IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.153126

GraphSense-RAG: An Intelligent Multi-Modal Document Question Answering System

Dr. M. Purnachandra Rao[1], D. Mastan Bee[2], B. Navya[3], B. Sravani[4], A. Bhargavi[5], B. Uma[6]

Associate Professor, Department of Information Technology, KKR & KSR Institute of Technology and Sciences Guntur, India[1]

Student, Department of Information Technology, KKR & KSR Institute of Technology and Sciences Guntur, India[2-6]

Abstract: The rapid growth in the digital documents made a critical challenge in retrieving accurate information from the large volumes of data. This question-answering system has become a challenging task for the companies and individuals who deals with the huge volume of information. To overcome these challenges, the proposed mainly focuses on Document Question Answering system based on Retrieval-Augmented Generation approach which is integrated with the embedding search. The main aim of this project is to provide accurate and reliable information for the user query from the document. The proposed system supports multi modal data processing and also uses hybrid retrieval technique which includes semantic search, keyword matching and metadata filtering. These techniques are used to enhance the retrieval performance and accuracy. The proposed system also included with a dedicated hallucination detection mechanism which validates the generated result by grounding them in retrieved evidence. In addition to this the system also supports automatic document summarization and indexing. It also allows the user for efficient ingestion of new documents. The proposed system provides scalable, reliable solution which can be effectively used for academic research and also for the enterprise companies.

Index Terms: Retrieval-Augmented Generation, Vector Embedding Search, Multi-Modal Question Answering, Hybrid In- formation Retrieval, Knowledge Graph Integration

I. INTRODUCTION

The huge development in the digital information over a few years significantly played a crucial role in transforming how information is created, organized and processed. The digital information may be huge information for some of the institutions and organizations. The digital information can be stored in different formats of data like PPT's, PDF's, word documents and images. By the increase in the digital information there is an increase in availability but at the same time it also introduced a new challenging for the users who are dealing with large documents. The challenge introduced by digital document is retrieval of accurate information for the user query from the large document. This reduced the efficiency and reliability of the system. The users have to spend time on the result generated by traditional systems to get the accurate result. Sometimes traditional systems generate the result which is irrelevant and incorrect according to the user query.

Modern developments in Natural Language Processing (NLP) and Large Language Models (LLMs) have enabled machines to understand and generate human-like text, offering practical solutions for question answering systems. These modern models produce a context-aware results but they commonly depend on their trained data. Therefore, these models may generate the information that is not accurate and reliable; this process is commonly known as hallucination. This drawback raises the challenges in reliability, particularly in academic research, enterprise decision-making where accuracy plays key role.

To overcome these challenges an efficient approach Retrieval-Augmented Generation (RAG) has been developed by integrating document retrieval with language generation. Instead of generating hallucination results, RAG system converts the document into segments then it retrieves the relevant document segment from the segments after that the RAG verifies the information accuracy and make this information as the final result for the user query. The results generated by this RAG system are both context-driven and evidence-driven. Based on this groundwork, the proposed system introduces an improved RAG framework which integrates hybrid retrieval strategies, multi-modal document processing, automatic summarization, and a hallucination detection mechanism.

The system enhances its retrieval precision and relevance by integrating with semantic search, keyword matching and metadata filtering. This system also contains the hallucination detection technique which validates the information

DOI: 10.17148/IJARCCE.2026.153126

generated by the RAG. Therefore, this hallucination detection increases the user's confidence. Therefore, the proposed system provides scalable, reliable and accurate information which are plays a key role in academic research and real-world enterprise knowledge management.

II. LITERATURE REVIEW

Mukhopadhyay et al. proposed a library-focused conversational search system (LibGPT) using open-source RAG architecture [1]. The main aim of this project is to present the key limitations of Large Language Models (LLMs) like hallucination. This system integrates the document retrieval with generative models to produce the hallucination free responses. It is an open-source tool but the main limitations of this system are it is only limited to text-based document, it doesn't support hybrid search. P. Lewis et al., proposed a system which is named as Retrieval-Augmented Generation for knowledge-Intensive NLP Tasks [4]. In this system the author used Retrieval-Augmented Generation (RAG) which combines information retrieval with text generation to improve performance on knowledge-intensive NLP tasks. The main drawback of this system is it has high computation cost and only used to process the one format of documents. J. Devlin et al., proposed a question answering system which is named as BERT: pre-training of Deep Bidirectional Transformers for Language Understanding [5]. The main purpose of this project is to pre-trainer the deep bidirectional language models with the help of Transformer encoders. The main drawback of this system is not suitable for text generation because it is an encoder. The input length must be only 512 tokens. To further enhance retrieval accuracy, the QuIM-RAG framework intro- duces an innovative Inverted Question Matching technique [2]. This system generates potential questions from documents to match the user query. After that the user's generated questions are compared with user's actual query. The main limitations of this system are it does not support multi modal data processing and it does not provide conversational memory for storing the context. Manjusha et al., proposed an innovative Multi- Document Question Answering System using RAG [3]. This system allows the user to upload the multiple documents for data processing. This system provides the conversational interface which is used for interactively asking questions to the system to the user. It uses Vector embeddings, ChromaDB, LangChain and Llama-based models for accurate responses. The main limitation of this system is it supports multiple documents of text-based only. It does not support multi-modal data.

\

From the above analysed research articles, it is proof that the existing RAG based systems improves accuracy and reduce hallucinations compared with the traditional methods but they are only limited to only single format of data which is text- based document, absence of hallucination detection and con- versational memory. To overcome all these challenges there is need of enhanced RAG-based system that must support multi- modal data, hybrid search, conversational memory, accurate response and scalable deployment.

III. METHODOLOGY

In this section we will discuss about the overall process and working of proposed question answering system. The components of this proposed system have its own importance and working principle. The system starts its processing by Document ingestion where data is collected and organised and processed. The data processing can be done by the integration of hybrid search and RAG. The output generated is validated by using hallucination detection. The final result is treated as the output given by proposed system.

A. Document Ingestion and Preprocessing

Document Ingestion is the starting component in the pro- posed system. In this step the system accepts the multiple formats of data like pdfs, word files, text-files, images and tabular formats. For every format of input we are using different techniques to extract data. To extract information from images and scanned documents we are using Optical Character Recognition (OCR).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.153126

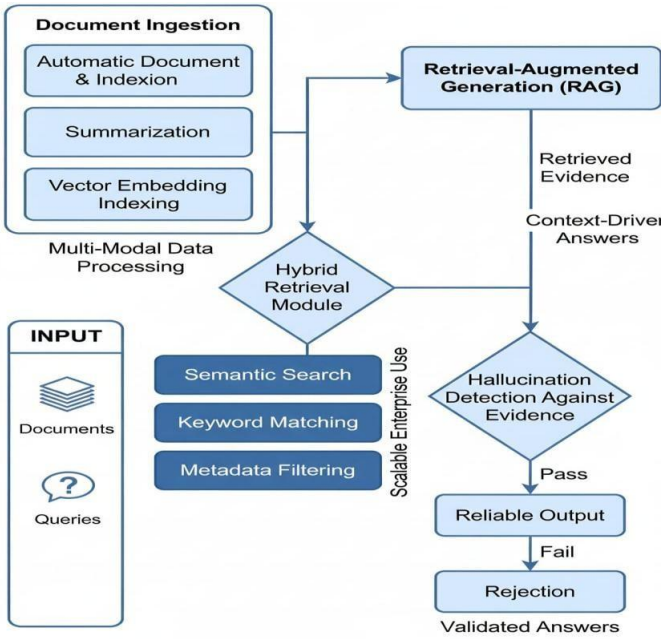

Fig. 1. System architecture of the proposed GraphSense-RAG framework

The input for this system is document and user query. After the text extraction from the documents the information is divided into smaller and similar segments. These segments are meaningful chunks. This segmentation process helps to improve accuracy and reliability. Not only these the Document Ingestion and preprocessing there is an feature of storing important data of the document like document title, author name, file type and main, sub headings.

B. Vector Embedding Generation

After the completion of preprocessing of data, each chunk or segment is converted into numerical representation. The processing of converting documents chunks into numerical chunks is called vector embeddings. The main purpose of this vector embedding is that the computers and system cannot understand human understandable language they can only understand binary language. The process for the solution generation is done by using this binary language. Another purpose of this vector embedding is to capture the meaning of text and allowing the system to understand the similar meaning between document chunk and user query. The vector databases like FAISS or ChromaDB are used to store these embeddings which improves fast and efficient similarity search during the process.

C. Hybrid Retrieval Strategy

Hybrid retrieval strategy plays an important role in the improvement of accuracy and reliability in the generated result. After the completion of data preprocessing and vector embeddings the data is processed by using the following searching techniques:

- **Semantic Search**: In this search the user's query is compared with the embeddings that are stored in the vector databases. In this semantic similarity takes place. The document segment with the highest similarity is retrieved.
- **Keyword Matching**: In this technique the user's query is compared with the document segment with a keyword in the user's query. The result is retrieved only when is a complete or partial matching of keyword. This is the technique that is using by the traditional systems. In this the technical terms and keywords are captured and stored.
- **Metadata Filtering**: In this technique the system uses the metadata of the document like document title, author, date, file type and so on. This metadata is used reduce the searching results which means sometimes searching technique may produce irrelevant information. To reduce incorrect results the metadata filtering is required.

The retrieved result by combining the sematic search, key- word matching and meta data filtering, the result is validated by ranking the result based on relevance, similarity search and keyword matching strength. The segment which got the highest ranking is treated as finalised result. The generated result is accurate according to the user query.

D. Answer Generation Using RAG

In this stage which we are using Retrieval Augmented Generation (RAG) to provide reliable and accurate results. Here

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.153126

the finalised and selected segments are given as an input to large language models as a context. After that the model generate the results which are completely depends on retrieved content. The responses generated are grounded to the document. For large language we are using RAG. This approach reduces the chances of generating of incorrect and irrelevant answers.

E. Hallucination Detection Mechanism

Hallucination detection mechanism is important and play a key role in generating reliable, accurate and trustworthy results. The generated results by the RAG are given as an input to this hallucination detection. This mechanism reduces the im- proper results. In this step each sentence in the generated result is verified by grounding them to the document. If the result seems to be irrelevant this mechanism will automatically make that irrelevant deletes and sometimes make it correct. This validation process reduces hallucination answers. Therefore, the result generated by the hallucination detection mechanism is accurate and reliable.

F. Automatic Summarization and Indexing

Whenever the new document is added to the proposed system then this system automatically summarizes the newly added document. This summarization process helps to reduce memory space and highlights the key points of the documents. This can make the result reliable and make easy to understand. In addition to this the system also contains another important feature, Indexing. When the result generated the system will automatically index which means gives the numerical values to the generated document. This reduces the ambiguity towards the generated result.

The following flow chart represents the working principles of proposed system also explains the formulas that are using in every step.

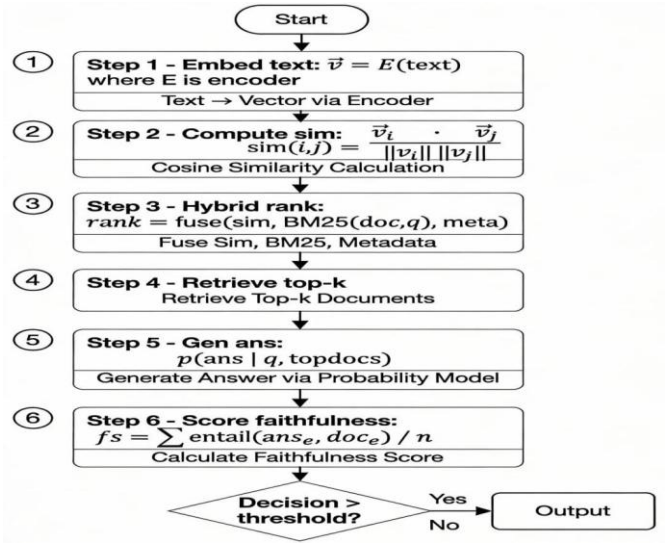

Fig. 2. Flowchart of the proposed GraphSense-RAG system workflow

IV. RESULT

This proposed system is using a multimodal documents like text documents, pdfs, word files, images. These tested documents include research papers, technical manuals and enterprise records. This proposed system is tested in different datasets. These datasets are helped to analyse the system's performance across multimodal documents structure, length, size. This proposed system is compared with the traditional key-word based search to analyse the accuracy and reliability of the generated result by the proposed system.

The primary purpose of the development of Question answering system is to focus on result accuracy, context relevance, response reliability and efficient retrieval. By observing the outcomes of testcases we identified that the hybrid retrieval-based RAG model always produces accurate, reliable and efficient responses to the user queries compared with the traditional systems. The proposed system integrates the semantic search with the keyword matching and metadata filtering which makes the system to identify relevant document segments.

The important result generated for the proposed system is the enhancement of reliable responses. The is due to the hallucination detection mechanism. The result generated by the proposed system is validated with the retrieved

DOI: 10.17148/IJARCCE.2026.153126

document evidence. This helps the system to reduce the occurrences of invalid or incorrect results. This validation process mainly used for large, complex and invalid user queries. Therefore, the proposed system generates reliable and accurate results.

The proposed systems use a vector database which helps the system to generate fast similarity search even when the document size increases. In addition to this, the system also handles multimodal document and multi-sentence queries ef- fectively and also generate accurate results for the multimodal document.

Therefore, the results of the experimental test stated that the proposed system offers an important enhancement over the traditional systems. By integrating hybrid retrieval, RAG and hallucination validations made the results to be more accurate and reliable and also made this system is well-suited for real- world enterprise applications and academics.

V. DISCUSSION

The results shows that proposed Document Question Answering System is an efficient and effective system because it brings together several retrieval and validations methods in one system. The hybrid retrieval method balances both semantic and exact keyword matching. Semantic searching technique is used to understand the user queries, keyword matching technique is used to match the keyword of the user's query with the document retrieved segment. Metadata filtering technique is used to improve precision and also by summarizing the results.

The most important technique and key role technique is hallucination detection mechanism. This technique plays an important role in improving system reliability. The proposed system validates the generated results with the retrieved doc- ument segment. This reduces the risk of incorrectness or misleading information. This is the most important feature which is important for academic research and real-world enterprise applications.

Additionally, the proposed system has another feature of automatic document summarization and indexing. In this feature the system will automatically summarize the retrieved information which can reduce the storage memory. Automatic indexing is the most important feature which means the system will automatically give the indexing to the generated result. This feature reduces the ambiguity for the large and complex documents. These features enhance the scalability and reliability of the generated result.

VI. CONCLUSION

Therefore, the proposed system provides a reliable, accurate and scalable Document Question Answering System. This sys- tem is integrated with the vector embeddings. The main use of vector embedding is use to convert document segments to the numerical in such a way that texts with the similar and same meanings have similar vector representations. The traditional systems which depend only on keyword-based searching give unreliable information whereas the proposed system ensures that answers are generated using verified document content which increases the reliability and accuracy.

The performance of the system can be improved by integrating hybrid retrieval method with semantic search, key- word matching and metadata filtering. Hallucination detection mechanism plays an important role in the enhancement of trustworthiness of the system. The most important feature of automatic summarization and indexing helps to reduce the manual effort of user.

Therefore, the proposed system proved it is an experimental and efficient solution for real-world enterprise application, academic research and also for the individuals and companies who deals with the large volumes of information for generating accurate and reliable results for their queries.

REFERENCES

[1]. P. Mukhopadhyay, "Designing Conversational Search for Libraries: Retrieval-Augmented Generation through OpenSource Large Language Models," *DESIDOC Journal of Library & Information Technology*, vol. 45, no. 2, pp. 123–134, Mar. 2025.

[2]. "QuIM-RAG: Advancing Retrieval-Augmented Generation with Inverted Question Matching for Enhanced QA Performance," *International Journal of Artificial Intelligence and Data Science*, 2024.

[3]. M. P. K. Manjusha, M. Deeksha, G. Deepika, M. Kaveri, H. Anusha, and L. Shree, "Multiple Document Based Q&A Using Retrieval-Augmented Generation," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 27,

DOI: 10.17148/IJARCCE.2026.153126

no. 2, ser. 1, pp. 6–10, Mar.–Apr. 2025, doi: 10.9790/0661-2702010610.

[4]. P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[5]. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Pro- ceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[6]. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *International Conference on Learning Representations (ICLR)*, 2013.

[7]. A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[8]. Y. Karpukhin, V. Oguz, S. Min, et al., "Dense Passage Retrieval for Open-Domain Question Answering," *Proceedings of EMNLP*, pp. 6769– 6781, 2020.

[9]. M. Peters, M. Neumann, M. Iyyer, et al., "Deep Contextualized Word Representations," *Proceedings of NAACL-HLT*, pp. 2227–2237, 2018.

[10]. S. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," *Proceedings of EACL*,pp. 874–880, 2021.

[11]. O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," *Proceedings of SIGIR*, pp. 39–48, 2020.

[12]. T. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877– 1901, 2020.

[13]. S. Min, D. Sanh, X. Du, et al., "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?," *Proceedings of EMNLP*, pp. 4918–4932, 2022.

[14]. H. Chen, A. Guu, L. Liu, et al., "Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index," *Proceedings of ACL*, pp. 2177–2189, 2021.

[15]. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of EMNLP*, pp. 3982– 3992, 2019.

