IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

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

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-As the rapid growth of online academic resources, identifying relevant and accurate information has become increasingly challenging for students and researchers. Traditional search systems often return large volumes of results without effectively capturing the intent of the query, leading to inefficient and time-consuming exploration. This work presents a research assistant system that improves information retrieval by combining the BM25 ranking algorithm with a controlled language model to generate answers based only on selected research documents. The system allows users to build a focused corpus of relevant papers and supports cross-document analysis to produce more comprehensive results. By integrating structured retrieval with guided answer generation, the proposed approach reduces manual effort, improves answer relevance, and enhances the overall research experience.

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

I. Introduction

A. Background and Motivation

The expansion of digital learning platforms and online repositories has made academic content more accessible than ever before. Students and researchers now have access to thousands of research papers, articles, and learning resources within seconds. However, this abundance of information introduces a significant challenge efficiently identifying useful and relevant knowledge from large datasets. Traditional search engines mainly rely on keyword matching, which often fails to understand the actual meaning behind user queries. As a result, users are required to manually explore multiple sources, making the process time-consuming and sometimes confusing. In many cases, important contextual relationships between documents are also missed.

Additionally, academic content is not limited to text alone. Diagrams, charts and graphs representations play a crucial role in explaining complex concepts. Existing systems generally ignore these components, resulting in incomplete understanding. This motivates the need for a system that can process multiple sources and provide meaningful, structured answers.

B. Need for Intelligent Automation

In the digital era, the rapid growth of educational resources such as textbooks, research papers, lecture notes, and multimedia materials has made it difficult for students and educators to efficiently locate relevant information. Traditional search and question answering systems mainly rely on keyword-based retrieval and often provide incomplete results, requiring users to manually search through multiple documents. Intelligent automation is needed to address this challenge by using technologies like Artificial Intelligence, Natural Language

Processing, and Machine Learning to automatically analyse user queries, retrieve relevant information from multiple sources, and generate meaningful responses. This approach improves information accessibility, reduces manual effort, and enables faster and more accurate knowledge discovery in educational environments.

C. Paper Objectives

The core objectives of this system that are targeted to achieve are:

1. To design a multimodal question answering framework capable of understanding both textual and visual educational content.
2. To enable cross-document retrieval and knowledge synthesis across multiple academic resources.
3. To implement transformer-based encoders for text and image representation.
4. To generate contextually relevant answers supported by both textual explanations and visual references.
5. To improve retrieval efficiency and answer accuracy by integrating relevance-based ranking mechanisms and context-driven response generation.

II. Survey of Literature

A. Multimodal Question Answering: This study focuses on improving question answering systems that can handle both textual and visual data such as images and charts. It introduces a method called Image and Chart Instruction Tuning (IC-Tuning), which helps language models better interpret combined inputs. The research also presents the ICT-QA dataset, which includes questions requiring reasoning across multiple data formats. Separate encoders are used for processing images and charts, and their outputs are aligned with textual representations to generate answers. The study highlights the importance of multimodal training for developing more effective question answering systems.

B. Cross-Modal Information Retrieval: This paper provides an overview of techniques used for retrieving information across different data types. It explains how a query in one format, such as text, can be used to retrieve data in another format, such as images. The study discusses various models, datasets, and evaluation methods used in this area. It also describes how deep learning techniques and shared embeddings help measure similarity between textual and visual data. The paper also identifies challenges such as the semantic gap between modalities and limited availability of large datasets. Overall, it emphasizes the importance of cross-modal retrieval in building advanced search and question answering systems.

C. Document Visual Question Answering; The SlideVQA research introduces a dataset designed for question answering on document images, particularly slide presentations. It includes thousands of slide decks, images, and questions, enabling complex reasoning across multiple pages. Unlike traditional approaches that focus on single images, this method requires understanding relationships between multiple slides. The study proposes a sequence-to-sequence model that performs both evidence selection and answer generation. This model supports multi-step reasoning and can handle numerical information present in slides.

D. Document Multimodal Transformer: LayoutLMv3 is a transformer-based model developed for document understanding tasks. It combines textual and visual features within a single framework by performing both masked language modelling and masked image modelling. The model introduces a word–patch alignment mechanism that helps in understanding the connection between text and image regions. Unlike earlier approaches, it avoids the use of convolutional neural networks, making it more flexible and efficient. This research demonstrates how unified multimodal training can improve performance across different document-based AI applications.

III. Proposed Work

A. Problem Definition

Despite the availability of large digital repositories of educational resources, existing question answering systems remain limited in their ability to process multimodal content and integrate knowledge from multiple documents. Traditional systems rely primarily on textual data and ignore visual elements such as diagrams, graphs, and charts that are essential for understanding academic concepts. As a result, students often receive incomplete answers that lack the contextual depth required for effective learning. The challenge is therefore to develop an intelligent system capable of interpreting both textual and visual information while simultaneously synthesizing knowledge from multiple academic documents.

B. Implementation Details

The system is implemented using a modern web-based architecture that integrates artificial intelligence and information retrieval techniques to assist users in academic research.

Frontend/UI: Developed using Next.js (React) to provide a responsive interface where users can enter research topics, review retrieved papers, and interact with the system.

Backend & Database: The backend manages query processing, document retrieval, and corpus construction. Research data and metadata are stored in a PostgreSQL database.

AI Engine: A large language model such as **Gemini** developed by **Google** is used to generate responses based on retrieved research documents.

Retrieval Mechanism: The system uses the **BM25** algorithm to retrieve relevant research papers during the exploration phase.

This modular architecture ensures efficient retrieval, reliable research assistance, and scalability for future system improvements.

C. Mathematical Model

1. Query Feature Extraction

For textual input, the system extracts important keywords from the user query to capture its semantic meaning. Instead of deep neural encoders, a lightweight preprocessing approach is used, which includes tokenization, stop-word removal, and normalization. This helps in identifying the most relevant terms required for document retrieval.

If $Q = [q_1, q_2, \dots, q_n]$represents the input query tokens, the extracted keyword set can be defined as:

$$K = fextract(Q) = \{k1, k2, \dots, km\}$$

2. Document Retrieval Model

For document retrieval, the system uses the BM25 to rank research papers based on their relevance to the query. Each document is scored using term frequency and inverse document frequency, ensuring that more relevant documents receive higher scores.

The BM25 scoring function is defined as:

$$Score(d,q) = \sum_{tq} IDF(t) . \frac{f(t,d).(k1+1)}{f(t,d) + k1 \cdot (1 - b + b \cdot avgdl \mid d \mid)}$$

where:
$f(t,d)$=frequency of term $t$in document d
$\mid d \mid$=length of document
$avgdl$=average document length
k_1, b= tuning parameters

3. Corpus Formation and Processing

The top-ranked documents are selected to form a research corpus. These documents are further processed by dividing them into smaller text segments to improve retrieval efficiency during query execution.

$$C = \{d1, d2, \dots, dN\}$$

$$Ch = fchunk(C) = \{c1, c2, \dots, cm\}$$

where:
• C= selected document corpus
• C_h= set of document chunks

4 Context-Based Answer Generation

The system generates answers using a language model such as Gemini, based only on the selected document chunks. Relevant chunks are first selected and combined into a context, which is then passed to the model.

$$X = \cup Cr$$
$$y\hat{} = fLLM(Q, X)$$

where:
• C_r= relevant chunks
• X= combined context
• $\hat{y}$= generated answer

5.Evaluation Metrics

The system performance is evaluated using standard metrics such as Precision, Recall, and F1 Score.

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1 = 2.(Precision \cdot Recall)/(Precision + Recall)$$

where:
• TP = True Positives
• FP = False Positives
• FN = False Negatives

6.Mean Reciprocal Rank (MRR)

The effectiveness of document retrieval is evaluated using Mean Reciprocal Rank (MRR), which measures how quickly the correct document appears in the ranked results.

$$MRR = \frac{1}{N}\sum_{I=1}^{N} 1/rank_i$$

D. Flow Diagram & Process Flow

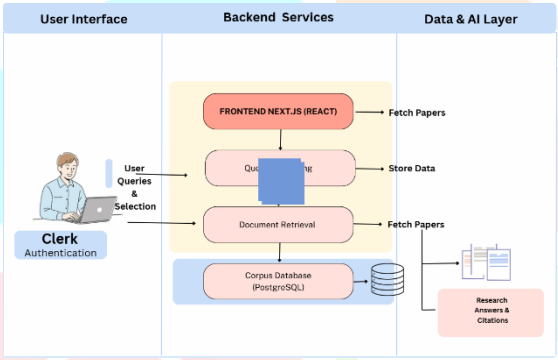


Fig1-System Design

Fig. 1 This figure illustrates the system architecture of the proposed research assistant, organized into three main layers: User Interface, Backend Services, and Data & AI Layer. In the User Interface layer, users interact with the system by submitting queries and selecting research papers, while authentication is securely managed. These inputs are then passed to the Backend Services layer, where the frontend built which processes user requests and forwards them to the query processing module. The system then performs document retrieval to fetch relevant research papers, which are stored in a structured Corpus. In the Data & AI layer, the stored research corpus is utilized to generate answers, where the system retrieves relevant documents and produces research-based responses along with proper citations. This layered architecture ensures a clear separation of responsibilities, enabling efficient query handling, secure data management, and accurate AI-driven answer generation.

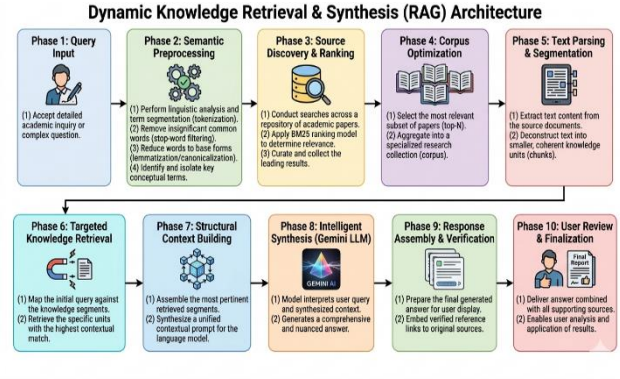


Fig 2-Process Flow

Fig. 2 The figure represents the complete workflow of the proposed Multi-Modal Cross Document Question Answering System, illustrating how a user query is transformed into a final research-based answer through ten sequential phases. The process begins with the user providing an academic query, and key concept extraction to better understand the intent of the query. The system then performs source discovery using the BM25 algorithm to retrieve and rank relevant academic papers, after which the most relevant documents are selected and organized into a focused research corpus. The selected documents are further processed by extracting and segmenting their content into smaller meaningful chunks for efficient retrieval. The system then performs targeted knowledge retrieval by matching the query with the most relevant segments, followed by structural context building where the retrieved information is organized into a meaningful prompt. This structured context is passed to a large language model which synthesizes a comprehensive and context-aware answer. Finally, the response is assembled with proper references and verified against the source documents before being presented to the user, allowing for accurate, reliable, and research-supported answers while enabling further analysis and validation.

E. Algorithms Used

The proposed system utilizes a combination of information retrieval algorithms and controlled generative AI workflows.

BM25 Retrieval Algorithm

The BM25 algorithm is used during the exploration phase to rank research papers according to query relevance.

The steps involved in the BM25 retrieval process include:

1. **Query Processing** – Extract keywords from the research topic.
2. **Document Scoring** – Calculate relevance scores using BM25 ranking functions.
3. **Ranking** – Rank documents based on computed relevance scores.
4. **Filtering** – Apply topical specificity filters to remove loosely related documents.
5. **Shortlisting** – Select the top ranked documents for user approval.

This algorithm ensures high precision retrieval while maintaining computational efficiency.

3.5.2. Corpus Retrieval Algorithm

During Research Mode, the system performs contextual retrieval within the constructed corpus.

Steps include:

1. Accept the user's research query.
2. Retrieve relevant document segments from the corpus.
3. Rank retrieved segments according to semantic similarity with the query.
4. Select the most relevant evidence passages.
5. Provide these passages as context for the AI reasoning module.

This step ensures that responses remain grounded in academic sources.

3. Controlled Generative Reasoning

The generative AI component follows a constrained reasoning approach.

The process includes:

1. Input: User research query and retrieved document context
2. Prompt Structuring: The system constructs a structured prompt containing retrieved evidence
3. Constraint Enforcement: The AI model is instructed to answer only from the provided sources
4. Response Generation: The AI synthesizes information across documents
5. Output: The system returns a research-supported answer with citations

This method enables cross-document synthesis while maintaining academic reliability.

IV. System Modules & Glimpses of Output

The proposed research assistant system is divided into different modules, where each module performs a specific task in the research workflow. These modules help the user explore research topics, retrieve relevant papers, process documents, and generate answers based on selected research content. This modular design makes the system easier to manage and allows future improvements to be added easily.

A. Topic Exploration Module

This module handles the initial interaction with the user. When a user enters a research topic or query, the system extracts important keywords and uses them to search for relevant research papers.

The system retrieves papers from academic sources such as arXiv. To rank the results, it uses the BM25 algorithm, which measures how relevant each document is to the search query.

A simple filtering step is also applied to ensure that the retrieved papers closely match the research topic.

B. Paper Selection Module

After retrieving the research papers, the system displays the most relevant papers to the user along with basic details such as title and abstract.

The user can review these papers and select the ones that are most useful for their research. To maintain efficiency and avoid information overload, the system limits the number of selected papers to around ten.

C. Document Processing Module

Once the papers are selected, the system processes the documents by extracting text from the PDF files. The extracted text is then divided into smaller sections so that the system can easily search through the content later.

These processed sections are stored in the system database and used during the research phase.

D. Research Query Module

This module allows users to ask detailed questions related to the selected research

papers. The system searches the processed documents to find the most relevant sections related to the user's query.

These sections are then analyzed by a large language model such as Gemini, developed by Google, to generate a response. The model is restricted to use only the retrieved research content, ensuring that the answers remain research-based.

E. Citation and Evidence Module

The final module ensures that all generated answers are supported by references from the selected research papers.

The system identifies which documents were used to generate the response and provides citations so that users can verify the information or read the original papers.

This helps maintain reliability and makes the system more useful for academic research.

V. Results and Outcomes

The implementation of the proposed research assistant system showed positive results in terms of research efficiency and information retrieval. The system was tested using different research queries to observe how effectively it retrieves relevant papers and generates answers based on selected documents.

A. Retrieval Relevance

The system improves retrieval relevance by using the BM25 algorithm to rank research papers according to their similarity with the user's query. This helps filter out unrelated papers and ensures that the retrieved results are closely related to the research topic.

B. Faster Literature Exploration

The system reduces the time required to explore research papers. Instead of manually searching through multiple sources, the system retrieves relevant papers from repositories such as arXiv and presents them to the user for selection.

C. Context-Based Answer

The system can generate answers based on the selected research papers. A large language model such as Gemini, developed by Google, analyzes relevant document sections and generates responses based on the retrieved research content.

This approach helps ensure that the answers are supported by academic sources and remain relevant to the research topic.

VI. Graphs

A. Model Accuracy Comparison

The system outputs demonstrate the practical functioning of the proposed research assistant. These outputs show how the system retrieves research papers, processes documents, generates answers, and provides supporting evidence in a structured manner. They also highlight the system's ability to deliver accurate and contextually relevant responses based on the selected research corpus.

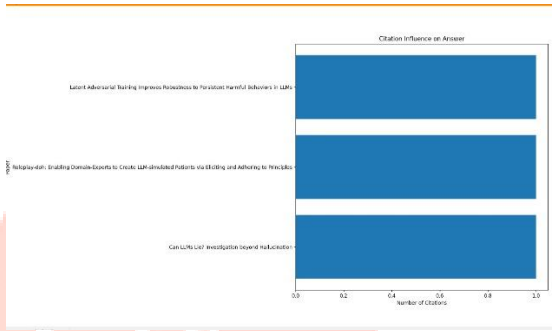


Fig.3 - Citation influence on answer

The citation influence graph illustrates how different research papers contribute to the final generated answer. In this graph, the horizontal axis represents the number of citations, while the vertical axis indicates the research papers considered by the system. It can be observed that a few papers have noticeably higher citation counts compared to others, suggesting that they play a more significant role in forming the final response. For example, some papers contribute extensively because they contain more relevant and frequently referenced information related to the query. On the other hand, some papers appear with lower citation counts, indicating that they are used only for supporting or supplementary details. This variation shows that the system does not treat all documents equally, but instead prioritizes sources based on their relevance. Overall, the graph helps in understanding how the system selectively uses information from multiple papers during answer generation.

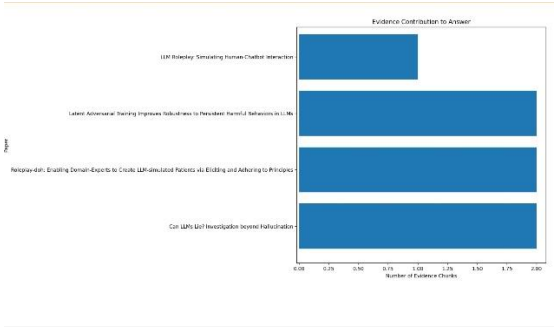
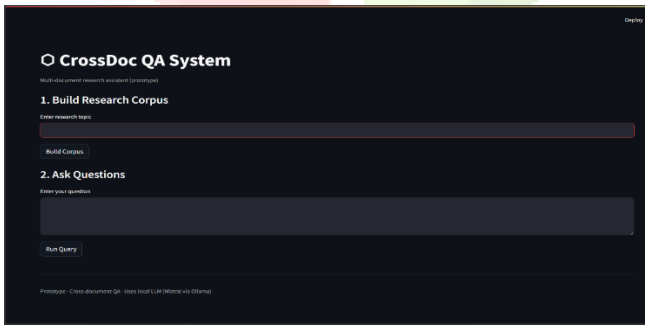


Fig.4-Evidence Contributions

The evidence contribution graph represents how different document segments are utilized during the answer generation process. Each segment contributes a certain amount of information depending on its relevance to the user's query. The graph shows that some segments have a higher contribution level, meaning they provide key information required for forming the answer.

In contrast, other segments contribute less and are mainly used to support or refine specific parts of the response. This distribution highlights the system's ability to identify and prioritize the most relevant content from the selected research corpus. It also reflects how the system combines multiple pieces of evidence to produce a well-structured and context-aware answer.

Additionally, this behaviour demonstrates that the system follows a selective approach rather than treating all information equally. Such prioritization helps improve the accuracy and overall quality of the generated responses.



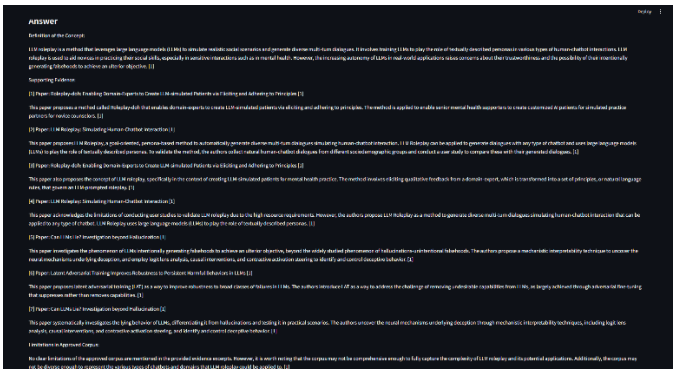


Fig.5 - Output Screenshots

B. User Interface Output

The user interface of the system is designed to provide a simple and structured workflow for performing research tasks. It is divided into two main sections: building the research corpus and asking questions.

In the first section, the user enters a research topic and clicks on the "Build Corpus" option. The system then retrieves relevant research papers and displays a confirmation message indicating the number of papers selected. This step ensures that the system works with a focused set of documents rather than a large and unstructured dataset.

The second section allows the user to enter a question related to the selected research topic. Once the query is submitted, the system processes the request and generates a response based on the constructed research corpus.

C. Generated Answer Output

The generated answer output represents the final response provided by the system. The answer is structured in a clear and organized format, starting with a definition or explanation of the concept, followed by supporting evidence.

The system includes references to multiple research papers used during answer generation. Each reference highlights how the information has been gathered from different sources, showing that the answer is based on actual research content.

This output demonstrates that the system does not generate answers randomly but instead uses the selected research corpus to produce accurate and relevant responses. The inclusion of supporting evidence improves the reliability of the answer and makes it more useful for academic purposes.

VII. Conclusion and Future Scope

Conclusion: The proposed cross-document research assistant system demonstrates how structured information retrieval can be effectively combined with controlled AI-based answer generation to improve the research process. By using the BM25 algorithm to identify relevant papers and limiting the language model to selected content, the system produces responses that remain accurate and grounded in reliable sources. The overall workflow, which includes building a focused research corpus, processing documents, and generating context-based answers, helps reduce the time and effort required for manual literature review. Users can access relevant information more efficiently without having to go through multiple documents individually.

The inclusion of supporting evidence and references enhances the credibility of the generated responses. This makes the system more suitable for academic use, where reliability and traceability of information are important. Overall, the system meets its objective of providing a simple and effective research assistance platform.

Future Scope: The proposed Cross-Document Research Assistant system can be further enhanced by integrating semantic search techniques that understand the context of queries instead of relying only on keywords. The answer generation process can be improved by fine-tuning language models for specific domains and adding fact-checking mechanisms to increase reliability. The system can also be extended to support multimodal data such as images, charts and tables for better understanding of complex topics. Additionally, real-time integration with research databases can ensure access to updated information.

VIII. 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) – Pretrains 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) – Pretrains 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.