Cross-Document Research Assistant using BM25 Retrieval and Controlled Language Models Review Paper

Authors: Prof. Dipmala Chaudhari, Prof. Ashwini Utture, Prof. Prachi Waghmare:
Associate Professors, Department of Computer Engineering, Nutan Maharashtra Institute of Engineering and Technology, Talegaon, Pune. Mr. Pratyush Jagtap, Mr. Varad Mule. Mr. Yashraj Gorde, Department of Computer Engineering, Nutan Maharashtra Institute of Engineering and Technology, Talegaon, Pune.

Abstract

With the rapid growth of online academic resources, it has become difficult for students and researchers to efficiently find relevant information and understand a topic without spending significant time on manual searching. Traditional search systems often return large volumes of results, many of which are not directly useful, leading to information overload. This project proposes a research assistant system that helps users explore research topics, retrieve relevant papers, and generate answers based on selected documents. The system uses the BM25 algorithm to rank research papers from sources such as arXiv, ensuring better relevance during retrieval. Users can then select a limited number of papers to form a focused research corpus. A large language model such as Gemini, developed by Google, is used to generate answers strictly based on the selected research content, reducing the chances of incorrect or unrelated outputs. The system aims to simplify literature review, improve retrieval accuracy, and provide reliable, context-based research assistance for students and researchers.

Keywords: Research Assistant, Information Retrieval, BM25 Algorithm, Academic Search, Cross-Document Question Answering, Natural Language Processing, Research Paper Analysis, AI-Based Learning System, Context-Based Answering.

1.Introduction

The digital transformation of education has resulted in a massive increase in the availability of online learning resources, including textbooks, lecture slides, research articles, and multimedia materials. While this abundance of information is beneficial, it also introduces challenges in retrieving precise and relevant knowledge efficiently. Traditional search engines and question answering systems primarily rely on keyword matching and textual retrieval, which often fail to capture the conceptual meaning of queries. Furthermore, educational content frequently includes visual elements such as diagrams, charts, and illustrations that play a crucial role in explaining complex concepts. Conventional QA systems are unable to interpret these visual components, resulting in incomplete or superficial answers.

A Students often need to consult multiple sources to understand a single academic concept. For example, understanding topics in biology, engineering, or data science may require reading textual explanations while simultaneously interpreting diagrams and charts. Existing QA systems treat these elements separately and lack the ability to integrate them into a unified response. Recent advancements in multimodal deep learning have demonstrated he potential of combining textual and visual representations to enable more comprehensive knowledge retrieval. Leveraging these technologies can significantly improve educational information systems by allowing them to reason across multiple modalities and documents.

2. Literature Survey

Paper 1

Title: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Author(s) & Year: Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., & Kiela, D. (2020)

Source: Neural Information Processing Systems (NeurIPS 2020)

This is the foundational RAG paper that introduced the concept of combining a retrieval mechanism with a generative language model. It forms the theoretical backbone of any system that retrieves document chunks and feeds them into an LLM to produce grounded answers exactly what your CrossDoc AI system does at its core.

Table 1 Literature Survey

Paper Title	Author(s) & Year	Focus	Key Contribution	Critique / Research Gap
Retrieval Augmented Generations for Knowledge-Intensive NLP Tasks.	Lewis et al. (2020)	Design Foundational RAG architecture.	First to combine dense retrieval with seq2seq generation for open-domain QA	Limited to single-model, no multi-document cross-referencing

Paper 2

Title: A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions

Author(s) & Year: Gupta, S., Ranjan, R., & Singh, S.N. (2024)

Source: arXiv:2410.12837

This paper presents a comprehensive study of RAG, tracing its evolution from foundational concepts to the current state of the art, combining retrieval mechanisms with generative language models to enhance output accuracy and addressing key limitations of LLMs.It serves as the primary literature survey reference for your project.

Table 2 Literature Survey

Paper Title	Author(s) & Year	Focus	Key Contribution	Critique / Research Gap
A Comprehensive Survey of RAG: Evolution, Current Landscape and Future Directions.	Gupta et al. (2024)	RAG evolution survey	Maps RAG from early retrieval models to modern LLM-integrated pipelines	Does not address multi-model routing or ensemble strategies

Paper 3

Title: Self-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection.

Author(s) & Year: Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024)

Source: International Conference on Learning Representations (ICLR 2024)

Self-RAG trains a model to dynamically decide when to retrieve, how many passages to retrieve, and how to critique its own outputs. Building on Self-RAG, several 2024 works taught models to decide when and how much to retrieve, improving accuracy without unnecessary retrieval calls.This directly informs your Chain-of-Thought and RAG+Rerank query strategies.

Table 3 Literature Survey

Paper Title	Author(s) & Year	Focus	Key Contribution	Critique / Research Gap
Self-RAG: Learning to Retrieve, Generate and Critique Through Self-Reflection.	Asai et al. (2024)	Adaptive retrieval & self-critique.	Model learns when to retrieve and critiques its own outputs dynamically.	Single model only; no cross-document or multi-model comparison.

Paper 4

Title: A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models.

Author(s) & Year: Fan, W., Ding, Y., Ning, L., et al. (2024).

Source: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024)

This systematic literature review analyzing 77 high-quality studies evaluates how RAG and LLM technologies address enterprise knowledge management challenges, finding that 63.6% of implementations utilize GPT-based models and 80.5% rely on standard retrieval frameworks such as FAISS or Elasticsearch. It benchmarks the technology stack choices for your project.

Table 4 Literature Survey

Paper Title	**Author(s) & Year**	**Focus**	**Key Contribution**	**Critique / Research Gap**
A Survey on RAG Meeting LLMs	Fan et al. (2024)	Enterprise RAG deployment	Benchmarks 77 studies on RAG+LLM for enterprise knowledge management.	80.5% rely on FAISS only; hybrid search and multi-model routing underexplored.

Paper 5

Title: MAIN-RAG: Multi-Agent Filtering Retrieval-Augmented Generation

Author(s) & Year: Li, J., et al. (2025)

Source: Proceedings of ACL 2025, Association for Computational Linguistics

MAIN-RAG proposes a multi-agent RAG framework consisting of three LLM agents — a Predictor, a Judge, and a Final-Predictor — to identify and filter noisy retrieved documents, improving overall answer quality across multiple QA benchmarks. This directly supports your ensemble voting and multi-model routing architecture.

Table 5 Literature Survey

Paper Title	Author(s) & Year	Focus	Key Contribution	Critique / Research Gap
MAIN-RAG: Multi-Agent Filtering Retrieval-Augmented Generation	Li et al. (2025)	Multi-agent document filtering	Three-agent pipeline (Predictor, Judge, Final-Predictor) filters noisy documents.	Agents use same base model; no heterogeneous multi-model ensemble.

Paper 6

Title: Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation.

Author(s) & Year: Abootorabi, M.M., Zobeiri, A., Dehghani, M., et al. (2025)

Source: ACL 2025 Findings

This survey offers a structured and comprehensive analysis of Multimodal RAG systems, covering datasets, benchmarks, metrics, evaluation methodologies, and innovations in retrieval, fusion, augmentation, and generation across multiple modalities including text, images, audio, and video. It guides the multi-format document ingestion design of your system

Table .6 Literature Survey

Paper Title	Author(s) & Year	Focus	Key Contribution	Critique / Research Gap
Ask in Any Modality: A Comprehensive Survey on Multimodal RAG	Abootorabiet al. (2025)	Multimodal RAG survey	Covers text, image, audio, video retrieval across all modality combinations.	Cross-modal alignment remains an open challenge; no unified benchmark.

Paper 7

Title: A Comprehensive Survey on Multimodal RAG: All Combinations of Modalities as Input and Output.

Author(s) & Year: Zhang, R., Liu, C., Su, Y., et al. (2025).

Source: TechRxiv

This paper conducts a comprehensive survey of multimodal RAG covering almost all combinations of modalities as input and output, presenting a taxonomy of MM-RAG methods and identifying four essential stages of the workflow, summarizing common approaches and discussing optimization strategies for each modality. It informs the heterogeneous document handling architecture of your system.

Table 7 Literature Survey

Paper Title	Author(s) & Year	Focus	Key Contribution	Critique / Research Gap
A Comprehensive Survey on Multimodal RAG: All Modality Combinations	Zhang et al. (2025)	MM-RAG taxonomy	Taxonomy of all input-output modality combinations with 4-stage workflow.	Lacks evaluation on real heterogeneous enterprise document datasets.

Paper 8

Title: Retrieval-Augmented Generation (RAG) for Enterprise Knowledge Management and Document Automation: A Systematic Literature Review

Author(s) & Year: Preprints.org Research Group (2025).

This study presents a systematic literature review analyzing 77 high-quality primary studies to evaluate how RAG and LLM technologies address practical enterprise challenges, formulating nine research questions targeting platforms, datasets, algorithms, and validation metrics to map the current landscape of enterprise RAG deployment. It directly maps to your enterprise document intelligence use case.

Table 8 Literature Survey

Paper Title	Author(s) & Year	Focus	Key Contribution	Critique / Research Gap
RAG for Enterprise Knowledge Management: A Systematic Literature Review	Preprints.org (2025)	Enterprise RAG SLR	Nine research questions covering platforms, datasets, algorithms, and metrics	Identifies significant lab-to-market gap; production readiness remains limited

Paper 9

Title: Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers.

Author(s) & Year: Multiple Authors (2025)

Source: arXiv: 2506.00054

This survey covers granularity-aware retrieval patterns, addressing retrieval precision by optimizing the unit of retrieval from full documents to fine-grained semantically aligned segments, along with retrieval-guided generation strategies that modulate generation based on retrieval metadata, task-specific cues, and agentic decision-making.It directly informs the chunking and reranking pipeline of your system.

Table 9 Literature Survey

Paper Title	Author(s) & Year	Focus	Key Contribution	Critique / Research Gap
RAG: A Comprehensive Survey of Architectures, Enhancements & Robustness Frontiers	Multiple Authors (2025)	RAG architecture robustness	Granularity-aware retrieval, query reformulation, and retrieval-guided generation.	Robustness against adversarial and counterfactual documents needs more work.

Paper 10

Title: Hybrid Retrieval, Agentic Orchestration, and Multimodal Search for Enterprise AI.

Author(s) & Year: Data Nucleus Research (2026).

Source: Data Nucleus.

This enterprise guide covers how RAG has evolved rapidly with graph-aware retrieval, agentic orchestration, and multimodal search, covering governance frameworks including ISO/IEC 42001 controls and GDPR compliance, making it a practical foundation for secure, ROI-driven workplace AI. It informs the security and compliance design of your system.

Table 10 Literature Survey

Paper Title	Author(s) & Year	Focus	Key Contribution	Critique / Research Gap
RAG in 2025: Hybrid Retrieval, Agentic Orchestration & Multimodal Search	Data Nucleus (2026)	Enterprise RAG governance	Covers hybrid retrieval, agentic RAG, GDPR and ISO/IEC 42001 compliance.	Security against poisoning attacks and adversarial documents still evolving.

3. Analysis of Existing Systems and Identified Gaps

3.1 Analysis of Existing Systems

1. **Single-Model Dependency:** Most existing systems are tightly coupled to a single underlying language model. Users have no ability to route queries to alternative models, compare responses, or fall back to a different model when one underperforms. This creates a critical single point of failure — if the model hallucinates, produces a biased answer, or fails on a specific document type, the user has no recourse within the same system.

2. **Isolated Document Processing:** Existing tools process documents in isolated silos. A query can only be answered from within a single notebook, project, or index at a time. There is no mechanism to cross-reference information across multiple unrelated document collections simultaneously, which severely limits the ability to draw unified insights from distributed knowledge sources.

3. **No Multi-Model Comparison:** None of the existing systems provide a native mechanism to run the same query across multiple LLMs at once and present their responses side by side. Users have no way to evaluate model agreement, identify conflicting interpretations, or score response

reliability all of which are critical for high-stakes decision making in legal, financial, and medical domains.

4. **Lack of Unified Interface:** Developer-focused frameworks require significant engineering effort to build even a basic user interface, while consumer tools offer rigid, non-customizable interfaces. There is no existing solution that combines a production-ready UI with a flexible multi-model backend forcing teams to either sacrifice usability or flexibility.

3.2 Identified Gaps

1. **Limited File Format Support:** Most existing tools are optimized for a narrow set of document formats, primarily PDF and plain text. Support for heterogeneous formats such as XLSX, PPTX, CSV, and MD within the same retrieval pipeline is either absent or requires significant custom preprocessing, making these systems inadequate for real enterprise document diversity.

2. **Weak Cross-Document Reasoning:** Existing systems retrieve chunks from individual documents and generate answers in isolation. They do not resolve entity coreferences across documents, identify contradictions between sources, or synthesize insights that span multiple documents simultaneously all of which are essential for tasks like financial auditing, legal contract analysis, and academic literature review.

3. **High Cost & Latency at Scale:** Current frameworks consume high token counts per query and introduce significant framework overhead, making them expensive and slow at enterprise scale. There is no built-in mechanism for query optimization, token budgeting, or intelligent routing that reduces cost without sacrificing answer quality.

4. **No Retrieval Transparency:** Existing consumer tools provide little to no visibility into how an answer was retrieved which chunks were used, what their similarity scores were, and which documents contributed most to the response. This lack of explainability makes it difficult to audit, validate, or trust AI-generated answers in regulated industries.

3.3 Proposed Contribution

This survey paper studies the limitations of existing research assistance and question answering systems, especially in handling information spread across multiple documents. Most traditional systems focus on keyword-based retrieval or generate answers directly without ensuring that the information is properly supported by reliable sources. This often leads to irrelevant results or incomplete understanding of a topic.

Based on the analysis of existing methods, this work highlights the importance of combining structured document retrieval with controlled answer generation. The proposed approach introduces a system that separates the process into two stages: first, selecting a set of relevant research papers, and second, generating answers only from those selected documents. This helps in reducing unnecessary information and ensures that the final output is more focused and reliable.

Another key contribution is the use of simple yet effective retrieval techniques along with basic filtering to improve relevance during the initial search stage. Instead of relying entirely on complex models, the system focuses on maintaining a balance between accuracy and efficiency. The inclusion of a limited research corpus and citation-based answers further improves transparency and makes the system more suitable for academic use.

Overall, this survey identifies existing gaps in research-oriented AI systems and presents a structured approach that improves relevance, reduces information overload, and provides answers that are clearly supported by research documents.

3.3.1 Addressing Gaps with ML [1-16]

From the survey of existing systems, it is observed that many research assistance tools face issues such as low retrieval relevance, lack of proper filtering, and generation of answers without sufficient supporting

evidence. These gaps mainly arise because traditional approaches either depend heavily on simple keyword matching or rely completely on generative models without controlling the source of information.

Machine learning techniques can be used to address these limitations in a more structured way. For example, learning-based ranking models can improve the relevance of retrieved documents by understanding patterns between user queries and document content. Instead of treating all keywords equally, ML models can learn which terms are more important based on past data and improve the quality of search results.

In addition, classification and clustering techniques can help in organizing research papers into meaningful groups, making it easier to filter out unrelated documents. This reduces noise during the retrieval phase and helps in building a more focused research corpus. Machine learning can also be applied to identify important sections within documents, allowing the system to extract only the most relevant information for answer generation.

Another important gap is the lack of reliability in generated answers. This can be addressed by combining machine learning with controlled reasoning, where the model is trained or guided to generate responses only from selected document content. This ensures that the output remains consistent with the available evidence.

Overall, machine learning provides practical solutions to improve retrieval accuracy, document filtering, and answer quality, helping in building more reliable and efficient research assistant systems.

4.Conclusion

The **Multi-Modal Cross-Document Answering System** has advances intelligent learning by integrating NLP, Computer Vision, and Deep Learning to understand and synthesize knowledge from both text and visuals across multiple documents. It overcomes the limitations of traditional text-only systems by providing contextually rich, visually grounded answers that enhance student engagement and learning outcomes. Built with scalable components like BERT, CLIP, cross-attention fusion, and FAISS retrieval, the model delivers accurate, explainable, and human-like reasoning. Though effective, it remains dependent on dataset quality and computational resources, suggesting future improvements through optimization and integration of large vision-language models for greater efficiency and generalization.

References

1. ICT-QA (2025) – Introduces a multimodal QA system handling images, charts, and text for context-rich question answering.

2. Chen, J., et al. (2025). MAIN-RAG: Multi-Agent Filtering Retrieval-Augmented Generation. Proceedings of ACL 2025. Association for Computational Linguistics.

3. Matsumoto, et al. (2025). Retrieval-Augmented Generation to Generate Knowledge Assets and Creation of Action Drivers. Applied Sciences, MDPI.

4. Cross-Modal Retrieval for Education (2024) – Uses multi-document fusion and reasoning to improve QA in educational contexts.

5. Gupta, S., et al. (2024). A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. arXiv:2410.12837.

6. Lewis, P., et al. (2024). A Systematic Review of Key RAG Systems: Progress, Gaps, and Future Directions.

7. Jiang, et al. (2024). Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers.

8. LayoutLMv3 (2022) – Pre-trains models jointly on text and images for document layout understanding in AI.

9. Cross-Document GNN (2022) – Applies graph neural networks for reasoning and fact verification across multiple documents.

10. CLIP (2021) – Trains visual models using natural language supervision to align image and text representations.

11. ViLBERT (2019) – Pre-trains models on large text-image data for versatile vision-language tasks.

12. HotpotQA (2018) – Provides a dataset for explainable multi-hop question answering across various data types.

Books & Frameworks

1. Schneider,etal.(2024). Retrieval-Augmented Generation (RAG). Business & Information Systems Engineering. Springer Nature.
2. Brown, T. B., et al. (2020). Language Models are Few-Shot Learners (GPT-3). NeurIPS 2020.
3. Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

Tools & Technologies

1. LangChain Documentation. (2024). LangChain: Building applications with LLMs.
2. LlamaIndex Documentation. (2024). LlamaIndex: Data framework for LLM applications.
3. OpenAI. (2024). GPT-4o Technical Report.
4. Anthropic. (2024). Claude 3.5 Model Card.