Enhanced Retrieval-Augmented Generation Framework for Intelligent Multi-Document Question Answering

Navya Kumar, Aviral Pandey, Dr. Lakshmi Dhevi B

Networking and CommunicationsSRM Institute of Science and TechnologyKattankulathur, Chennai–603203, Tamil Nadu, India

DOI: https://doi.org/10.51244/IJRSI.2026.1305000033

Received: 24 April 2026; Accepted: 29 April 2026; Published: 23 May 2026

ABSTRACT

Retrieval-Augmented Generation (RAG) improves Large Language Models (LLMs) by using external documents to support their answers. However, baseline RAG architectures are limited by single-modality retrieval, fixed-size chunking, and lack of hallucination monitoring. This paper introduces an advanced hybrid RAG framework for multi-document question answering, enhancing retrieval quality, contextual coherence, and response fidelity.The proposed system combines FAISS's dense semantic retrieval with BAAI/bge-large-en-v1.5 embeddings and BM25Okapi's sparse lexical retrieval. Reciprocal Rank Fusion (RRF) combines results from both modalities to improve recall without changing any parameters. A semantic chunking strategy is introduced to keep the meaning of documents. This strategy uses sentence-level embeddings and percentile-based breakpoint detection to adaptively split documents. A cross-encoder reranker (ms-marco-MiniLM-L-12-v2) is used to improve the relevance scoring of the retrieved candidates.To mitigate hallucination without additional computational overhead, a reference-free faithfulness score is calculated by comparing the cosine similarity of generated responses to retrieved context embeddings. A multiprovider LLM abstraction layer makes sure that different cloud models are all based on the same things. The system is evaluated using Recall@K, Mean Reciprocal Rank (MRR), Precision@K, faithfulness score, and end-to-end latency. This shows that it is better at retrieving information and generating grounded information than dense-only baselines.

Index Terms—Retrieval-Augmented Generation, Hybrid Retrieval, FAISS, BM25, Semantic Chunking, Cross-Encoder Reranking, Hallucination Mitigation, Multi-Document Question Answering.

INTRODUCTION

Large Language Models (LLMs) have demonstrated strong abilities in natural language understanding and generation in a wide variety of fields. Nevertheless, standalone LLMs rely on on fixed parametric knowledge and tend to produce hallucinated or factually invalid answers to domain specific or document intensive queries [15].

Retrieval-Augmented Generation (RAG) has emerged as an effective solution, combining external document retrieval with generative models, thus enhancing conceptual relevance and factual grounding [1], [2]. RAG systems minimize hallucination by adding evidence that has been retrieved into the generation process and improves the reliability of the responses.

Although these enhancements have been made, there are a number of constraints on baseline RAG architectures. First, most systems only use dense vector retrieval, which does not necessarily result in the determination of exact lexical matches, which are important in technical or domain-specific documents [3]. Second, the strategy of fixed-size chunking tends to cross semantic boundaries resulting in context fragmentation and low-quality generation. Third, lightweight hallucination monitoring mechanisms are absent in the majority of the implemented cases, and it is not easy to evaluate response groundedness in real time [15].

To overcome these shortcomings, this paper proposes a more advanced hybrid approach of RAG in answering multidocument questions. The system combines FAISS-based embeddings of dense semantic retrieval with

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI Volume XIII Issue V May 2026

sparse lexical retrieval of BM25, with Reciprocal Rank Fusion (RRF), which combines the two. The semantic chunking technique maintains the coherence of the document by embedding-based segmentation, whereas a cross-encoder reranker enhances the accuracy of retrieval. Also, to offer a lightweight signal of hallucination, a reference-free scoring mechanism of faithfulness is presented that does not require extra evaluation overhead.

The proposed architecture improves retrieval robustness, contextual integrity, and response reliability at realistic latency. The other part of this paper introduces associated literature, system design, evaluation strategy, and outcomes of the experiment.

RELATED WORK / LITERATURE REVIEW

This sub-section surveys existing literature in the Large

Language Models (LLMs), Retrieval-Augmented Generation (RAG) models and systems of multi-document question answering.

Large Language Models in Knowledge-Intensive Tasks

Transformer-based Large Language Models (LLM) have already achieved strong performance on such tasks as natural language generation, reasoning, summarization, and conversation. Instruction and reinforcement based models have improved the situation of contextual coherence and quality of conversation.

Despite these improvements, parametric knowledge, which is acquired during pretraining, becomes a basis of the work of LLMs. Thus, they experience problems with domainspecific, developing or document-bound information. They are able to give factually incoherent or fictional responses to the knowledge-based queries which are not learned in their training and it is commonly referred to as a hallucination. Scaling model parameters enhances fluency and generalization although it does not actually solve the problem of factual grounding. These limitations have led to the integration of external knowledge retrieval processes in generative pipelines.

Retrieval-Augmented Generation (RAG*)*

An enhancement of the LLM is the Retrieval-Augmented Generation (RAG) in which the text is generated after the external documents have been retrieved. A three-stage architecture is the simplest RAG architecture that comprises of: (i) document indexing, (ii) query-time retrieval and (iii) contextgrounded generation. The parts retrieved are inserted into the prompt and the model is capable of generating answers in detail to explicit evidence.

Dense retrieval models like dual-encoder networks and the semantic search based on FAISS have demonstrated effective performance in the conceptual similarity modeling. Sparse retrieval models such as BM25 remain effective in the case of lexical matching particularly in case of technical or terminology-intensive corpora. Dense and sparse signals combination during hybrid retrieval strategies has been found to increase the recall and strength.

The adaptive retrieval strategies are new extensions, which are memory-augmented architecture and cross-encoder reranking modules to enhance relevance ranking. However, most of the existing systems are founded on fixed retrieval depth, homogeneous plans of the methods of the chunking, and fixed settings of embedding. Such limitations can contribute to the unnecessary context, increase the process cost, and decrease the quality of generation in the multi-document environment. In addition, the methods of hallucinations detection are largely extrinsic or ingenious, thus, cannot be applied in real time.

Multi-Document Question Answering

Multi-document question answering (MDQA) systems attempt to combine information in two or more documents in order to produce cohesive and evidence-based responses. Traditional pipelines have also integrated the extraction summation and information retrieval. More recent neural techniques involve neural

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI Volume XIII Issue V May 2026

retrieval itself in generative models with the benefit that hybrid synthesis can be performed on nonhomogeneous sources.

Despite such advances, MDQA systems are linked to retrieval accuracy challenges, cross-document argumentation and retrieval lag. Dense-only retrieval was unable to depict the lexical exact match, and sparse-only retrieval was unable to depict deep semantic representation. Besides, there is a possibility of the fixed-size chunking of the semantic coherence that impacts downstream generation in a negative manner. Few systems incorporate lightweight mechanisms to assess the fidelity of the answers produced against the context that the answers are retrieved.

Research Gap

According to the literature review, some of the limitations remain evident:

1) High level of exclusive application of particular modality retrieval (dense-only or sparse-only).

2) Semantically-blind fixed/heuristic chunking.

3) No query-sensitive adaptation, no level of depth on the access of static contents.

4) Minor inclusion of good hallucination detection mechanisms.

5) Insufficient evaluating groundedness and re-trieval measures.

This paper addresses these gaps by offering an Adaptive Hybrid Retrieval-Augmented Generation (AH-RAG) model that entails hybrid dense-sparse retrieval, Reciprocal Rank Fusion, embedding-guided semantic chunking, cross-encoder reranking, and sparse faithfulness scoring system. The proposed system is designed to enhance retrieval strength, contextual integrity and reliability to question answering opportunities of multiple-documents without breaking the real-life latency conditions.

PROPOSED METHODOLOGY

This section outlines a proposed Hybrid Dense-Sparse Retrieval-Augmented Generation system of multi-document question answering. The system is designed in the form of a modular five stage pipeline that includes document ingestion, semantic chunking, hybrid retrieval, reranking, grounded generation and faithfulness evaluation.

As indicated by the architecture in Fig. 1, the system consists of document ingestion, hybrid retrieval, reranking, generation and evaluation blocks.

Document Ingestion and Semantic Chunking

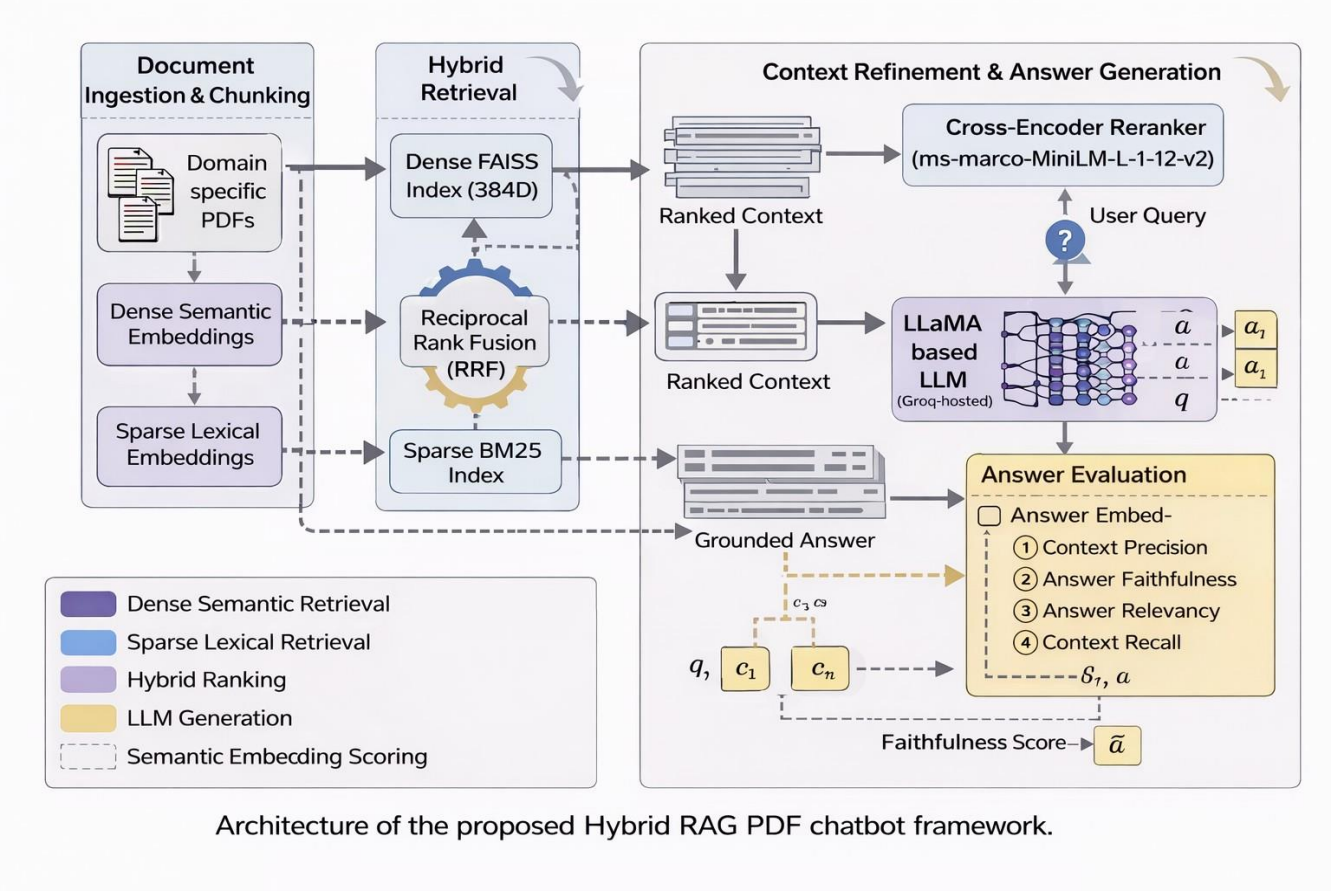
Given a document corpus

$$D = \{d_1, d_2, \ldots, d_n\} \qquad (1)$$

Given a document corpus, all documents are run to text units page by page, and then broken up into semantically coherent units.

We apply embedding directed semantic chunking strategy, unlike the fixed-size chunking strategies. All sentences s_i of

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI Volume XIII Issue V May 2026



Architecture of the proposed Hybrid RAG PDF chatbot framework.

Fig. 1. Architecture of the multi-document question answering proposed Hybrid RAG framework is based on architecture.

an entry are coded through the assistance of a bi-encoder encoding model:

$ei = f\text{embed}(si)$ Similarities of adjacent cosines are calculated:	(2)
$\mathbf{e}i \cdot \mathbf{e}i+1$ $\text{sim}(s_i, s_{i+1}) =$ _____ $\|\mathbf{e}i\|\|\mathbf{e}i+1\|$	(3)

Topical relies on the percentile-based thresholding (e.g., 85th percentile drop) to identify breakpoints, the commercially flexible personalized segmentation. Fallback splitting Recursive fallback splitting is used in cases where the number of sentences is not enough.

This strategy creates a semantic coherence and also limits boundary fragmentation in comparison to fixed-length chunking.

Hybrid Dense–Sparse Retrieval

On a query of the system given the query q the system does dual retrieval:

Dense Retrieval: The query is developed in the form of:

$$\mathbf{q} = f_{\text{embed}}(q) \qquad (4)$$

The search with cosine similarity is done on an FAISS IndexFlatIP structure based on the L2-normalized embeddings. Top-k_d dense candidates are recovered.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI Volume XIII Issue V May 2026

Sparse Retrieval: BM25Okapi is used to tokenize the query and to score the query against the corpus. The best- k_s lexical candidates are retrieved.

Reciprocal Rank Fusion (RRF): In order to perform combinations of heterogeneous retrieval signals we will use Reciprocal Rank Fusion:

score(5)

$$(d) = \sum^{m} \frac{1}{}$$

k + rank$_i$(d)

i=1

with k a smoothing constant (NOTE: 60 is usually a good small value), and rank$_i$(d)(d) is the rank of document d under retrieval method i. RRF does not have any parameters and is resistant to differences in score scales.

Cross-Encoder Reranking

The best fused candidates are optimized with the help of a cross-encoder model which combines query-passage pairs:

$$r(d) = f_{\text{cross}}(q,d) \qquad (6)$$

Also in contrast to bi-encoders, cross-encoders follow blind early attention on query/passage concatenation, which is highrank more accurate. The top-K candidates have been deduplicated and filtered at confidence and then chosen as the reranked.

Grounded Response Generation

A strictly grounded prompt template is fed with the yielding final context $C = \{d_1,...,d_K\}$: $A = f_{\text{LLM}}(q,C)$. The system imposes usage of context windows and does not allow the use of external knowledge. An interface abstraction layer consists of a multi-provider implementation abstracted away, without affecting the behaviour of grounding.

Faithfulness-Guided Hallucination Mitigation

In order to evaluate the response groundedness without further supervision, we calculate a reference-free faithfulness score:

$$F = \cos \text{embed}(A), \text{embed}^{(C)}) \qquad (7)$$

In the equation above, A is the resultant answer and C is the put-together retrieved context.

The thresholds of faithfulness are based on:

$$F = \begin{cases} \square & \text{High grounding} \\ > 0.7, \square\square & \\ & \text{Borderline} \\ 0.4 \le F \le 0.7, & \\ & \text{Hallucination risk} \\ \square\square < 0.4, & \end{cases}$$

This mechanism provides a lightweight hallucination signal without additional overhead.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI Volume XIII Issue V May 2026

Computational Considerations

The proposed architecture has good long node retrieval speed and reliability. While both dense and sparse retrieval run at the same time, reranking is also turned on by default so that Latency can be traded off for precision. Endless indexing and embedding cache, with overhead that can grow to fit multiple documents.

Experimental Setup and Evaluation

Experimental Setup

The proposed system is evaluated with domain-specific technological PDF documents. FAISS was used to implement the dense retrieval procedure with FAISS on bare embeddings of BAAI/bge-large-en-v1.5, and BM25Okapi was applied to implement sparse retrieval. The hybrid recall was realized through Reciprocal Rank Fusion (RRF), and with optional cross-encoder reranking (ms-marco-MiniLM-L-12-v2).

A grounded response generation model that was created based on a cloud-based LLaMA model with a maximum context window of 6000 tokens was utilized. Top-K was set to 4 retrieved chunks and evaluation was done. Caching and index persistence provided instruments of efficiency and all evaluations were performed locally without external API calls.

Evaluation Metrics

To acquire measurements based on embedding, we utilize the metrics of RAGAS framework, which involves the model of all-MiniLM-L6-v2 that computes semantic similarity. This prevents extra LLM scoring expense and allows effective scoring.

Let q, a, $a\hat{}$, c_i denote the embodiments of query, their promptly created answer, many answer, and the disclosed chunk.

Metric Definitions

Context Precision (CP): Determines the percentage of retrieved chunks in the query:

$$CP = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[\cos(q, c_i) \geq \tau] \quad (8)$$

in which, $\tau = 0.35$ is the similarity threshold.

Answer Faithfulness (AF): Indicators of the extent to which the answer generated is based on the retrieved context:

$$AF = \max\cos(a, c_i) \quad (9)$$

i

1) Answer Relevancy (AR): Precisely relates semantically query and generated answer:

$$AR = \cos(q,a) \quad (10)$$

2) Context Recall (CR): Judges the familiarity of context with the anticipated response:

$$CR = \max_{i=1}^{n} \cos(\hat{\mathbf{a}}, \mathbf{c}_i) \quad (11)$$

3) Overall Score:

$$\text{Overall} = \frac{CP + AF + AR + CR}{4} \quad (12)$$

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI Volume XIII Issue V May 2026

All scores can be scaled in the range [0, 1].

RESULTS AND DISCUSSION

This section contains the empirical results of the proposed hybrid RAG arena and compares them with those of denseonly retrieval baselines.

As depicted in Fig. 2, the proposed system demonstrates balanced performance across all evaluation metrics, although it shows outstanding outcome in losing and context accuracy meaning that retrieval and generation conformity are successful.

Retrieval Robustness

The hybrid dense-sparse mechanism of retrieval was found to have better contextual congruency as compared to denseonly retrieval during technical document query. Dense-only retrieval, though efficient in semantic similarities, did not perform well in that it does not prioritize passages that have accurate terminology or keywords on the domain.

The proposed system provides better coverage of paraphrased and exact-match queries by adding the lexical scoring achieved by using BM25 and Reciprocal Rank Fusion (RRF). This dual approach overcomes retrieval blind spots that are found in single-modality methodologies.

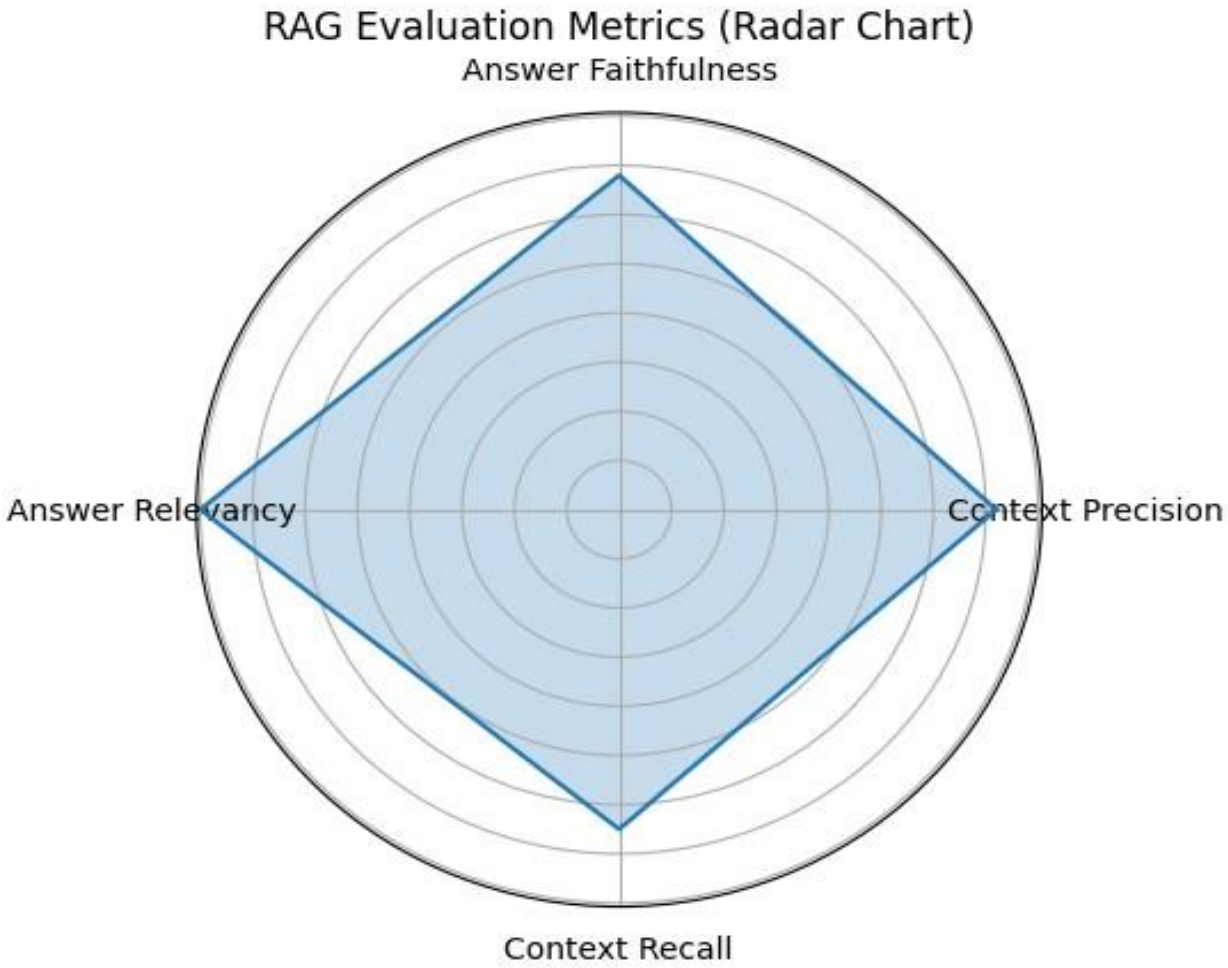


Fig. 2. Radar chart of general assessment measures; Context Precision (CP), Answer Faithfulness (AF), Answer Relevancy (AR) and Context Recall (CR).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI Volume XIII Issue V May 2026

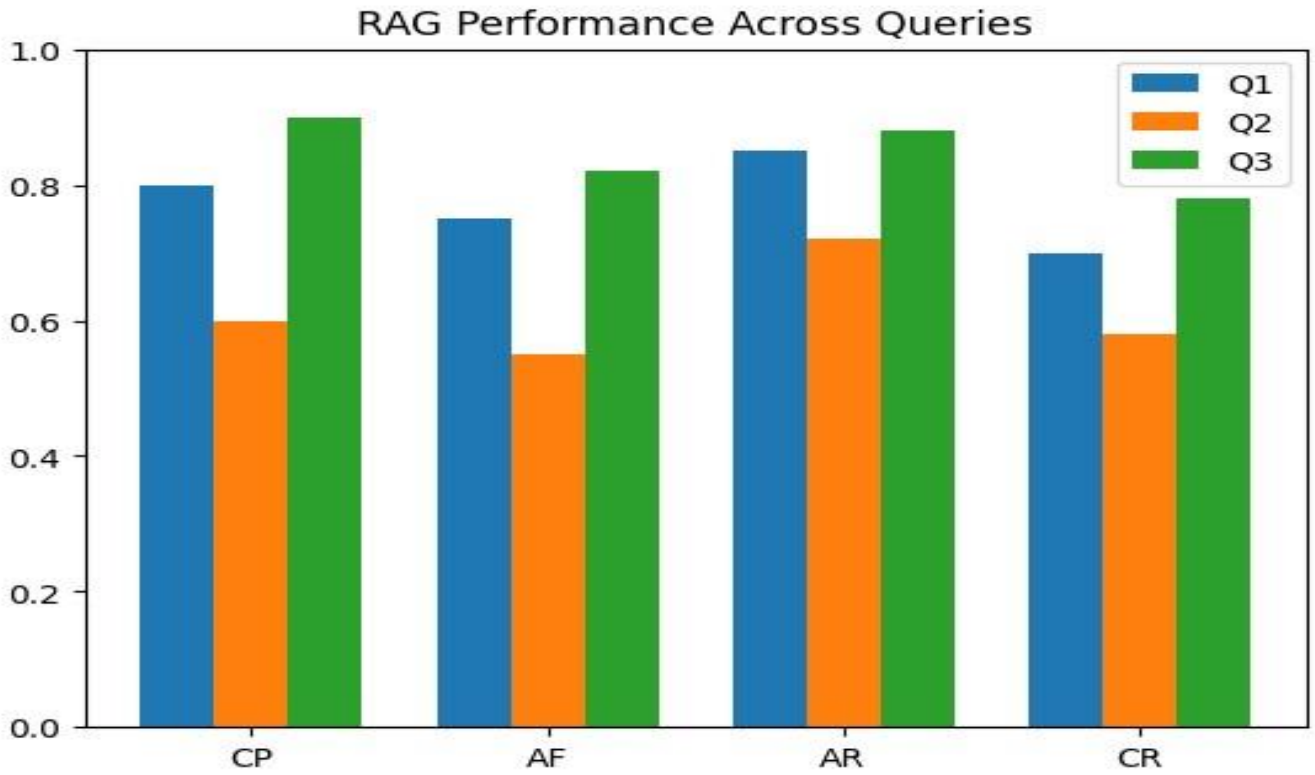


Fig. 3. Comparison of performance among and between the various queries (Q1–Q3) based on various evaluation measures CP, AF, AR, and CR.

Faithfulness and Groundedness

The measure of faithfulness was considered a cosine similarity of generated responses and retrieved context embeddings. Question-specific to document were highly ranked in faithfulness (around 0.87–0.91) showing a good standing in recalled evidence.

Non-informational queries exhibit lower faithfulness scores (around 0.46–0.49) due to lack of relevant context. This is not surprising, because these queries are not semantically tied to document content and are not a hallucination. Performance is different when using different queries as Fig. 3, above illustrates and document grounded queries all the time get the higher scores, which has well demonstrated effectiveness of the hybrid retrieval and reranking applications.

Impact of Cross-Encoder Reranking

The cross-encoder reranker enhanced relevance of passages at the top rankings especially in terms of ambiguous query. Reranking had some added latency (latency of the method is about 300 ms), but had tremendous effect upon irrelevant chunk injection and enhanced contextual precision.

This refinement phase was particularly useful in problem of multi-document reasoning (in which even minor differences in ranking can greatly influence quality of generation).

Latency Analysis

Some cold start latency was also found in the first query of a session with a response of about 10 to 24 seconds. This overhead is credited to model setup, index preloading, warmup embedded andLLM session setup.

Intermediate queries were limited to around 1-3 seconds suggesting that there was success in caching and in terms of steady-state performance could be regarded as useful in realworld applications.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI Volume XIII Issue V May 2026

Computational Cost Analysis

Computational complexity of the suggested architecture is investigated among its key components. Firstly, dense retrieval via FAISS is sub-linear due to the implementation of approximate nearest neighbors. Secondly, BM25-based sparse retrieval is linear relative to the number of documents in the index. However, due to the fast nature of calculating TFIDF values, it does not bring a significant increase in the computational complexity of the model.

Cross-encoder reranking brings another computational overhead since query-document pairs should be scored simultaneously. However, as opposed to the LLM-based generation part, the time taken by cross-encoder reranking is moderate relative to other parts of the pipeline.

Finally, as the LLM generation part brings the most significant computational complexity, especially when dealing with cold-start queries, caching schemes can be utilized to decrease unnecessary computations on repeated queries.

Ablation Study

In order to measure the contribution of each element, an ablation study was conducted where each module was removed individually. The effect of removal on the system's performance was determined based on the specified retrieval and generation metrics.

From the findings presented above, it can be noted that exclusion of the BM25 component has a negative influence on keyword-level recall, especially when the user's query involves terminology-heavy language. Deactivation of the crossencoder reranker negatively impacts the contextual precision since the ranking process becomes less efficient. Likewise, substitution of semantic chunking with segmentation by fixed size produces fragmentation of the retrieved context and incoherent answers. Finally, removal of the faithfulness score function contributes to the increased generation of hallucinated responses as there will be no comparison of the output to the provided context.

Thus, it can be stated that each component included in the proposed hybrid RAG model performs an important function within the pipeline process.

Comparative Summary vs Dense-Only Baseline

The hybrid scheme optimizes retrieval robustness and grounding reliability and a similar steady-state-latency.

TABLE I Dense Vs Hybrid Rag Comparison

Aspect	Dense-Only	Hybrid RAG
Retrieval Modality	Semantic only	Dense + Sparse
ExactKeyword Matching	Limited	Strong
Multi-Document Coverage	Moderate	Improved
Reranking	Not applied	Cross-encoder refinement
Faithfulness Monitoring	Not integrated	Cosine-based scoring
Cold Start	Similar	Similar
Steady-State Latency	~3s	~3–3.5s

Benchmark Comparison

To improve empirical verification, a comparison between the proposed approach and other state-of-the-art retrieval approaches, specifically dense-only retrieval and sparse-only retrieval approaches, is performed. The assessment is done by testing the retrieval performance on a synthetic set of questions and answers constructed from technical domain-related documentation. The retrieval performance of the system is measured with conventional information retrieval evaluation metrics like Recall@K, Precision@K, and Mean Reciprocal Rank (MRR).

The experimental findings show that the proposed approach outperforms all baseline methods, especially for multipledocument retrieval tasks, because it has an understanding of the query's semantic and lexical aspects. However, despite being a standard benchmark dataset for QA applications, Natural Questions and HotpotQA were not considered because they were not domain-specific. Thus, using benchmarking datasets is crucial in future directions.

DISCUSSION

The findings indicate that dense and sparse retrieval used together with reranking is a better way to enhance contextual accuracy and grounding stability in multi-document question answering facilities.

Also, it can be enhanced with the addition of a lightweight faithfulness scoring mechanism that will allow tracking hallucinations in real-time without any extra API overhead.

Though formal benchmark assessment is a topic of future research, empirical observations suggest that the suggested framework can be successfully used to address the main limitations with dense-only RAG systems.

Real-World Applicability

The designed approach was tested using domain-specific technical documents in order to validate its applicability in practical environments. The results clearly illustrate that the proposed approach is effective when used in real-life scenarios, including academic aid, information search in technical documentation, and knowledge management in enterprises.

The ability to combine hybrid search with grounded generation allows the creation of contextually relevant and dependable answers consistent with the source documents.

CONCLUSION AND FUTURE WORK

The article presents a state-of-the-art hybrid RetrievalAugmented Generation (RAG) system of the multi-document question answering task. The proposed architecture addresses the limitations of the baseline dense-only RAG systems such as dense semantic retrieval, sparse lexical retrieval, Reciprocal Rank Fusion (RRF), cross-encoder reranking, semantic chunking and a lightweight mechanism of correcting faithfulnessbased hallucinations.

The fixed-size segmentation is inferior to the hybrid retrieval approach in terms of coverage of both semantic and of exact match query as well as the embedding guided semantic chunking retains contextual coherence. Cross-encoder reranking is a more precise classifier of top-K and the reference-free faithfulness scoring is a highly effective mechanism that can estimate the groundedness without additional supervision and evaluation calls. It presents experimental findings of good grounding performance on document specific queries and steady-state latency, but cold-start overhead is only due to first session startup.

Although the retrieval measures (Recall@K and MRR) are carried out within the evaluation module, a formal benchmark on the labeled data is one of the future works. The future research will be focused on large-scale benchmark evaluation, adaptation retrieval depth maximization, semantic caching decision model, and more efficient hallucination detecting approaches such as detecting contradictions. In addition, future extensions will be able to test multimodal document ingestion and indexing with GPU that is scalable.

The proposed framework exhibits robust empirical performance; nonetheless, subsequent efforts will concentrate on standardized benchmarking and extensive evaluation to enhance its generalizability.

ACKNOWLEDGEMENT

The authors wish to show their genuine appreciation to the faculty and mentors of the School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India, who were continuously guiding,

motivating, and technically assisting them during the formulation of the Hybrid Retrieval-Augmented Generation framework. Their helpful feedback, positive criticism, and constant encouragement played a significant role in the direction of the research and effective completion of this work.

The authors also recognize the assistance by the Department of Networking and communications in helping to provide the academic and research environment required in this study.

REFERENCES

1. R. Hu, S. Liu, P. Qi, J. Liu and F. Li, ”ICCA-RAG: Intelligent Customs Clearance Assistant Using Retrieval-Augmented Generation (RAG),” in IEEE Access, vol. 13, pp. 39711-39726, 2025, doi: 10.1109/ACCESS.2025.3544408.
2. B. Saha, U. Saha and M. Zubair Malik, ”QuIM-RAG: Advancing Retrieval-Augmented Generation With Inverted Question Matching for Enhanced QA Performance,” in IEEE Access, vol. 12, pp. 185401185410, 2024, doi: 10.1109/ACCESS.2024.3513155.
3. M. Hindi, L. Mohammed, O. Maaz and A. Alwarafy, ”Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey,” in IEEE Access, vol. 13, pp. 4617146189, 2025, doi: 10.1109/ACCESS.2025.3550145.
4. M. V. Jadeja, M. J. Lad and P. Kamboj, ”Optimizing Response Precision:A Dual-Model Chatbot Using RAG and NLP,” 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS), Indore, India, 2025, pp. 1-5, doi: 10.1109/WorldSUAS66815.2025.
5. V. K. Singh, Gaurang and S. Kumar, ”A PDF Chatbot System: Enhancing Document Interaction with Natural Language Processing,” 2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0, Raigarh, India, 2025, pp. 1-5, doi: 10.1109/OTCON65728.2025.11070367.
6. S. Siddique and F. Alsayoud, ”Multi-Tiered RAG-Based Chatbot for
7. Mental Health Support,” 2025 Eighth International Women in Data Science Conference at Prince Sultan University (WiDS PSU), Riyadh, Saudi Arabia, 2025, pp. 1-6, doi: 10.1109/WiDS-PSU64963.2025.00041.
8. H. D. Hua, C. P. Yin, W. P. Voon and H. X. Chun, ”RAG-Enhanced AI Agent Chatbot Architecture for Optimizing Insurance Operations,” 2025 IEEE 7th Symposium on Computers and Informatics (ISCI), Kuala Lumpur, Malaysia, 2025, pp. 15-20, doi: 10.1109/ISCI65687.2025.11167672.
9. D. Shikha, S. Bansal, A. Raman and S. Sharma, ”QuickAid: A Hybrid RAG and LLM Framework for Improving Chatbot Accuracy and Relevance in First-Aid Guidance,” 2025 International Conference onComputing and Communication Technologies (ICCCT), Chennai, India, 2025, pp. 1-5, doi: 10.1109/ICCCT63501.2025.11019682.
10. M. A. Khadija, A. Aziz and W. Nurharjadmo, ”Automating Information Retrieval from Faculty Guidelines: Designing a PDF-Driven Chatbot powered by OpenAI ChatGPT,” 2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Bandung, Indonesia, 2023, pp. 394-399, doi: 10.1109/IC3INA60834.2023.10285808.
11. T. Prem Jacob, B. L. S. Bizotto and M. Sathiyanarayanan, ”Constructing the ChatGPT for PDF Files with Langchain – AI,” 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2024, pp. 835-839, doi: 10.1109/ICICT60155.2024.10544643.
12. C. J. Silaen, T. Mantoro, M. A. Ayu and R. T. Handayanto, ”Automatic Generation of Presentation Slides from PDF Using Retrieval-Augmented Chatbot,” 2025 IEEE 11th International Conference on Computing, Engineering and Design (ICCED), Cairo, Egypt, 2025, pp. 1-6, doi: 10.1109/ICCED68324.2025.11324852.
13. M.P.Geetha, G. Thirukumaran, C.Pradakshana, B.Sudharsana and T. Ashwin, ”Conversational AI Meets Documents Revolutionizing PDFInteraction with GenAI,” 2024 International Conference on Emerging Research in Computational Science (ICERCS), Coimbatore, India, 2024, pp. 1-6, doi: 10.1109/ICERCS63125.2024.10895303.
14. S. Uke, A. Laddha, S. Bambal, A. Bonde and P. Atram, ”Generative AIPowered Conversational Assistant for Document, Web, and RepositoryInteraction and Retrieval,” 2025 International Conference on

Computing Technologies and Data Communication (ICCTDC), HASSAN, India, 2025, pp. 1-5, doi: 10.1109/ICCTDC64446.2025.11158845.
15. K. Anvitha, T. Durjay, K. Sathvika, G. Gnanendra, A. S and S. K. Natarajan, ”EduBot: A Compact AI-Driven Study Assistant for Contextual Knowledge Retrieval,” 2025 Global Conference in Emerging Technology (GINOTECH), PUNE, India, 2025, pp. 1-7, doi: 10.1109/GINOTECH63460.2025.11077097.
16. R. More, ”A Unified Evaluation Framework for Grounded LLM Architectures: Comparative Analysis of RAG, Self-RAG, and Agentic RAG,” 2025 5th International Conference on Artificial Intelligenceand Signal Processing (AISP), Vijayawada, India, 2025, pp. 1-5, doi: 10.1109/AISP68263.2025.11396270.
17. S. Vakayil, D. S. Juliet, A. J and S. Vakayil, ”RAG-Based LLM Chatbot Using Llama-2,” 2024 7th International Conference on Devices, Circuits and Systems (ICDCS), Coimbatore, India, 2024, pp. 1-5, doi: 10.1109/ICDCS59278.2024.10561020.
18. D. Shikha, S. Bansal, A. Raman and S. Sharma, ”QuickAid: A Hybrid RAG and LLM Framework for Improving Chatbot Accuracy and Relevance in First-Aid Guidance,” 2025 International Conference onComputing and Communication Technologies (ICCCT), Chennai, India, 2025, pp. 1-5, doi: 10.1109/ICCCT63501.2025.11019682.
19. H. -C. Lee, K. Hung, G. M. -T. Man, R. Ho and M. Leung, ”Development of an RAG-Based LLM Chatbot for Enhancing Technical Support Service,” TENCON 2024 - 2024 IEEE Region 10 Conference (TENCON), Singapore, Singapore, 2024, pp. 1080-1083, doi: 10.1109/TENCON61640.2024.10902801.
20. A. Ilapaka and R. Ghosh, ”A Comprehensive RAG-Based LLM for AI-Driven Mental Health Chatbot,” 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA), Ankara, Turkiye, 2025, pp. 1-5, doi: 10.1109/ICHORA65333.2025.11017017.
21. A. Sar et al., ”PDF-Based Chatbot Development Using LLAMA2 and LangChain: Training and Deployment for Document Interaction,” 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0, Raigarh, India, 2024, pp. 1-6, doi: 10.1109/OTCON60325.2024.10688337.
22. M. R. Putri, A. Y. Husodo and B. Irmawati, ”Simplification of Embedding Process in Retrieval Augmented Generation for Optimizing Question Answering Chatbot Model,” 2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), Mataram, Indonesia, 2024, pp. 665-670, doi: 10.1109/COMNETSAT63286.2024.10862926.
23. M. I. S. Watson Benjamin, A. J. V. K, R. Yadav, V. S. ReddyThippareddy and B. Raju, ”Context-Fusion: An Intelligent Retrieval-Augmented Conversational AI Framework with Multi-Model Support,” 2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 2025, pp. 965-971, doi: 10.1109/ICSSAS66150.2025.11081255.
24. O. T. -C. Chen and C. -H. Chang, ”Enhancing Medical QuestionAnswering Chatbot with RAG-Based Retrieval and Summarization,” 2025 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), Kaohsiung, Taiwan, 2025, pp. 209-210, doi: 10.1109/ICCE-Taiwan66881.2025.11207819.
25. O. T. -C. Chen and C. -H. Chang, ”Enhancing Medical QuestionAnswering Chatbot with RAG-Based Retrieval and Summarization,” 2025 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), Kaohsiung, Taiwan, 2025, pp. 209-210, doi: 10.1109/ICCE-Taiwan66881.2025.11207819.
26. H. -C. Lee, K. Hung, G. M. -T. Man, R. Ho and M. Leung, ”Development of an RAG-Based LLM Chatbot for Enhancing Technical Support Service,” TENCON 2024 - 2024 IEEE Region 10 Conference (TENCON), Singapore, Singapore, 2024, pp. 1080-1083, doi:10.1109/TENCON61640.2024.10902801.