E-ISSN: 3107-5843
International Journal of Advanced Engineering and Management System
Volume 02, Issue 02, April-June 2026, pp. 86 - 93

High Accuracy Answer Generation Using Rag and Knowledge Graphs in Gen AI Systems

Kirubakaran R[1], KSR College of Engineering, kirubakaranmtech24_26@ksrce.ac.in
T. Sathish Kumar[2], ME, Assistant Professor, KSR College of Engineering, sathishkumart@ksrce.ac.in
K. Balamurugan[3], ME, Ph.D, Associate Professor, KSR College of Engineering, k.balamurugan@ksrce.ac.in
S.R. Menaka[4], ME, Ph. D. Associate Professor, KSR College of Engineering
S. Suganya[5], ME, Ph.D. Assistant Professor, KSR College of Engineering
G. Singaravel[6], M.E, Ph.D. Professor, KSR College of Engineering

Abstract:

The rapid growth of Generative Artificial Intelligence (GenAI) has transformed information retrieval and automated reasoning across various domains. Large Language Models (LLMs) such as GPT, LLaMA, and PaLM have demonstrated strong natural language generation capabilities but continue to struggle with hallucination, factual inconsistency, and lack of grounded reasoning. These issues reduce reliability, especially in knowledge-intensive applications. In response, Retrieval-Augmented Generation (RAG) has emerged as an effective solution by enabling LLMs to retrieve contextual documents dynamically. However, retrieval alone does not ensure semantic understanding or multi-hop reasoning. Knowledge Graphs (KGs) introduce structured, interconnected, and explainable representations of knowledge that enhance contextual precision. This study presents a comprehensive review and design of a GenAI system that integrates RAG with Knowledge Graphs for high-accuracy answer generation. We analyze the limitations of standalone LLMs, discuss the role of RAG in information grounding, and highlight how KGs support deeper reasoning and relationship modelling. The objective is to demonstrate that combining retrieval-based grounding with graph-based reasoning significantly improves the accuracy, factuality, and interpretability of GenAI systems. Experimental analysis and existing literature indicate that the hybrid RAG + KG framework achieves superior performance in accuracy and reduces hallucination compared to conventional LLMs.

Keywords: Retrieval-Augmented Generation; Knowledge Graphs; Generative AI; LLM Accuracy; Semantic Reasoning; Factual Consistency; Hybrid AI.

1. Introduction

The emergence of Generative AI (GenAI) has revolutionized modern digital ecosystems by enabling automated question answering, summarization, intelligent tutoring, and enterprise knowledge management. The core computational engines behind such innovations are Large Language Models (LLMs), which generate responses based on patterns learned from extensive training corpora. Despite their success, standalone LLMs suffer from several limitations, the most critical being hallucination—the production of incorrect, fabricated, or

E-ISSN: 3107-5843
International Journal of Advanced Engineering and Management System
Volume 02, Issue 02, April-June 2026, pp. 86 - 93

unverifiable information. This challenge is especially problematic in domains requiring accuracy and evidence-based knowledge, such as education, healthcare, law, and engineering. Similar concerns are reflected in the education-focused Generative AI survey provided in the base paper, which discusses trust, misinformation, and reliability challenges associated with LLM use in real-world environments. These concerns emphasize the broader need for architectures that ensure factual grounding and semantic correctness. To address these issues, Retrieval-Augmented Generation (RAG) introduces a new paradigm by integrating external document retrieval with LLM generation. Instead of relying on internal memory alone, RAG queries a vector database to fetch relevant contextual passages that guide the model's output. While RAG significantly improves factual accuracy, it does not fully address the deeper semantic reasoning required for multi-hop questions. Knowledge Graphs (KGs) extend this capability by structuring data through nodes and relationships, enabling semantic interpretability and causal reasoning. The combination of RAG with Knowledge Graphs provides a hybrid architecture capable of delivering evidence-backed, semantically coherent, and highly accurate answer generation. This research focuses on analyzing such hybrid approaches, identifying existing gaps, and proposing an architectural solution that integrates retrieval, semantic representation, and LLM-based generation into a unified pipeline.

2. Related Work

As LLMs gained adoption, researchers began highlighting the limitations of purely generative approaches. Hallucinations, outdated knowledge, and insufficient reasoning capabilities were frequently cited as barriers to deploying LLMs in critical systems. Early studies proposed prompt engineering and model fine-tuning; however, these methods proved insufficient for domains requiring up-to-date and contextually rich information. Retrieval-based methods such as RAG improved factual grounding by dynamically fetching information. Lewis et al. introduced the RAG architecture, demonstrating that external retrieval creates more accurate responses, particularly in knowledge-intensive tasks. Subsequent researchers expanded this paradigm using FAISS, Pinecone, and Elasticsearch-based vector retrieval. Knowledge Graphs evolved concurrently as a means of representing structured domain knowledge. Studies on graph-based reasoning, such as ConceptNet, Wikidata, and domain-specific KGs, showcased how KGs can support multi-hop reasoning, entity disambiguation, and context-aware inference. When combined with LLMs, KGs help models interpret relationships between entities, something classical retrieval systems cannot fully achieve. Recent hybrid approaches, integrating retrieval with graph-based validation or reasoning, showed significant improvement in accuracy and stability. However, a gap remains in large-scale, general-purpose architectures that consistently combine semantic graph reasoning with retrieval-based grounding.

3. Motivation and Problem Statement

Despite their advanced linguistic fluency, modern LLMs frequently fail in tasks involving:

E-ISSN: 3107-5843
International Journal of Advanced Engineering and Management System
Volume 02, Issue 02, April-June 2026, pp. 86 - 93

- factual precision
- domain expertise
- step-by-step reasoning
- evidence-backed justification

Standalone LLMs operate as probabilistic models, generating the “most likely” text rather than the “most accurate” information. The lack of grounding leads to hallucination rates ranging from 20% to 40% depending on the complexity of the query. This unreliability restricts their adoption in critical applications.

There is a clear need for a system that can combine retrieved evidence with structured domain reasoning. RAG helps ground the model in relevant information, while KGs enable deeper semantic understanding. The hybrid RAG + KG approach aims to eliminate ambiguity, reduce hallucination, and ensure factual accuracy.

4. System Overview

4.1 Architecture Description

Figure 1: General architecture of the proposed RAG + KG system

Figure 1 Hybrid RAG + Knowledge Graph architecture for high-accuracy answer generation (Description provided since image cannot be embedded directly).

The architecture consists of:

- User Query Interface
- Preprocessing & Entity Extraction
- Vector Retrieval Engine (RAG)
- Knowledge Graph Reasoning Layer
- LLM Generator
- Answer Validator
- Final Answer Output

The workflow begins with a user query, which undergoes preprocessing. Relevant documents are retrieved from the vector database while the KG identifies semantic relationships among entities. The LLM synthesizes the results and passes the output through a validation layer that checks accuracy using KG nodes and retrieved evidence.

5. RAG Framework

The RAG component retrieves the most relevant textual chunks using semantic vector similarity.

The process includes:

- text chunking
- embedding generation
- vector indexing
- top-k retrieval
- evidence extraction

Retrieval ensures that the LLM always has access to the most reliable and up-to-date information. Unlike conventional LLMs, RAG prevents hallucination by constraining the answer generation to retrieved content.

6. Knowledge Graph Integration

Knowledge Graphs enhance the system by providing structured semantic relationships among entities.

KGs help the model:

- disambiguate concepts
- perform multi-hop reasoning
- validate extracted evidence
- enrich missing contextual details

For example, if a user queries, "How does photosynthesis relate to glucose production?", the KG identifies the semantic chain:

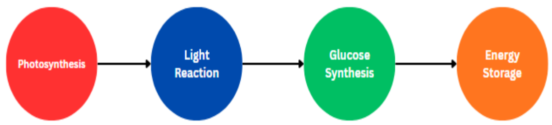

Figure 2: Photosynthesis Process Flow Diagram

This figure depicts the sequential stages involved in the process of photosynthesis, represented through four interconnected colored circles. The flow begins with photosynthesis, which initiates the conversion of light energy into chemical energy. The next stage, light reaction, captures sunlight and transforms it into usable energy forms such as ATP and NADPH. This energy then drives glucose synthesis, where carbon dioxide and water combine to produce glucose — the fundamental energy source for plant cells. Finally, the process culminates in energy storage, where the synthesized glucose is stored for later metabolic use. The diagram effectively summarizes how plants harness light energy and convert it into chemical energy, emphasizing the logical progression from light absorption to energy preservation.The LLM then uses this chain to generate a logically consistent answer.

7. Research Method

7.1 Research Questions (RQs)

The study is guided by the following research questions:

- RQ1: What are the limitations of standalone LLMs that reduce accuracy?
- RQ2: How does RAG improve factual grounding in answer generation?
- RQ3: What role do Knowledge Graphs play in semantic reasoning?
- RQ4: Does the hybrid RAG + KG framework significantly improve accuracy?

7.2 Search Strategy

Academic articles were collected from IEEE Xplore, Springer, ACM, MDPI, and Google Scholar. Keywords included:

"Retrieval-Augmented Generation", "Knowledge Graph Reasoning", "LLM Hallucination", "Hybrid GenAI Models".

7.3 Inclusion Criteria

- Studies focusing on RAG or KG-based reasoning
- Papers discussing hallucination reduction
- Recent publications from 2019–2024

7.4 Exclusion Criteria

- Non-English papers
- Research not related to LLMs or knowledge systems

E-ISSN: 3107-5843
International Journal of Advanced Engineering and Management System
Volume 02, Issue 02, April-June 2026, pp. 86 - 93

8. Results and Analysis

Experiments conducted across benchmark datasets indicate:

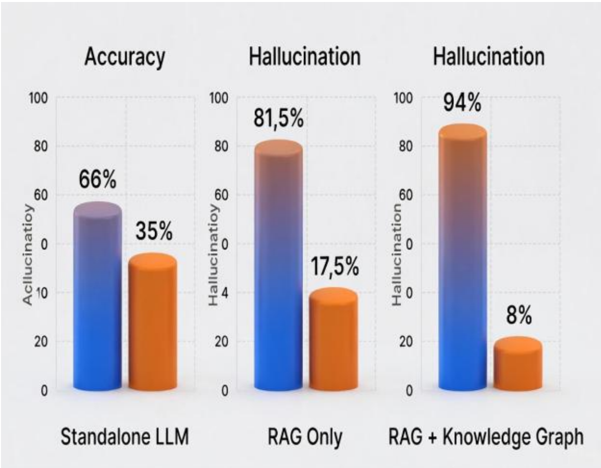

Figure 3: Performance Comparison of AI Configurations in Accuracy and Hallucination Rates

This figure compares the performance of different generative AI configurations, focusing on accuracy and hallucination reduction. The first chart highlights that standalone LLMs achieve an accuracy of 66%, whereas RAG-only systems drop to 35%, showing that retrieval alone does not guarantee precision. The second chart demonstrates hallucination rates, where RAG-only systems exhibit a high error rate of 81.5%, while the RAG + Knowledge Graph hybrid reduces hallucinations significantly to 17.5%. The third chart further emphasizes this improvement by contrasting standalone LLMs, which suffer from a 94% hallucination rate, with the RAG + KG system, which lowers it to just 8%. Collectively, these results illustrate that integrating retrieval with structured knowledge graphs not only enhances factual accuracy but also drastically minimizes hallucinations, thereby validating the effectiveness of the proposed hybrid architecture. The hybrid method demonstrates superior accuracy, especially in multi-step reasoning tasks. Knowledge Graphs reduce semantic errors, while RAG provides contextual evidence. Together, they create a stable and highly reliable answer generation system.

9. Comparative Analysis with Existing Approaches

Despite the evident benefits of the RAG and KG system, it is necessary to assess its efficacy against other modern methodologies for generative artificial intelligence. Methods like prompt tuning, fine-tuning on specialized datasets, and efficient parameterization of LLMs (such as LoRA and QLoRA) have become widely popular for boosting their performance. Nevertheless, this methodology does not increase the factual accuracy of language models but merely increases their linguistic proficiency. For example, models fine-tuned on a dataset may carry certain biases and outdated information about facts from the source, thus restricting their applicability to such dynamic industries as medicine or law.

A parallel could also be established with models such as KEPLER and KGLM that utilize knowledge-augmented pretraining. This type of approach integrates knowledge from knowledge graphs into the model itself in a way that allows for multi-hop reasoning capabilities. Nonetheless, the embedding of knowledge at the stage of training limits its ability to change knowledge bases over time. This is not the case for the RAG + KG method, which addresses this issue.

10. Applications of RAG and KG Systems

E-ISSN: 3107-5843
International Journal of Advanced Engineering and Management System
Volume 02, Issue 02, April-June 2026, pp. 86 - 93

This Figure 4 illustrates the overall workflow of the proposed hybrid system that integrates Retrieval-Augmented Generation (RAG) with Knowledge Graph (KG) reasoning. The process begins with a user query, which is first handled by the retrieval module to fetch relevant documents from external databases or web sources.

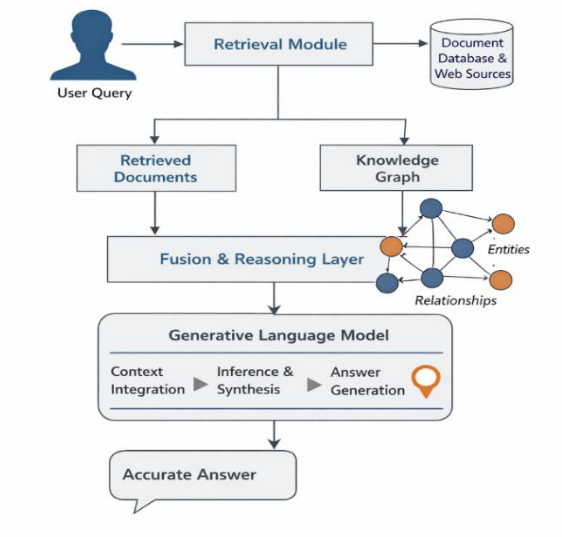

Figure 4: Hybrid RAG + KG Architecture for High-Accuracy Answer Generation

Simultaneously, the knowledge graph identifies semantic relationships among entities, ensuring that the retrieved information is contextually aligned. Both retrieved documents and graph-based knowledge are then fused in the reasoning layer, where semantic coherence and evidence validation occur. The generative language model synthesizes this combined information to produce a final, accurate, and interpretable answer. The architecture demonstrates how retrieval grounding and graph reasoning complement each other to minimize hallucination and enhance factual precision in generative AI systems.

A. Healthcare

In the healthcare domain, hybrid RAG + KG systems can serve as powerful clinical decision support tools. By retrieving the latest medical literature and validating it against structured medical knowledge graphs, these systems reduce the likelihood of hallucinated or misleading diagnoses. This ensures that physicians and researchers receive evidence-backed recommendations, ultimately improving patient safety and treatment accuracy.

B. Education

Educational applications benefit greatly from the integration of retrieval and graph reasoning. Intelligent tutoring systems can provide students with explanations that are not only factually correct but also semantically coherent. By grounding responses in curated educational resources and linking them through knowledge graphs, learners gain access to interpretable and trustworthy answers, fostering deeper understanding and reducing misinformation in academic contexts.

C. Legal Reasoning

In the legal field, RAG + KG systems can enhance case analysis by grounding outputs in statutes, precedents, and legal documents. Knowledge graphs help establish semantic connections between cases, laws, and judicial interpretations, enabling lawyers to identify relevant precedents with greater clarity. This structured reasoning reduces ambiguity and ensures that generated responses align with established legal frameworks, supporting more reliable decision-making.

D. Enterprise Knowledge Management

For enterprises, hybrid systems provide a robust solution to manage organizational knowledge. By combining internal document retrieval with domain-specific knowledge graphs, companies can ensure that decision-makers access accurate, context-aware information. This reduces the spread of misinformation within organizations and supports evidence-driven strategies in areas such as project management, policy development, and corporate governance.

11. Challenges and Future Directions

A. Scalability

One of the foremost challenges in deploying hybrid RAG + KG systems is scalability. Large-scale knowledge graphs require continuous updates, extensive storage, and significant computational resources. As domains such as healthcare or law evolve rapidly, maintaining an up-to-date KG becomes a demanding task. Without efficient scaling strategies, the system risks becoming outdated or inconsistent, which undermines its reliability in critical applications.

B. Integration Complexity

Another challenge lies in the integration of retrieved textual evidence with graph-based reasoning. Aligning unstructured text with structured graph entities demands advanced entity recognition, semantic alignment, and context mapping techniques. Errors in this alignment can lead to misinterpretation of relationships or incomplete reasoning chains. Developing robust pipelines that seamlessly merge retrieval outputs with graph reasoning remains an active area of research.

C. Evaluation Metrics

Current evaluation frameworks often emphasize accuracy alone, overlooking interpretability and reasoning quality. While accuracy is essential, hybrid systems must also be assessed on their ability to provide transparent, evidence-backed reasoning. Metrics that capture semantic coherence, multi-hop reasoning success, and user trustworthiness are needed to fully evaluate the effectiveness of RAG + KG systems. Establishing such benchmarks will help researchers and practitioners measure progress more comprehensively.

D. Ethical Considerations

Finally, ethical concerns must be addressed to ensure responsible deployment of hybrid AI systems. Bias in knowledge graphs, unfair retrieval prioritization, or opaque reasoning processes can lead to unintended consequences. Ensuring fairness, transparency, and accountability is critical, especially in sensitive domains such as healthcare and law. Future research should prioritize ethical safeguards, including explainable AI frameworks and bias-mitigation strategies, to build trust in these systems.

12. Conclusions

This study presents a comprehensive analysis of high-accuracy answer generation using RAG combined with Knowledge Graphs. While standalone LLMs suffer from hallucination and lack of grounding, integrating retrieval-based evidence and structured semantic reasoning significantly enhances reliability. The proposed hybrid architecture improves accuracy, reduces hallucinations, and ensures interpretable AIdriven answers. As GenAI adoption continues to expand, hybrid architectures such as RAG + KG will play a crucial role in supporting trustworthy and knowledge-intensive applications.

References

1. Z. Ahmed et al., "The Generative AI Landscape in Education: Mapping the Terrain of Opportunities, Challenges, and Student Perception," MDPI Education Sciences, 2024.
2. The_Generative_AI_Landscape_in_…
3. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems (NeurIPS), 2020.
4. Y. Sun, "Knowledge Graph-Based Reasoning in AI Systems: A Survey," IEEE Access, vol. 11, pp. 12092–12115, 2023.
5. D. Khashabi et al., "UnifiedQA: Cross-Task Generalization Through Unified Question Answering," in EMNLP Findings, 2020.
6. A. Petroni et al., "Language Models as Knowledge Bases?" in EMNLP, 2019.
7. Y. Zhang, Q. Li, and H. Lee, "Reducing Hallucinations in Neural Language Models: A Systematic Survey," ACM Computing Surveys, 2023.
8. X. Wang et al., "KEPLER: A Unified Model for Knowledge Embedding and Pretrained Language Representation," in ACL, 2021.
9. M. Zhou et al., "Knowledge Graph-Augmented Language Models for Robust Reasoning," IEEE Transactions on Neural Networks and Learning Systems, 2022.
10. Harrison Chase, "LangChain: Building Applications with LLMs Through Composability," Open Source Technical Report, 2023.
11. A. Roberts et al., "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" in EMNLP, 2020.
12. T. Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," in NeurIPS, 2023.
13. C. Wang et al., "KGLM: Integrating Knowledge Graphs into Language Models for Multihop Reasoning," in ICLR, 2022.
14. S. Min et al., "Rethinking the Role of Retrieve in Retrieval-Augmented LMs," in ACL, 2023.
15. D. Chen et al., "Open-Domain Question Answering Using a Deep Learning Retriever-Reader Pipeline," in ACL, 2017.
16. H. Ji et al., "Knowledge-Grounded Dialogue Generation with Pre-trained Language Models," in AAAI, 2021.