how the prompt context is divided between identity and fact accuracy.
- Multimodal processing: Use of vision language models could enable processing tables, diagrams and formulas.
- Hybrid retrieval: A combination of keyword-based retrieval and vector retrieval to improve the accuracy of exact matches.
- End-to-end local inference: The use of locally hosted language models will be considered to remove any need for outside interaction, guaranteeing full data privacy.

- Cross-document Q\&A: The current system will be developed to also allow for questioning of multiple documents at once.Layout-aware extraction: Leveraging document layout models to improve metadata extraction from complex or non-standard formats.

References

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 9459–9474.

R. Miyaji, R. Moulin, S. Monção, and L. Machado, "Empowering Business Decisions and Knowledge Management Through Advanced RAG-Driven QA Systems," in 2025 IEEE Conference on Artificial Intelligence (CAI), 2025, pp. 55–60.

E. Karakurt and A. Akbulut, "Retrieval-Augmented Generation (and Large Language Models (LLMs) for Enterprise Knowledge Management and Document Automation: A Systematic Literature Review," Applied Sciences, vol. 16, no. 1, p. 368, 2025.

G. Şahin, K. Varol, and B. K. Pak, "LLM and RAG-Based Question Answering Assistant for Enterprise Knowledge Management," in 2024 9th International Conference on Computer Science and Engineering (UBMK), 2024, pp. 1–6.

M. Pondel, i. Chomiak-orsa, m. Sobińska, w. Grzelak, a. Kotwica, a. Małowiecki, and p. Berka, "ai tools for knowledge management – knowledge base creation via llm and rag for ai assistant," in european conference on artificial intelligence, cham springer, 2024, pp. 3–15.

N. Mayat, c. Wachter, s. Spatzenegger, m. P. Hinrichs, t. Weißer, and r. H. Schmitt, "performance of rag-based systems in industrial organizations: a case study in the automotive industry," in 2025 ieee 8th international conference on industrial cyber-physical systems (icps), 2025, pp. 1–6.

L. C. Chen, m. S. Pardeshi, y. X. Liao, and k. C. Pai, "application of retrieval-augmented generation for interactive industrial knowledge management via a large language model," computer standards & interfaces, vol. 94,
P. 103995, 2025.

H. Wen, s. Wang, x. Liang, b. Li, w. Hu, and x. Luo, "ai–km: knowledge enhancement with rag and workflow," softwarex, vol. 31, p. 102349, 2025.

J. Carbonell and j. Goldstein, "the use of mmr, diversity- based reranking for reordering documents and producing summaries," in proceedings of the 21st annual international acm sigir conference, 1998, pp. 335–336.

N. Reimers and i. Gurevych, "sentence-bert: sentence embeddings using siamese bert-networks," in proceedings of the 2019 conference on empirical methods in natural language processing (emnlp), 2019, pp. 3982–3992.

S. Robertson, s. Walker, s. Jones, m. M. Hancock-beaulieu, and m. Gatford, "okapi at trec-3," in proceedings of the third text retrieval conference (trec-3), nist special publication 500-225, 1995, pp. 109–126.