International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 11 | Nov - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Knowledge Graph-Augmented RAG for Multi-Hop Question Answering

Sanchita Dwivedi, Akhilesh A. Waoo*, Chandra Sekhar Gautam
*akhileshwaoo@gmail.com, ORCID: 0000-0001-6788-710X

ABSTRACT

Retrieval-Augmented Generation (RAG) has emerged as an effective approach for enhancing the factual grounding of Large Language Models (LLMs) by incorporating external knowledge. However, conventional RAG systems often face challenges in capturing complex relationships, integrating distributed evidence, and supporting multi-hop reasoning, particularly for complex questions that need several connected facts. Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG) addresses these limitations by integrating organized knowledge representations with semantic retrieval and language generation. This review examines recent developments in KG-RAG approaches, focusing on architectures, retrieval strategies, knowledge integration mechanisms, reasoning techniques, and evaluation methodologies. Representative frameworks, including GraphRAG, HippoRAG, KG²RAG, LightRAG, and Think-on-Graph, are analyzed to identify their methodological contributions and limitations. The review presents a taxonomy of KG-RAG approaches based on knowledge graph construction, retrieval mechanisms, evidence integration, and answer generation. It also discusses benchmark datasets and evaluation metrics relevant to retrieval effectiveness, answer faithfulness, and multi-hop reasoning. The literature synthesis identifies major challenges involving knowledge graph incompleteness, evidence reliability, hybrid retrieval, computational scalability, and domain adaptation. Prospects for future research include confidence-aware evidence fusion, adaptive retrieval strategies, reasoning-path verification, and robust domain-specific evaluation. Overall, this review provides a structured understanding of KG-RAG developments and highlights opportunities for building more reliable, efficient, and context-aware question-answering systems.

Keywords: Hybrid Retrieval, Knowledge Graphs, Large Language Models, Multi-Hop Reasoning, Question answering , Retrieval-Augmented Generation.

1. INTRODUCTION

Large Language Models (LLMs) have significantly advanced natural language processing, particularly in retrieving information and responding to inquiries. However, these models still face challenges related to factual accuracy, outdated knowledge, and understanding complex relationships within specialized domains. Retrieval-Augmented Generation (RAG) addresses some of these restrictions by incorporating external knowledge into the response generation process. Nevertheless, conventional RAG systems primarily rely on semantic similarity-based retrieval, which may struggle to identify relationships between information distributed across multiple documents (Tang & Yang, 2024).

Knowledge Graphs (KGs) provide a structured representation of information through entities and their relationships. Integrating KGs with RAG has led to Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG), which combines semantic retrieval with graph-based knowledge to support relationship-aware retrieval and multi-hop reasoning. Representative frameworks, such as GraphRAG, HippoRAG, LightRAG, and KG²RAG, differ in the ways that they build graphs, retrieve evidence, and pass it to the LLM

KG-RAG is particularly relevant to multi-hop question answering, where accurate responses often require specialized knowledge and connections among multiple concepts. However, challenges remain in knowledge graph construction, incomplete relationships, retrieval efficiency, and evidence integration (Zhou et al., 2025). Evaluating these systems also requires considering retrieval quality, respond to faithfulness, and response relevance (Saad-Falcon et al., 2024).

This review examines existing KG-RAG approaches for multi-hop question answering, focusing on their architectures, retrieval strategies, knowledge integration mechanisms, and evaluation methods. It presents a taxonomy of KG-augmented RAG techniques, discusses representative datasets and performance measures, and identifies key

limitations and research gaps. The review aims to provide a structured understanding of current developments and highlight future directions for improving the reliability, efficiency, and adaptability of KG-RAG systems.

2. REVIEW METHODOLOGY

This review adopts a structured literature-based approach to examine recent developments in Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG) for multi-hop question answering. The analysis concentrates on graph-based retrieval, hybrid knowledge integration, multi-hop reasoning, and evaluation methodologies. The selected studies are examined to identify existing approaches, their contributions, limitations, and emerging research challenges.

2.1 Research Objectives

The primary objective of this review is to examine existing Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG) approaches for multi-hop question answering. The specific objectives are:

1. To examine existing KG-RAG architectures and retrieval strategies used in question-answering systems.
2. To analyze the role of KG in supporting multi-hop reasoning and evidence integration.
3. To investigate the datasets and evaluation metrics utilized to evaluate KG-RAG performance.
4. To compare existing KG-RAG approaches based on their methodologies, contributions, and limitations.
5. To identify key research gaps, technical challenges, and potential directions for upcoming studies in multi-hop question answering.

2.2 Literature Search Strategy

The literature considered in this review covers research published between 2023 and 2026. Relevant studies are identified through academic sources such as Google Scholar, arXiv, ACL Anthology, and major conference and journal publications. The search terms include *"Knowledge Graph-Augmented RAG," "GraphRAG," "Hybrid Retrieval," "Knowledge Graph Question Answering,"* and *"Multi-Hop Reasoning."* Priority is given to studies directly related to knowledge graph integration, retrieval strategies, and question-answering performance.

2.3 Inclusion and Exclusion Criteria

Studies are considered relevant when they investigate graph-based retrieval, knowledge graph integration with LLMs, multi-hop question answering, or RAG evaluation. Research articles, conference papers, and relevant preprints published during the selected period are eligible for consideration. Studies unrelated to retrieval-augmented question answering, in addition to duplicate publications and papers lacking sufficient methodological information, are excluded.

2.4 Study Selection and Data Extraction

The selected papers are analyzed based on their proposed architectures, knowledge graph sources, retrieval mechanisms, reasoning strategies, datasets, and evaluation metrics. Their reported findings and limitations are compared to identify common methodological patterns and unresolved challenges. The extracted information is arranged into a taxonomy of KG-RAG approaches and used to support the comparative discussion and prospects for upcoming studies presented in subsequent sections.

3. FOUNDATIONS OF KNOWLEDGE GRAPH-AUGMENTED RAG

Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG) combines the contextual retrieval capabilities of traditional RAG with the structured representation of knowledge graphs. This integration enables the retrieval of both semantically relevant information and explicitly connected entities, supporting more informed

question answering and multi-hop reasoning. Recent frameworks demonstrate how graph-based knowledge organization can complement conventional retrieval methods (Edge et al., 2024; Zhu et al., 2025).

3.1 Conventional RAG

RAG enhances LLM by obtaining outside data relevant to a user's query before generating a response. A conventional RAG pipeline typically includes document preprocessing, text chunking, embedding generation, vector-based retrieval, and LLM-based answer generation. The retrieved passages provide additional context that can improve the factual grounding of responses that were produced.

However, conventional RAG systems may retrieve isolated text passages without adequately capturing relationships among entities or information distributed across multiple documents. This limitation becomes particularly important for complex questions requiring multi-hop reasoning, where multiple pieces of evidence must be connected to produce an accurate answer (Tang & Yang, 2024).

3.2 KG and LLM

KGs represent information through interconnected entities and connections, commonly expressed as triples consisting of a subject, predicate, and object. This structured representation enables systems to identify semantic connections and explore relationships that might not be explicitly preserved in text-based retrieval.

Integrating KGs with LLMs provides access to structured external knowledge that can support relational reasoning and evidence retrieval (Gutiérrez et al., 2024; Sun et al., 2024).

3.3 Knowledge Graph-Augmented RAG Architecture

KG-RAG extends conventional retrieval pipelines by incorporating graph construction, graph-based retrieval, and knowledge integration. Documents may be processed to extract entities and connections, which are subsequently arranged in a knowledge graph. During question answering, the system retrieves relevant graph entities, relationships, or subgraphs, often alongside semantically similar text passages. The evidence that was recovered is then incorporated into the LLM's context for answer generation.

Different KG-RAG frameworks implement these stages in distinct ways. Edge et al. (2024) used graph communities and hierarchical summaries to support query-focused summarization. Zhu et al. (2025) proposed knowledge graph-guided chunk expansion and organization, while Guo et al. (2025) integrated graph-based indexing with efficient retrieval mechanisms. These approaches demonstrate how graph structure can enhance the organization and selection of evidence for RAG.

3.4 Role of Knowledge Graphs in Multi-Hop Question Answering

Multi-hop question answering requires combining information from multiple interconnected entities, relationships, or documents to generate an accurate response. Unlike single-hop questions, which can often be answered using a single piece of evidence, multi-hop questions require several reasoning steps and the additon of distributed information.

Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG) supports this process by depicting entities and their relationships in a structured form. Graph-based retrieval can help identify connected evidence, while semantic retrieval can provide additional contextual data from unstructured documents. However, the efficiency of KG-RAG depends on graph completeness, retrieval accuracy, and evidence reliability. Zhou et al. (2025) examined the consequences of knowing incompleteness on KG-based RAG, emphasizing the significance of graph quality for reliable question answering.

4. LITERATURE REVIEW

Recent research on KG-Augmented Retrieval-Augmented Generation (KG-RAG) has explored the incorporation of structured knowledge, semantic retrieval, and LLM to improve retrieval information and responding to inquiries. This section describes the main studies in four groups: RAG assessment and comparison, knowledge graph-based reasoning, graph-augmented retrieval architectures, and domain-specific and benchmark studies. Each method is described once here, and later sections refer back to this section.

4.1 RAG Evaluation and Benchmarking

Research on RAG has increasingly focused on retrieval quality, multi-hop question answering, and systematic evaluation. Tang and Yang (2024) introduced MultiHop-RAG, a benchmark designed to assess the ability of retrieval-augmented systems to answer questions that require evidence from several documents. Their findings highlight the difficulties faced by conventional retrieval approaches when relevant information is distributed across interconnected sources.

Saad-Falcon et al. (2024) proposed ARES, an automated evaluation framework that assesses RAG systems using context relevance, respond to faithfulness, and relevance. More recently, Han et al. (2026) presented a systematic evaluation of RAG and GraphRAG, offering comparative insights into the capabilities and limitations of graph-based retrieval approaches.

4.2 Knowledge Graph-Based Reasoning with LLMs

Several studies use knowledge graphs to support logic and retrieval in LLMs. Sun et al. (2024) introduced Think-on-Graph, in which the LLM explores a knowledge graph step by step and uses the discovered relations to reason toward an answer. Ma et al. (2025) extended this idea in Think-on-Graph 2.0 by combining knowledge-graph guidance with retrieval from unstructured text, so that structured and unstructured evidence support each other during reasoning.

Gutiérrez et al. (2024) proposed HippoRAG, which is inspired by human long-term memory and uses graph-based associative retrieval to connect information across documents for multi-hop question answering. He et al. (2024) developed G-Retriever, which retrieves a relevant subgraph and gives it to the LLM for question answering over textual graphs.

4.3 Graph-Augmented Retrieval Architectures

A second group of studies proposes graph-based architectures to overcome the boundaries of semantic retrieval. Edge et al. (2024) introduced GraphRAG, which extracts entities and relationships from a document collection, groups them into communities, and creates hierarchical summaries to support query-focused summarization. Guo et al. (2025) presented LightRAG, which combines graph-based indexing with retrieval designed for efficient access to knowledge. Zhu et al. (2025) proposed KG²RAG, which combines semantic retrieval with knowledge graph-guided chunk expansion and organization, so that retrieved information is more relevant and coherent. Linders and Tomczak (2025) studied knowledge graph-extended RAG for question answering, which is another way to bring structured knowledge into retrieval-augmented generation.

4.4 Domain-Specific and Benchmark Studies

The use of KG-RAG in specialized question answering requires attention to graph quality, retrieval effectiveness, and evaluation. Zhou et al. (2025) evaluated KG-based RAG methods under knowledge incompleteness and examined how missing relationships affect retrieval and response quality, which shows the need to test graph-based systems under imperfect knowledge. Li et al. (2024) proposed GraphReader, which represents a long document as a graph and lets an agent explore it to answer questions that need information from many parts of the text.

Although these studies show the potential of graph-based retrieval, challenges remain in domain adaptation, incomplete knowledge graphs, multi-hop reasoning, and computational efficiency. These challenges motivate the taxonomy and the analysis in the part that follows.

5. TAXONOMY OF KG-AUGMENTED RAG

Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG) methods can be categorized according to their knowledge graph sources, retrieval strategies, knowledge integration mechanisms, and reasoning techniques. This taxonomy provides a structured framework for understanding the methodological differences among existing KG-RAG systems and their suitability for multi-hop question answering.

5.1 Overview of the Taxonomy

Existing KG-RAG approaches adopt different strategies for representing and retrieving knowledge. Some methods construct KG from unstructured documents, while others use existing structured knowledge sources. Similarly, retrieval mechanisms range from graph traversal and subgraph selection to hybrid strategies that combine graph-based and semantic retrieval. The integration of retrieved knowledge into LLMs also varies across frameworks (Edge et al., 2024; He et al., 2024; Zhu et al., 2025).

The proposed taxonomy organizes KG-RAG methods into four major dimensions: knowledge graph sources and construction, retrieval strategies, knowledge integration and evidence fusion, and reasoning and answer generation.

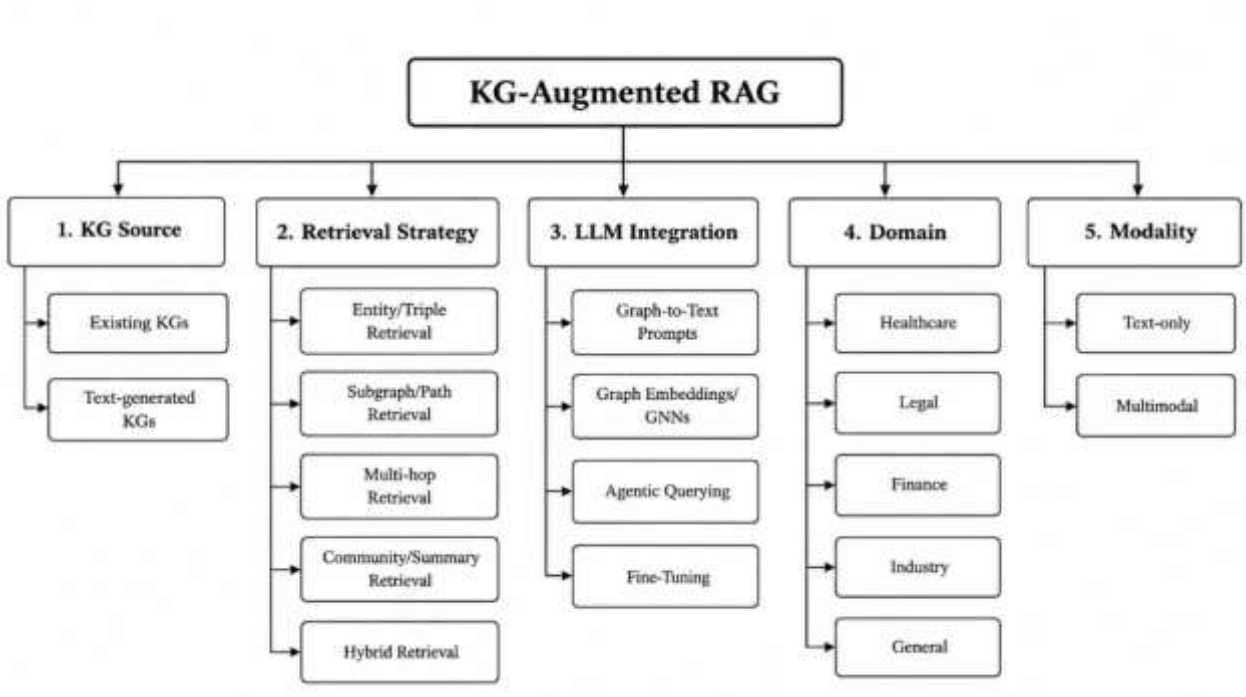


Figure 1: Taxonomy of KG-Augmented RAG

Table 1: Taxonomy of Knowledge Graph-Augmented RAG approaches

KG Sources	Retrieval Strategies	Knowledge Integration	Reasoning and Generation
Existing knowledge graphs	Graph-based retrieval	Triples and subgraphs	Single-hop QA
Document-derived graphs	Vector-based retrieval	Text chunks and summaries	Multi-hop reasoning
	Hybrid retrieval	Evidence fusion	LLM answer generation

5.2 Knowledge Graph Sources and Construction

Knowledge graph construction is a fundamental component of KG-RAG systems. Existing approaches may use predefined knowledge graphs or automatically identify linkages and items from unstructured text. For example, GraphRAG and LightRAG build graphs from document collections, whereas Think-on-Graph reasons over existing knowledge graphs.

The knowledge graph completeness and quality influence retrieval effectiveness. Incomplete relationships and inaccurate entity extraction may reduce the availability of relevant evidence, particularly for questions that require multiple reasoning steps (Zhou et al., 2025).

5.3 Graph-Based, Vector-Based, and Hybrid Retrieval

Retrieval strategies represent a major distinction among KG-RAG architectures. Graph-based retrieval identifies relevant entities, relationships, or subgraphs, while vector-based retrieval uses semantic resemblance with retrieve related text passages.

Gutiérrez et al. (2024) proposed HippoRAG, which employs graph-based associative retrieval to identify relevant information. He et al. (2024) introduced G-Retriever, which selects relevant subgraphs for graph question answering. Zhu et al. (2025) proposed KG²RAG, combining semantic retrieval with knowledge graph-guided chunk expansion and organization.

Hybrid retrieval approaches integrate structured and unstructured information to provide complementary evidence. Guo et al. (2025) explored graph-based indexing and retrieval through LightRAG, while Ma et al. (2025) combined knowledge-guided reasoning with external information retrieval in Think-on-Graph 2.0. These approaches illustrate The possibility of merging semantic relevance with explicit relationships.

5.4 Knowledge Integration and Evidence Fusion

Knowledge integration refers to the process of incorporating retrieved graph information and textual evidence into the LLM's input context. Different frameworks employ distinct mechanisms, including structured triples, retrieved subgraphs, organized text chunks, and graph-derived summaries.

Edge et al. (2024) used graph community summaries to organize information for global question answering. Zhu et al. (2025) employed graph-guided chunk organization to improve the coherence of retrieved evidence. Linders and Tomczak (2025) investigated the extension of RAG with knowledge graph information for question answering.

Evidence integration's efficacy is dependent upon the relevance, consistency, and completeness of retrieved information. Combining graph-based and textual evidence may provide broader contextual support, but irrelevant or conflicting evidence can also affect answer quality.

5.5 Reasoning and Answer Generation

Reasoning and answer generation constitute the final stage of KG-RAG systems. Retrieved knowledge is used to support responses requiring factual understanding, relationship identification, or multi-hop reasoning.

Sun et al. (2024) proposed Think-on-Graph, enabling iterative exploration of knowledge graph relationships. Ma et al. (2025) extended this approach through Think-on-Graph 2.0, incorporating knowledge-guided retrieval for more complex reasoning tasks. Gutiérrez et al. (2024) demonstrated graph-based associative retrieval in HippoRAG, while Zhu et al. (2025) organized connected evidence to support answer generation.

Despite these developments, reliable multi-hop reasoning remains dependent on graph quality, retrieval accuracy, and evidence selection. Consequently, future KG-RAG systems must address incomplete knowledge, irrelevant retrieval results, and computational burden to increase their efficiency when responding to multi-hop questions.

6. DATASETS AND EVALUATION

The evaluation of Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG) systems is essential for assessing their effectiveness in retrieving relevant information, supporting multi-hop reasoning, and generating reliable answers. Existing studies employ different datasets, evaluation metrics, and experimental settings, making systematic comparison important for understanding the strengths and limitations of KG-RAG approaches.

6.1 Benchmark Datasets

Benchmark datasets provide standardized tasks for evaluating retrieval and question-answering performance. MultiHop-RAG, introduced by Tang and Yang (2024), focuses on questions requiring evidence from multiple interconnected documents. It is particularly relevant for assessing the limitations of conventional retrieval methods in multi-hop question answering.

Gutiérrez et al. (2024) evaluated HippoRAG on knowledge-intensive question-answering tasks, including multi-hop benchmarks. Zhu et al. (2025) examined KG²RAG using multi-hop question-answering datasets to assess knowledge graph-guided chunk expansion and organization. He et al. (2024) evaluated G-Retriever on textual graph question-answering tasks, demonstrating the relevance of graph-structured benchmarks for retrieval and logic.

These datasets support different evaluation objectives. While multi-hop benchmarks emphasize the integration of distributed evidence, textual graph datasets concentrate on obtaining and interpreting structured relationships. Consequently, dataset selection should reflect the reasoning complexity, document characteristics, and domain requirements of the target application.

6.2 Evaluation Metrics

KG-RAG systems are evaluated using metrics that measure retrieval effectiveness, answer quality, and computational efficiency. Common retrieval metrics include precision, recall, and ranking-based measures, which assess whether relevant evidence is successfully retrieved.

Answer quality is evaluated through measures such as exact match, F1-score, answer relevance, and faithfulness. Exact match and F1-score assess the correspondence between generated answers and reference answers, whereas faithfulness evaluates whether generated claims are supported by retrieved evidence. Saad-Falcon et al. (2024) introduced ARES, an automated RAG evaluation framework that considers context relevance, faithfulness, and relevance.

For graph-based approaches, additional evaluation considerations include multi-hop reasoning accuracy, evidence connectivity, and the relevance of retrieved graph paths. Computational metrics, such as retrieval latency, indexing time, and resource consumption, are also important for assessing the practical feasibility of KG-RAG systems, particularly in multi-hop question-answering systems that require the retrieval and integration of evidence from several different sources.

6.3 Comparison of Evaluation Settings

Existing KG-RAG approaches demonstrate different strengths depending on their architectures and evaluation settings. Edge et al. (2024) emphasized query-focused summarization using graph communities, while Gutiérrez et al. (2024) investigated associative graph-based retrieval for knowledge-intensive question answering. Zhu et al. (2025) focused on graph-guided evidence expansion and organization, whereas Guo et al. (2025) explored efficient graph-based indexing and retrieval.

Additionally, there has been a growing interest in comparing RAG and GraphRAG.. Han et al. (2026) examined differences between conventional and graph-based retrieval approaches, while Zhou et al. (2025) investigated how knowledge incompleteness affects KG-based RAG performance. These findings demonstrate how crucial it is to take into retrieval accuracy, reasoning complexity, graph quality, and computational requirements when evaluating different frameworks.

However, direct numerical comparisons across studies remain challenging because datasets, language models, retrieval configurations, and evaluation procedures may differ. Therefore, the reviewed approaches should be compared based on their reported evaluation settings rather than assuming that performance improvements observed in one benchmark will generalize across all domains.

7. RESEARCH CHALLENGES AND GAPS

Despite significant advancements in Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG), several challenges remain in achieving reliable, efficient, and scalable multi-hop question. Existing approaches have improved knowledge organization, retrieval, and reasoning; however, limitations related to graph quality, evidence reliability, retrieval integration, computational efficiency, and domain adaptation require further investigation.

7.1 Knowledge Graph Quality and Incompleteness

The effectiveness of KG-RAG systems depends heavily on the accuracy, completeness, and consistency of their knowledge graphs. Automatically constructed graphs may contain missing relationships, incorrectly identified entities, or incomplete domain information. Such limitations can affect evidence retrieval and logic, particularly when answering complex questions requiring multiple connected facts.

Zhou et al. (2025) investigated the performance of KG-based RAG methods under knowledge incompleteness, emphasizing the significance of graph quality of retrieval-augmented question responses. Although approaches such as GraphRAG and LightRAG support graph-based knowledge organization (Edge et al., 2024; Guo et al., 2025), maintaining reliable graph information remains an important practical challenge.

A key research opportunity is the development of KG-RAG systems that can identify missing or uncertain relationships and adapt their retrieval strategies accordingly.

7.2 Multi-Hop Logic and Evidence Reliability

Multi-hop question requires connecting information across multiple entities, relationships, or documents. Conventional retrieval methods may struggle to retrieve all necessary evidence, while graph-based approaches can encounter difficulties when relevant reasoning paths are incomplete or ambiguous.

Tang and Yang (2024) highlighted the challenges of multi-hop retrieval through the MultiHop-RAG benchmark. Sun et al. (2024) and Ma et al. (2025) explored knowledge graph-guided reasoning through Think-on-Graph and Think-on-Graph 2.0, respectively.

However, retrieving connected evidence does not automatically guarantee that the generated answer is logically correct or fully supported. Further research is needed on verifying reasoning paths, detecting unsupported conclusions, and ensuring that generated answers remain consistent with the evidence that recoverd.

7.3 Hybrid Retrieval and Evidence Fusion

Hybrid KG-RAG approaches combine semantic retrieval with graph-based retrieval to improve information coverage and contextual relevance. Zhu et al. (2025) proposed KG²RAG, which uses knowledge graph-guided chunk expansion and organization, while Guo et al. (2025) introduced LightRAG with graph-based indexing and retrieval mechanisms.

Despite these developments, selecting the most appropriate evidence from structured and unstructured sources remains challenging. Graph-based evidence may capture explicit relationships, whereas vector-based retrieval may identify relevant information that is not represented in the graph. Combining these sources can also introduce redundant, inconsistent, or irrelevant information.

An important research direction is confidence-aware adaptive evidence fusion, in which retrieval contributions are adjusted according to evidence relevance, reliability, and the complexity of the user's query. This represents a potential extension of existing hybrid retrieval approaches rather than an entirely unexplored research area.

7.4 Computational Cost and Scalability

KG-RAG systems may introduce additional computational requirements through entity extraction, relationship identification, graph construction, indexing, and graph traversal. These operations can increase preprocessing complexity and affect retrieval latency, especially when processing large or frequently updated document collections.

Edge et al. (2024) introduced graph-based community organization for query-focused summarization, while Guo et al. (2025) explored efficient graph-based retrieval through LightRAG. Although these approaches provide different strategies for organizing and accessing knowledge, computational efficiency remains an important consideration when selecting architectures for practical applications.

Future research should investigate incremental graph updates, selective graph exploration, efficient indexing, and adaptive retrieval mechanisms that balance answer quality with processing time and resource consumption.

7.5 Generalization Across Multi-Hop Question-Answering Tasks

The efficiency of KG-RAG systems may vary across multi-hop question-answering tasks because of differences in reasoning complexity, document structure, knowledge availability, and graph connectivity. A structure that performs effectively on questions requiring two reasoning steps may not achieve equivalent performance on more complex questions involving longer reasoning chains.

Incomplete relationships, ambiguous entities, and variations in evidence distribution can further influence retrieval and answer generation. Although existing approaches demonstrate Graph-based retrieval's potential and reasoning (Gutiérrez et al., 2024; He et al., 2024; Zhu et al., 2025), additional evaluation is needed to determine how consistently these methods perform across different multi-hop reasoning scenarios.

Future research should investigate adaptive retrieval strategies, reasoning-path verification, and evaluation benchmarks covering different levels of multi-hop complexity. These developments may improve the robustness and generalization of KG-RAG systems across diverse question-answering tasks.

8. DISCUSSION AND FUTURE RESEARCH DIRECTIONS

The reviewed literature demonstrates that Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG) provides a promising approach for improving retrieval of information and logic in multi-hop question-answering systems. By integrating structured knowledge with LLM, KG-RAG can support relationship-aware retrieval, evidence organization, and the integration of information across multiple reasoning steps. However, existing approaches differ considerably in their architectures, retrieval mechanisms, computational requirements, and evaluation methodologies.

GraphRAG empha sizes graph community organization and query-focused summarization (Edge et al., 2024), while HippoRAG employs graph-based associative retrieval for knowledge-intensive question answering (Gutiérrez et al., 2024). KG²RAG incorporates knowledge graph-guided chunk expansion and organization (Zhu et al., 2025), whereas LightRAG explores efficient graph-based indexing and retrieval (Guo et al., 2025). These approaches illustrate that the effectiveness of KG-RAG depends not only on integrating knowledge graphs but also on how relevant evidence is selected, organized, and incorporated into answer generation.

An important observation is that no single retrieval strategy can be assumed to perform optimally across all question types and domains. Graph-based retrieval can identify explicit relationships, while semantic retrieval can capture relevant information that might not be represented in the graph. Consequently, future research should investigate adaptive hybrid retrieval mechanisms that dynamically select or combine retrieval strategies according to query complexity, evidence availability, and knowledge graph completeness.

Another important direction is improving the reliability of multi-hop reasoning. Although Think-on-Graph and Think-on-Graph 2.0 demonstrate the potential of knowledge-guided reasoning (Sun et al., 2024; Ma et al., 2025), reliable answer generation requires more than retrieving connected entities. Future systems should incorporate evidence verification, reasoning-path validation, and mechanisms for identifying unsupported or conflicting information.

Knowledge graph incompleteness also presents a significant challenge. Zhou et al. (2025) examined KG-based RAG under incomplete knowledge conditions, stressing the importance of evaluating retrieval systems beyond ideal graph representations. Future approaches could incorporate uncertainty estimation, missing-relation detection, and fallback retrieval from textual sources when graph evidence is insufficient.

Furthermore, computational efficiency and scalability require greater attention. Graph construction, indexing, and retrieval may introduce additional processing overhead, particularly in large or continuously updated knowledge collections. Incremental graph maintenance, selective subgraph retrieval, and resource-aware retrieval strategies could improve the practicality of KG-RAG for real-world applications.

Finally, standardized evaluation remains essential for understanding the KG-RAG systems' generalizability. Current evaluation frameworks and benchmarks, such as MultiHop-RAG and ARES, offer helpful bases for evaluating multi-hop retrieval and response quality.(Tang & Yang, 2024; Saad-Falcon et al., 2024). Future evaluation should combine retrieval accuracy, answer faithfulness, reasoning reliability, computational efficiency, and robustness under incomplete knowledge conditions.

Overall, the findings suggest that future KG-RAG research should move beyond simply combining graph-based and semantic retrieval toward adaptive, uncertainty-aware, and evidence-grounded architectures. Such developments may improve the reliability, efficiency, and applicability of KG-RAG systems across specialized knowledge domains.

9. CONCLUSION

This review examined recent dev elopments in Knowledge Graph-Augmented Retrieval-Augmented Generation (KG-RAG) for multi-hop question, focusing on existing architectures, retrieval strategies, knowledge integration techniques, evaluation methods, and research challenges. The reviewed studies demonstrate that integrating knowledge graphs with RAG provides a promising approach for supporting relationship-aware information retrieval, structured evidence organization, and multi-hop logic.

Existing frameworks, including GraphRAG, HippoRAG, KG²RAG, LightRAG, and Think-on-Graph, demonstrate different approaches to combining structured knowledge with LLM. (Edge et al., 2024; Gutiérrez et al., 2024; Sun et al., 2024; Zhu et al., 2025; Guo et al., 2025). However, their effectiveness depends on factors such as knowledge graph quality, retrieval accuracy, reasoning complexity, and computational efficiency.

The review identifies several important challenges, including incomplete knowledge graphs, unreliable multi-hop evidence, difficulties in integrating structured and unstructured information, scalability limitations, and domain adaptation. Addressing these challenges requires further investigation into adaptive hybrid retrieval, confidence-aware evidence fusion, reasoning-path verification, and efficient graph maintenance.

Overall, KG-RAG represents a valuable direction for developing more dependable and context-aware question-answering systems. Future research should emphasize robust knowledge integration, standardized evaluation, and domain-specific adaptability to enhance the accuracy, efficiency, and trustworthiness of RAG.

REFERENCES

[1] Zhu, X., Xie, Y., Liu, Y., Li, Y., & Hu, W. (2025, April). Knowledge graph-guided retrieval augmented generation. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 8912-8924).

[2] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... & Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130.

[3] Gutiérrez, B. J., Shu, Y., Gu, Y., Yasunaga, M., & Su, Y. (2024). Hipporag: Neurobiologically inspired long-term memory for large language models. Advances in neural information processing systems, 37, 59532-59569.

[4] Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2025, November). LightRAG: Simple and Fast Retrieval-Augmented Generation. In EMNLP (Findings) (pp. 10746-10761).

[5] Shrivastava, R., & Gautam, C. S. (2026). Enhanced Heart Disease Prediction using Ensemble of Machine Learning Models. *International Journal of Computer Applications*, *187*(104), 40-46.

[6] He, X., Tian, Y., Sun, Y., Chawla, N. V., Laurent, T., LeCun, Y., ... & Hooi, B. (2024). G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. Advances in Neural Information Processing Systems, 37, 132876-132907.

[7] Linders, J., & Tomczak, J. M. (2025). Knowledge graph-extended retrieval augmented generation for question answering. Applied Intelligence, 55(17), 1102.

[8] Zhou, D., Zhu, Y., Wang, X., He, Y., Chen, J., Staab, S., & Kharlamov, E. (2025). Evaluating knowledge graph based retrieval augmented generation methods under knowledge incompleteness. arXiv preprint arXiv:2504.05163.

[9] Ma, S., Xu, C., Jiang, X., Li, M., Qu, H., Yang, C., ... & Guo, J. (2025, May). Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. In International Conference on Learning Representations (Vol. 2025, pp. 52782-52806).

[10] Sun, J., Xu, C., Tang, L., Wang, S., Lin, C., Gong, Y., ... & Guo, J. (2024, May). Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In International Conference on Learning Representations (Vol. 2024, pp. 3868-3898).

[11] Gutiérrez, B. J., Shu, Y., Qi, W., Zhou, S., & Su, Y. (2025). From rag to memory: Non-parametric continual learning for large language models. arXiv preprint arXiv:2502.14802.

[12] Li, S., He, Y., Guo, H., Bu, X., Bai, G., Liu, J., ... & Zheng, B. (2024, November). Graphreader: Building graph-based agent to enhance long-context abilities of large language models. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 12758-12786).

[13] Tang, Y., & Yang, Y. (2024). Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. arXiv preprint arXiv:2401.15391.

[14] Waoo, A. A., & Soni, B. K. (2021). Performance analysis of sigmoid and relu activation functions in deep neural network. In *Intelligent systems: proceedings of SCIS 2021* (pp. 39-52). Singapore: Springer Singapore.

[15] Tiwari, M., & Waoo, A. A. (2024). Transforming healthcare: the synergistic fusion of AI and IoT for intelligent, personalized well-being. In *Revolutionizing healthcare: AI integration with IoT for enhanced patient outcomes* (pp. 109-149). Cham: Springer Nature Switzerland.

[16] Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., ... & Tang, J. (2024). Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*.

[17] Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024, June). Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 338-354).

[18] Han, H., Ma, L., Wang, Y., Shomer, H., Guo, K., Lei, Y., ... & Tang, J. (2026, August). Rag vs. graphrag: A systematic evaluation and key insights. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2* (pp. 8966-8977).