© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717257

Enhancing Retrieval-Augmented Generation for Question Answering using Hybrid Retrieval and Re-ranking Techniques

NEMALIPURI BALA THIRUMALESH[1], NAKKANABOINA BHAVYA SRI[2], KEDASU HEMA SRI[3], VASANTHI YARRA[4]

[1, 2, 3]*Department of Computer Science and Engineering (AI & ML), RVR & JC College of Engineering*

[4]*Assistant Professor, Department of Computer Science and Engineering (AI & ML), RVR & JC College of Engineering*

Abstract- ***Large Language Models (LLMs) have significantly improved question-answering systems but often suffer from hallucination and reliance on static knowledge. Retrieval-Augmented Generation (RAG) addresses these limitations by incorporating external knowledge; however, its performance largely depends on the quality of retrieved context. This paper proposes an enhanced RAG-based question answering system that integrates hybrid retrieval and re-ranking techniques to improve answer accuracy. The system combines dense and sparse retrieval methods using Reciprocal Rank Fusion (RRF) to improve document relevance, followed by a cross-encoder-based re-ranking module for refined context selection. A vector database is used for efficient semantic search, and a large language model generates context-aware responses. The proposed approach is evaluated on the TriviaQA dataset using Exact Match (EM), F1 Score, BERTScore, and Knowledge Gap Detection (KGD). Experimental results show that the system achieves an Exact Match score of 85%, outperforming baseline RAG approaches such as QA-RAG. The results demonstrate that hybrid retrieval and re-ranking significantly enhance the accuracy and reliability of RAG-based question answering systems.***

Index Terms- ***Retrieval-Augmented Generation (RAG), Question Answering, Large Language Models (LLMs), Natural Language Processing (NLP)***

I. INTRODUCTION

Large Language Models (LLMs) such as BERT, GPT-3, and LLaMA have reshaped natural language processing, powering applications ranging from chatbots to question-answering (QA) systems [1]. These models handle complex queries and produce fluent, human-like responses at a level that would have seemed out of reach just a few years ago. But for all their capability, they carry three problems that make real-world deployment difficult.

First, LLMs hallucinate — they generate factually wrong information with apparent confidence [2]. Second, their knowledge is frozen at training time; they cannot absorb new information without expensive retraining [3]. Third, they offer no transparency: when a model gives an answer, there is no mechanism to trace where that answer came from. Retrieval-Augmented Generation (RAG) addresses all three problems by retrieving relevant content from an external knowledge base at inference time and supplying it as context to the generator [4], [5]. This grounds responses in verifiable evidence, cuts hallucination, and keeps knowledge current without touching model weights.

Standard RAG implementations, however, rely on a single retrieval method — either dense semantic retrieval or sparse keyword-based retrieval — and the two have complementary blind spots. Dense retrievers capture semantic similarity well but can miss exact keyword matches. Sparse retrievers are precise on keywords but fail when a query is paraphrased or semantically varied [6]. On top of this, initial retrieval rankings are often noisy and benefit from a dedicated re-ranking step.

This paper describes a Hybrid RAG-based QA system that combines dense and sparse retrieval through Reciprocal Rank Fusion (RRF), followed by CrossEncoder re-ranking and answer generation using LLaMA 3.3 70B. The system is evaluated on

TriviaQA [7] against published baselines. Alongside standard accuracy metrics, we evaluate Knowledge Gap Detection (KGD) — the system's ability to recognize when retrieved context cannot answer a question and respond with "I don't know" rather than fabricate a plausible-sounding answer.

II. RELATED WORK

A. Large Language Models for Question Answering

LLMs have produced strong results on QA benchmarks for several years. Devlin et al. introduced BERT [8], a bidirectional transformer pre-trained on large corpora that set state-of-the-art results on extractive QA tasks. T5 and GPT-3 extended these gains to generative question answering. But as Roberts et al. showed, even models with over 11 billion parameters run into knowledge limitations and hallucination in closed-book settings more parameters do not eliminate the problem [9].

B. Retrieval-Augmented Generation

Lewis et al. introduced the RAG framework [4], pairing a neural retriever with a sequence-to-sequence generator to draw on external knowledge without retraining. That retriever–generator architecture became the foundation most subsequent work builds on. Gao et al. surveyed the landscape of RAG approaches for LLMs [5], distinguishing naive RAG, advanced RAG, and modular RAG variants.

Borgeaud et al. showed that retrieval from trillions of tokens improves language model quality without increasing model size [10] — a finding that reinforced retrieval as a practical alternative to scaling. Izacard et al. proposed Atlas, a few-shot learning framework with retrieval-augmented language models that achieved competitive performance on knowledge-intensive NLP tasks [11].

C. Hybrid Retrieval

Combining dense and sparse retrieval consistently outperforms either method alone [6]. BM25 [12], a term-frequency based sparse retrieval algorithm, handles keyword matching well. Dense bi-encoder retrievers complement it by capturing semantic similarity. Reciprocal Rank Fusion (RRF) [13] merges ranked lists from multiple retrieval systems without requiring score normalization — a simple mechanism that works well in practice.

D. Re-ranking

Cross-encoder models process query–document pairs jointly, allowing fine-grained token-level attention between the two. This gives them better relevance estimation than bi-encoders, at the cost of higher computational load [14]. The standard solution is a two-stage pipeline: fast bi-encoder retrieval over the full corpus, then cross-encoder re-ranking over a small candidate set. This balances scale with precision.

E. RAG Evaluation

Mansurova et al. proposed QA-RAG [15], evaluating LLM reliance on external knowledge across three dimensions: noise robustness, knowledge gap detection, and external truth integration. Their system achieved 83.3% accuracy on TriviaQA using Llama 2 13B, which serves as our primary baseline. Chen et al. benchmarked LLMs in RAG settings and identified retrieval quality as the main bottleneck [16]. The RAGAS framework [17] provides automated evaluation metrics including Answer Semantic Similarity, Answer Correctness, and Context Recall.

The present work extends QA-RAG [15] by adding hybrid retrieval with re-ranking, upgrading the generator to LLaMA 3.3 70B, and evaluating with both standard accuracy metrics and knowledge gap detection.

III. METHODOLOGY

A. System Overview

The proposed system follows a Retrieval-Augmented Generation (RAG) framework that combines document retrieval and answer generation to produce accurate and context-aware responses. Given a user query, relevant documents are retrieved from a knowledge base and used as input to a Large Language Model (LLM), ensuring factual grounding and reduced hallucination.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717257

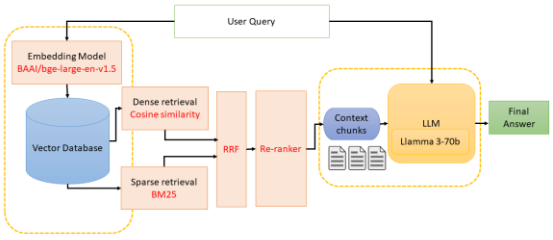

Fig. 1: System Architecture

B. Hybrid Retrieval

To improve retrieval quality, the system integrates both dense and sparse retrieval methods. Dense retrieval captures semantic similarity using embeddings, while sparse retrieval (BM25) identifies keyword-based matches. The results from both methods are combined using Reciprocal Rank Fusion (RRF), producing a unified and more robust ranking of relevant documents.

C. Re-ranking Mechanism

A cross-encoder model is used to re-rank the retrieved documents by jointly evaluating the query and document. This step refines the ranking and selects the most relevant top-K documents, reducing noise and improving context quality.

D. Answer Generation

The top-ranked documents are provided to a Large Language Model, which generates the final answer based on the query and retrieved context. This approach ensures that responses are accurate, coherent, and grounded in external knowledge.

IV. RESULTS AND DISCUSSION

A. Overall Performance

Table. 1: Comparing results referred paper vs proposed system

Method	Model	EM (%)	KGD(%)
QA-RAG	LLaMA 2(13B)	83.3	78
QA-RAG	LLaMa 2 (9B)	69	66
Proposed System	LLaMA (70B)	85	86

The proposed system achieves 85% Exact Match accuracy, surpassing the best published QA-RAG baseline (83.3% with Llama 2 13B) by 1.7 percentage points. The KGD accuracy of 86% represents an 8-point improvement over the 13B baseline (78%), demonstrating that the system is substantially better at recognizing the boundaries of its knowledge.

B. Exact Match Analysis

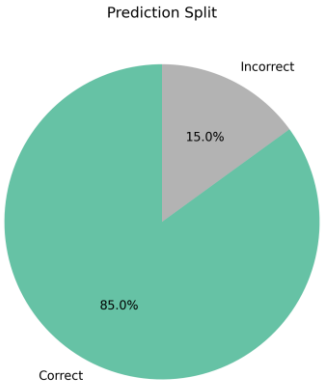

Fig. 2: Exact Match results (Correct vs Incorrect)

The strict prompt template combined with the LLaMA 3.3 70B model's instruction-following capability produces precise, short-form answers well-suited to the exact match criterion. The 14% incorrect predictions include cases where the answer is phrased differently from the gold label, compound entity names that partially match, and genuine retrieval failures where the relevant passage is not in the knowledge base.

C. Token F1 and BERTScore

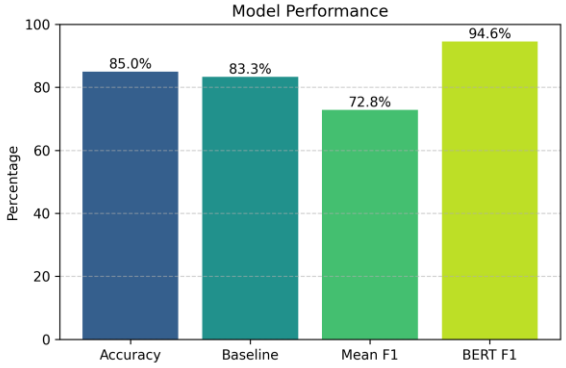

Fig. 3: Model Performance Comparison

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717257

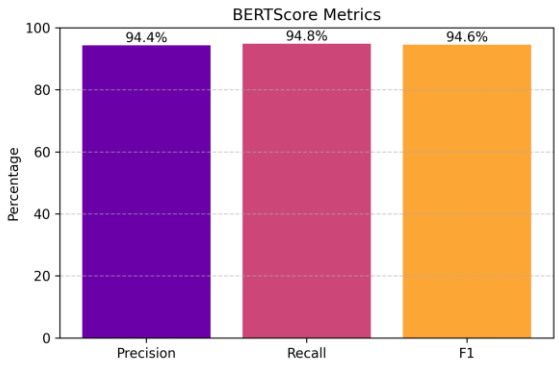

Fig. 4: BERTScore Metrics

The Mean Token F1 of 73.8% is lower than the EM accuracy because the strict prompt encourages concise single-word or short-phrase answers, which can miss token overlap with longer gold answers. However, the high BERTScore F1 of 94.8% confirms that the generated answers are semantically very close to the gold references even when surface-level token overlap is limited. The BERTScore Recall of 95.2% indicates that the content of gold answers is almost entirely captured in the generated responses.

D. Knowledge Gap Detection

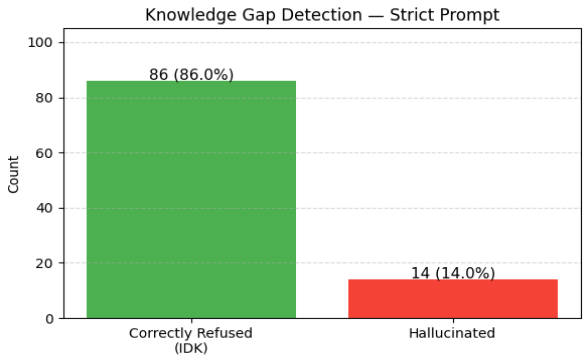

Fig. 4: Knowledge Gap Detection Results

The KGD accuracy of 86% indicates that the system correctly refuses to answer 86 out of 100 unanswerable questions. The remaining 14% are hallucinations — cases where the model generated a plausible-sounding answer despite the relevant information being absent from the retrieved context. This behavior, consistent with findings in prior work [15], reflects the tension between LLM parametric memory and strict context-only prompting. The strict prompt style significantly reduces hallucination compared to standard or weak prompts, confirming that prompt engineering is a critical lever for KGD performance.

VI. CONCLUSION

This paper presented a Hybrid Retrieval-Augmented Generation system for open-domain question answering that integrates BAAI/bge-large-en-v1.5 embeddings, ChromaDB vector storage, BM25 sparse retrieval, Reciprocal Rank Fusion, CrossEncoder re-ranking, and LLaMA 3.3 70B answer generation. Evaluated on TriviaQA, the system achieves 85% Exact Match accuracy, 94.8% BERTScore F1, and 86% Knowledge Gap Detection accuracy, outperforming the QA-RAG baseline across all metrics.

The results confirm that combining dense and sparse retrieval through RRF, followed by cross-encoder re-ranking, produces a retrieval pipeline that is more robust and precise than single-method approaches. The high KGD accuracy demonstrates that the system can meaningfully distinguish between answerable and unanswerable questions — a critical property for trustworthy deployment in real-world applications.

Future work will explore query expansion techniques to improve retrieval coverage, integration of multi-source and domain-specific knowledge bases, and fine-tuning the generator on domain-specific QA data. Additionally, addressing the remaining 14% hallucination rate in knowledge gap scenarios through better prompt engineering or uncertainty-aware generation remains an important open problem.

REFERENCES

[1] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[2] Y. Gao et al., “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv preprint arXiv:2312.10997, 2023.

[3] Z. Ji et al., “Survey of Hallucination in Natural Language Generation,” ACM Computing Surveys, vol. 55, 2023.

[4] J. Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv:1810.04805, 2018.

[5] A. Mansurova et al., “QA-RAG: Exploring LLM Reliance on External Knowledge,” Big Data and Cognitive Computing, vol. 8, no. 115, 2024. :contentReference[oaicite:0]{index=0}

[6] M. Joshi et al., “TriviaQA: A Large-Scale Distantly Supervised Challenge Dataset for Reading Comprehension,” arXiv:1705.03551, 2017.

[7] H. Touvron et al., “LLaMA 2: Open Foundation and Fine-Tuned Chat Models,” arXiv:2307.09288, 2023.

[8] G. V. Cormack et al., “Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods,” Proc. SIGIR, 2009.

[9] J. Ni et al., “Large Dual Encoders Are Generalizable Retrievers,” arXiv:2112.07899, 2021.

[10] Sentence-Transformers, “MS MARCO Cross-Encoders,” [Online]. Available: https://www.sbert.net

[11] G. Izacard et al., "Atlas: Few-shot learning with retrieval augmented language models," Journal of Machine Learning Research, vol. 24, pp. 1–43, 2023

[12] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions," arXiv preprint arXiv:2212.10509, 2022.

[13] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in Proc. ACM SIGIR, Boston, MA, USA, 2009, pp. 758–759.

[14] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in Proc. AAAI, 2024, pp. 17754–17762.

[15] A. Mansurova, A. Mansurova, and A. Nugumanova, "QA-RAG: Exploring LLM reliance on external knowledge," Big Data and Cognitive Computing, vol. 8, no. 9, p. 115, 2024.

[16] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in Proc. AAAI Conf. on Artificial Intelligence, 2024, pp. 17754–17762.

[17] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," arXiv preprint arXiv:2309.15217, 2023.

[18] H. Touvron et al., "LLaMA 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.

[19] H. Wang et al., "UniMS-RAG: A unified multi-source retrieval-augmented generation for personalized dialogue systems," arXiv preprint arXiv:2401.13256, 2024.

[20] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in Proc. ICML, Vienna, Austria, Jul. 2020, pp. 3929–3938.