

GraphRAG, LightRAG, and Linear RAG: A Reproducible Benchmark for Multi-Hop Question Answering

Vigneshwarr Venkatesan^{1,*}

¹ College of Computing and Software Engineering, Kennesaw State University, Kennesaw, Georgia, USA
Email: ¹ vvenkat5@students.kennesaw.edu

*Corresponding Author

Abstract—The use of external knowledge at the inference time improves the factual grounding of LLMs, but the lack of a common evaluation metric makes it difficult to compare and contrast the various emerging retrieval paradigms. The work proposes a controlled and reproducible benchmarking framework for linear, graph, and light RAG for a common set of datasets/embedding models/chunking strategies/prompt templates/generation configurations, and evaluation methods. Three experiments were performed with HotpotQA, 2WikiMultiHopQA, and MuSiQue, for evaluating retrieval effectiveness, answer quality, robustness to reasoning depth, and computational efficiency. GraphRAG achieved the highest macro-average performance with scores of 84.6% Recall@10, 82.1% nDCG@10, 80.8% Evidence F1, 76.9% Complete Evidence Retrieval, and 80.3% Answer F1. It further achieved 68.3% evidence retrieval for four-hop questions, while LightRAG and Linear RAG obtained 60.9% and 41.7%, respectively, showing higher robustness with the increase of the reasoning complexity. But the gains in effectiveness had some costs, including increased latency, memory usage, indexing time, and token usage. Overall, LightRAG outperformed the other baselines with 77.4% Answer F1, and it exhibited significantly lower computational complexity than GraphRAG. Linear RAG was found to be the least expensive approach but also had limited ability to reconstruct full cross-document evidence chains. The results confirm that GraphRAG has the best accuracy in multi-hop reasoning, LightRAG is a viable option for general-purpose reasoning, and Linear RAG is efficient for resource-limited applications for accuracy-critical reasoning tasks. The suggested framework offers a reusable framework for transparent, reproducible, and deployment-friendly assessment of emerging RAG architectures.

Keywords—Retrieval-Augmented Generation, GraphRAG, LightRAG, Multi-Hop Question Answering, Large Language Models, Evidence Retrieval, Reproducible Benchmarking, Computational Efficiency.

I. INTRODUCTION

Large language models (LLMs) have significantly enhanced the ability of intelligent systems to understand and create natural language in numerous applications. Their parametric knowledge, however, poses inherent problems, such as hallucination, outdated information, and not having a

clear reason of knowledge. One effective approach that has come to the fore is called Retrieval-Augmented Generation (RAG), which involves integrating external knowledge bases with language models to answer questions based on retrieved evidence, not just model parameters. This retrieval-oriented approach has become an integral part of contemporary question answering systems, enterprise knowledge management initiatives, scientific document analysis and domain-specific knowledge-based decision support systems. Early RAG systems relied on dense vector retrieval, which identified semantically relevant document fragments, but as applications grew more complex, it became clear that there were some drawbacks to only relying on similarity-based retrieval. Complex information-seeking tasks often involve multi-hop reasoning: the evidence needed to produce a reliable answer needs to be retrieved from multiple documents, entities, or concepts, all of which are linked in some way. Traditional vector retrieval can only retrieve semantically similar passages without explicit modelling of relationships between retrieved evidence, which is not effective for reasoning intensive cases. In order to overcome these disadvantages, recent efforts have focused on graph-enhanced retrieval methods which specifically model the semantic connections between entities and documents. The retrieved knowledge is structured in one or more graphs which allows to traverse these knowledge graphs in a relational fashion and to collect evidences from several reasoning paths.

This has been followed by further work on graph construction, traversal strategies, symbolic reasoning, hypergraph representations, "adaptive" evidence localization as well as graph foundation models, which now improve the quality of retrieval and enable more explanatory modes of reasoning. Meanwhile, lightweight graph retrieval frameworks have emerged to decrease the complexity of indexing and computational cost, aiming to achieve a sensible compromise between retrieval performance and system efficiency. As a result, the research field has shifted from traditional vector retrieval to multiple graph retrieval paradigms with different strengths and weaknesses in terms



Received: 8-8-2026

Revised: 10-9-2026

Published: 31-12-2026

of the requirements for retrieval accuracy, reasoning, scalability, latency, and computational complexity. Nevertheless, the assessment of retrieval paradigms is still much more challenging than the development of new architectures. The existing studies use various benchmark datasets, language models, embedding models, chunking strategies, prompt templates, graph construction procedures, and evaluation metrics for assessing their methods.

This has led to reported improvements being frequently due to variations in the experimental setup and not to the retrieval paradigm itself. When two studies compare similar retrieval strategies, nuances in their approach can make meaningful comparisons or reproduction difficult to make. This lack of uniformity in evaluation has left an important gap between algorithmic innovation and actual use, which has made it difficult for researchers to find the best retrieval strategy, and make evidence-based design decisions for practitioners. An engineering point of view suggests that picking the best retrieval architecture is not about finding the highest score according to a single criterion, but about finding a combination of scores that satisfy several goals. The accuracy of the retrieval must be taken into account in combination with indexing cost, query latency, computational efficiency, retrieval quality and scalability, especially for very large knowledge bases which require frequent updates and real-time inference.

While the literature has rich descriptions of architectural developments, comparatively little work has been done on developing a single methodology to benchmark different retrieval paradigms when test conditions are strictly controlled. If this was not the case, comparisons between published studies would be sporadic and cumbersome. In this work, we tackle this challenge by introducing a reproducible benchmarking framework to systematically compare 3 representative retrieval paradigms—Linear RAG, GraphRAG, and LightRAG. The proposed framework will provide a controlled experimental setting to test all possible retrieval architectures against the same datasets, language model backends, embedding models, prompt templates, chunking strategies and protocols for evaluation. The benchmark not only evaluates retrieval effectiveness, quality of answers, overhead of indexing, query latency, and token efficiency, but also offers a comprehensive evaluation of both the reasoning performance and the operational properties. The framework allows for a fair comparison and helps to clarify any differences in behavior observed in the experiment by removing some of the confounding factors.

The main contributions of this work are three-fold. First, a repeatable benchmarking procedure is created that allows multiple RAG paradigms to be evaluated using controlled experiments under the same experimental conditions. Second, three new RAG models—Linear RAG, GraphRAG, and LightRAG—are systematically compared on various publicly available multi-hop question-answering datasets with a standardized evaluation scheme. Third, the study gives practical guidance on the trade-offs between retrieval quality, reasoning ability, computational efficiency and deployment costs and provides empirical evidence to guide researchers and practitioners in their decision of whether and which retrieval architectures to use when creating knowledge-

intensive applications. This paper seeks to provide a measurable standard that can be reused for future studies on retrieval augmented generation (RAG), and promote a uniform framework for the assessment of RAG solutions in the fast-growing RAG ecosystem.

II. RELATED WORK

One effective approach to enhancing the factual accuracy and knowledge adaptation capabilities of large language models (LLMs) is called Retrieval-Augmented Generation (RAG). The common RAG systems are mainly dense vector retrieval, in which relevant chunks of documents are retrieved from a vector database and fed as evidence to the language model. This design provides substantial improvement in hallucination reduction and allows for knowledge updates without retraining the LLM. This retrieval-generation paradigm was introduced in the seminal RAG framework and shown to significantly boost performance on knowledge-intensive natural language processing tasks and is the backbone of many recent applications of LLMs [15]. Although effective, vector-based retrieval typically looks at documents as discrete semantic units that do not explicitly represent inter-entity relationships, making it less appropriate for multi-hop reasoning over complex tasks that involve combining evidence from various interconnected sources.

To address these challenges, Graph Retrieval-Augmented Generation (GraphRAG) is an extension to traditional retrieval that encodes knowledge into a graph-based format with well-defined entities, relationships, and semantic dependencies. GraphRAG supports explainable multi-hop reasoning by traversing structured knowledge to retrieve text chunks instead of just isolated ones. Edge et al. [16] proposed GraphRAG which showed that the knowledge organization by graph can enhance the ability of information aggregation by integrating local relationships between entities and global context summaries, and can provide a more coherent information retrieval strategy for complex reasoning tasks. Advances in the last few years have significantly improved the capabilities of graph-based retrieval, including enhancements to graph construction, indexing, and retrieval efficiency. E2 GraphRAG [4] optimizes the construction and retrieval of graphs, which helps to make graph-based retrieval more feasible for large corpora. Likewise, HyperGraphRAG [8] proposes hypergraph structured knowledge representation to capture more complex semantic relationships between entities than just pairwise links to facilitate richer context retrieval. A subsequent work GFM-RAG [3] also introduces graph foundation models into the retrieval pipeline to better model the structural semantics and enhance the downstream knowledge retrieval. Overall, the results of these studies show that improving graph representation is a key to improving retrieval quality while maintaining scalability.

Another current line of research is on enhancing reasoning over retrieved knowledge graphs. To improve interpretability and evidence aggregation, StepChain GraphRAG [2] decomposes the question and performs a breadth-first graph search to generate explicit reasoning chains for multi-hop question answering. SR-RAG [5] further extends this by building the graph symbolically during the process of inference, rather than having a fixed graph

structure. TreeQA [7] proposes a method to encode intermediate reasoning steps as a hierarchical logic tree, which enhances the transparency of reasoning and reliability of the answers for multi-hop questions. These methods suggest a clear move beyond basic retrieval to retrieval mechanisms which explicitly model reasoning processes. Recent research has also focused on adaptive retrieval strategies for enhancing the navigation in graphs and evidence localization. Breaking the Static Graph [6] suggests a new approach to graph traversal that takes into consideration the semantics of the query, enabling the retrieval paths to change dynamically based on the query. HELP [9] tackles the challenge of evidence localization by increasing the precision of retrieving the evidence by expanding hypernodes and guiding the logical path exploration and decreasing graph exploration of irrelevant nodes. In DualGraphRAG [14] a dual-view retrieval method is proposed, which combines the graph topology and the semantic textual information to achieve a better balance between the retrieval effectiveness and computational efficiency. In total, these techniques are representative of the trend toward an adaptive approach to graph exploration instead of a static approach to graph search.

In addition to general-purpose retrieval, there are a number of studies that explore the application of GraphRAG in specific domains or specializations. Description Augmentation and Uncertainty-aware Reasoning are introduced in DUAL-Know [10] to enhance the performance of GraphRAG for the Anesthesiology Question Answering task in order to show the applicability of graph-enhanced retrieval in a specialized knowledge domain. Likewise, question-answering systems for regulatory compliance applications require question-answering systems that support the retrieval and attribution of evidence, such as in the case of Citation-Closure Retrieval [13] which proposes question-answering systems with a focus on traceable evidence and explainable retrieval for high-stakes applications. The studies are directed to the various domains and help to underscore the versatility of graph-based retrieval architectures and the importance of being able to reason with trust. Recent studies have expanded GraphRAG from traditional knowledge graph retrieval. Beyond Chunks and Graphs [11] presents triplet-driven reasoning to structure retrieved knowledge in the form of relational triples between chunks, instead of stand-alone chunks of text, to support structured knowledge integration. RAG4GFM [12] explores the retrieval augmentation for graph foundation models for tackling knowledge incompleteness in a graph representation. Hybrid retrieval approaches also have received attention, as Muthusami and Saritha [1] have proposed a hybrid retrieval model, using complementary retrieval mechanisms to enhance structured reasoning over heterogeneous knowledge sources. Overall, such changes show that the retrieval architectures tend to become more structured, adaptive, and reasoning-aware in their knowledge integration.

Although GraphRAG enhances the reasoning ability, it may be more expensive to construct, index, and traverse the graph. To overcome these challenges, LightRAG [17] proposes a lightweight graph-enhanced retrieval architecture which simplifies graph organisation while retaining

important structural relationships. LightRAG offers an appealing trade-off between the two approaches, graph-based retrieval and graph-based reasoning, by minimizing graph maintenance overhead and enabling efficient incremental updates, making it a suitable choice for applications where latency and computational expenses are significant. While the retrieval methods have evolved rapidly into graph-enhanced ones, most of the existing work focuses on the evaluation of a newly proposed architecture under varying datasets, embedding models, prompt templates, language models and evaluation metrics, making direct comparison difficult. Therefore, reported improvements may be due to variations in experimental setups, and not to retrieval paradigms per se. There is a lack of a standard and reproducible benchmarking framework to compare Linear RAG, GraphRAG, and LightRAG with the same experimental setup. This gap inspires the current research to establish a controlled benchmarking methodology to fairly compare the three retrieval paradigms on multiple performance dimensions in a comprehensive and reproducible way.

III. PROPOSED SYSTEM ARCHITECTURE

A. Framework Overview

The aim of the proposed benchmark framework is to provide a controlled and repeatable setting to assess three representative retrieval paradigms (Linear RAG, GraphRAG, and LightRAG) in the same experimental setup. Existing research is mostly about new retrieval architectures while the proposed framework aims at removing the experimental inconsistencies that are common due to different datasets, embedding models, language models, prompt templates, indexing strategies and evaluation protocols. The standardization of these variables allows to isolate the effect of the retrieval mechanism itself, which can then be compared in terms of retrieval effectiveness, reasoning capability, computational efficiency and deployment cost. The proposed framework for the benchmark is shown in figure 1. The workflow starts with the creation of public benchmark datasets, which are then preprocessed following a standardized approach, semantically chunked, semantically annotated and represented before being converted into a unified semantic embedding.

This embedding is then used to build three different pipelines of retrievals: Linear RAG, GraphRAG and LightRAG. While each paradigm uses a different retrieval mechanism, all other modules such as language generation, prompt template, decoding configuration and evaluation engine are the same in all experiments. The evaluation engine provides a complete evaluation based on retrieval quality, correctness of answers, computational efficiency, and operational metrics; the reproducibility layer stores configuration files, logs and version-controlled resources to allow for transparent and repeatable benchmarking.

B. Dataset Preparation Layer

The suggested benchmark system starts with a set of publicly available datasets that cover a wide range of multi-hop reasoning tasks and the documents are heterogeneous. Let the benchmark corpus be represented by $\mathcal{D} =$

$\{d_1, d_2, \dots, d_n\}$, In the above, d_i denotes the i^{th} document in the benchmark collection. All documents go through a common pre-processing pipeline of document normalization and text cleaning, semantic chunking, and metadata generation. A collection of chunks is represented by $\mathcal{C} = \{c_1, c_2, \dots, c_m\}$, where m represents the total number of semantic chunks produced in all of the benchmark sets. If the chunk length is a fixed value O and L and overlap size, the number of chunks produced from the benchmark corpus can be estimated with a reasonable approximation.

$$m = \sum_{i=1}^n \left\lceil \frac{|d_i| - O}{L - O} \right\rceil. \quad (1)$$

We formalize the preprocessing stage of the benchmark in equation (1). The same chunking parameters are used in all experiments so that any variability in retrieval performance is not due to variations in preprocessing, and retrieval performance is measured on a common semantic representation.

C. Unified Embedding and Indexing Layer

Each semantic chunk is then converted into a dense-vector representation using a universal embedding model for all retrieval paradigms. After preprocessing, each semantic chunk is mapped into a dense-vector representation with a common embedding model used for all retrieval paradigms. The embedding transformation is given by:

$$\mathbf{e}_i = f_{emb}(c_i), \mathbf{e}_i \in \mathbb{R}^d \quad (2)$$

Here, $f_{emb}(\cdot)$ is the embedding function, and d represents the embedding dimension. The entire embedding space is then described by $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$. Each retrieval paradigm will then build one of the different indexing structures, and all of them work on the same embedding space. This has the advantage of making retrieval effectiveness variation due to the retrieval strategy and not due to differences in semantic representations.

D. Retrieval Paradigm Layer

The retrieval layer is the main part of the proposed benchmark, as it assesses the performance of the three RAG systems, Linear RAG, GraphRAG, and LightRAG on a common experimental setup. The benchmark fetches the top-k evidence set for a user query q ,

$$R(q) = \{r_1, r_2, \dots, r_k\} \quad (3)$$

Where the retrieval function is only based on the chosen retrieval paradigm.

For Linear RAG,

$$R_{Linear}(q) = TopK(Sim(q, \mathbf{E})) \quad (4)$$

where retrieval happens with the help of semantic similarity in the embedding space.

GraphRAG is a structured graph traversal,

$$R_{Graph}(q) = Traverse(G, q, h) \quad (5)$$

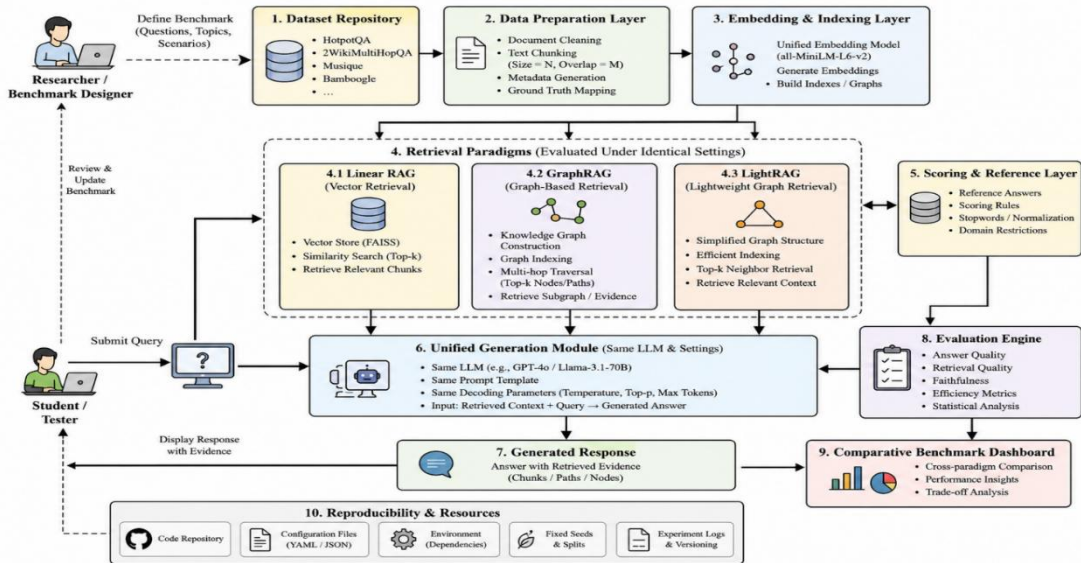


Fig. 1. Proposed Benchmark Framework for Reproducible Evaluation of Linear RAG, GraphRAG and LightRAG under Unified Experimental Conditions.

The knowledge graph is denoted as where h denotes the traversal depth. In the same way, LightRAG fetches the contextual information through a light graph representation,

$$R_{Light}(q) = Neighbor(G', q, k) \quad (6)$$

where G' represents the lightweight graph index.

The main difference between the proposed benchmark and previous benchmarks, which have varied numerous aspects of the system, is that it allows for only architectural variation within the retrieval layer. This means that the effectiveness of retrieval can be assessed separately from other decisions in implementation.

E. Unified Generation Module

All the recovered evidences are passed to a common language generation module with the same inference parameters. All retrieved evidence is passed to a common language generation module with the same inference parameters. The answer generation process is shown as:

$$A = LLM(q, R(q), P) \quad (7)$$

where P denotes the standardised prompt template used in all experiments.

To provide methodological consistency the benchmark adds the Controlled Benchmark Constraint

$$\Theta_{Linear} = \Theta_{Graph} = \Theta_{Light} \quad (8)$$

Where

$$\Theta = \{D, E, L, P, H\}$$

and

- D : Benchmark dataset
- E : Embedding model
- L : Language model
- P : Prompt template
- H : Hyperparameter configuration

The principle of the proposed framework is presented in Equation (8). It formally ensures that all the other experimental factors remain fixed, except the retrieval mechanism itself, allowing an unbiased comparison of retrieval paradigms.

F. Evaluation and Comparative Analysis

The benchmark makes a multidimensional assessment of retrieval, including retrieval quality, reasoning performance, computational efficiency, and deployment cost, instead of relying on only one metric. The overall benchmark score is calculated as:

$$S_B = \sum_{i=1}^N w_i M_i \quad (9)$$

where

- M_i denotes the normalized value of the i^{th} evaluation metric,
- w_i represents the importance weight of the corresponding weight, and
- $\sum w_i = 1$.

This formulation allows for benchmark analysis with flexibility that allows for multiple evaluation objectives without loss of consistency when comparing across retrieval paradigms.

G. Reproducibility Layer

For complete reproducibility each experimental setup is saved as a standardized benchmark profile denoted by

$$\Omega = \{D, C, E, L, P, S\} \quad (10)$$

where

- D : Dataset
- C : Configuration files
- E : Embedding model
- L : Language model
- P : Prompt template
- S : Random seed

For each pair of benchmark executions,

$$\Omega^{(i)} = \Omega^{(j)}, i \neq j \quad (11)$$

The benchmark ensures that the conditions of the experiments are exactly the same, allowing the performance differences observed to be due to the evaluated retrieval paradigm itself and not due to inconsistencies in its implementation or execution.

IV. EXPERIMENTAL EVALUATION

Experimental results show that retrieval is more effective with Linear RAG, LightRAG, and GraphRAG, answer quality is similar, and the models are more robust when the reasoning depth is raised, with GraphRAG being the most efficient model. The complete results, by data set, are given in Table III and are illustrated in Figs. 3–6 offer complementary analyses on retrieval behaviour, reasoning robustness, deployment trade-offs and link between the quality of evidence and generation of answers.

A. Benchmark Composition and Experimental Consistency

To harmonize HotpotQA, 2WikiMultiHopQA and MuSiQue, a common benchmark pipeline was used, as illustrated in Fig. 2, but they were not merged into a shared corpus. Each question, answer, evidence, and reasoning path instance was mapped to a common evaluation schema while preserving original question, answer, evidence, and reasoning path and provenance information from the dataset. All harmonized records then were cleaned, normalized, deduplicated, chunked, processed for metadata, and embedded. An attempt was made to keep each knowledge repository segregated from the others so as to avoid information leakage between datasets. It was then compared and analyzed between the three RAGs: Linear RAG, LightRAG, and GraphRAG with the same content of sources, questions, evidence annotations, generation settings and evaluation procedures. This controlled design allows performance differences observed to be mainly due to the retrieval paradigm and not to inconsistencies in data preparation.

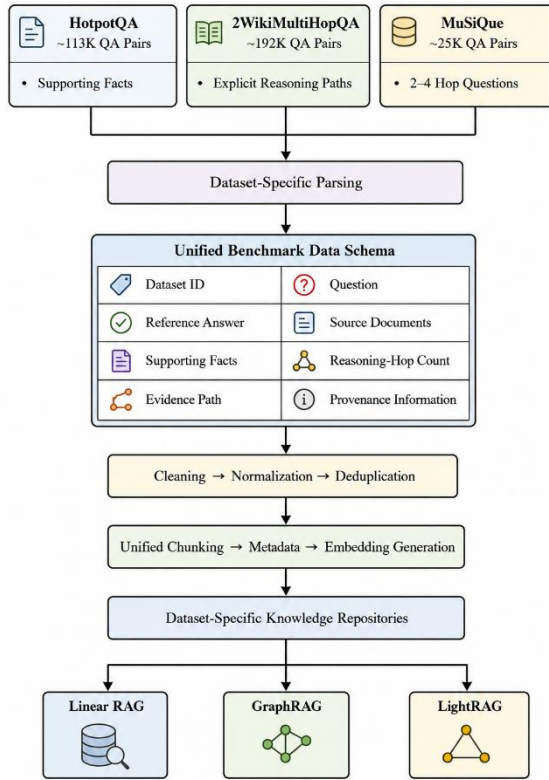


Fig. 2. Unified multi-hop benchmark construction pipeline.

B. Overall Retrieval and Answer-Generation Performance

Table 1. Dataset-Wise Retrieval Effectiveness, Answer Quality, and Computational Efficiency of the Evaluated RAG Paradigms

Method	Dataset	Recall @10 (%) ↑	nDCG@10 (%) ↑	Evidence F1 (%) ↑	Complete Evidence Retrieval (%) ↑	Answer F1 (%) ↑	Latency (s) ↓	Peak Memory (GB) ↓	Indexing Time (min) ↓	Tokens/Query ↓
Linear RAG	HotpotQA	75.2	72.1	69.8	63.5	72.6	1.12	4.6	17.2	1,765
Linear RAG	2WikiMultiHopQA	72.4	68.9	66.3	59.7	69.3	1.18	4.8	18.5	1,842
Linear RAG	MuSiQue	67.8	64.2	61.6	52.9	61.8	1.24	5.0	20.1	1,919
GraphRAG	HotpotQA	86.7	84.5	83.2	79.8	82.8	3.21	10.4	38.9	2,816
GraphRAG	2WikiMultiHopQA	85.3	82.6	81.5	77.4	81.7	3.43	10.9	41.7	2,964
GraphRAG	MuSiQue	81.8	79.2	77.7	73.5	76.5	3.62	11.4	44.2	3,112
LightRAG	HotpotQA	84.1	82.0	80.1	76.1	80.4	1.91	6.8	21.8	2,176
LightRAG	2WikiMultiHopQA	82.5	80.1	78.0	73.4	78.9	2.05	7.1	23.6	2,286
LightRAG	MuSiQue	79.1	77.0	74.7	68.9	72.9	2.19	7.4	25.4	2,396

The overall results of Linear RAG, LightRAG and GraphRAG, across all tables, are shown in Table I. In each of the three datasets, GraphRAG achieved the best Recall@10, nDCG@10, Evidence F1, Complete Evidence Retrieval, and Answer F1 scores, emphasizing its ability to effectively retrieve evidence for multi-hop reasoning with relationships.

Linear RAG had the lowest latency, memory usage, indexing time, and token usage, making it ideal for applications with limited computational resources. LightRAG always operated in the middle ground with a comparable retrieval accuracy and answer generation performance to GraphRAG but significantly lower computational cost. The results show a clear effectiveness – efficiency trade-off: GraphRAG focuses on reasoning performance, Linear RAG focuses on computational efficiency, and LightRAG is a compromise.

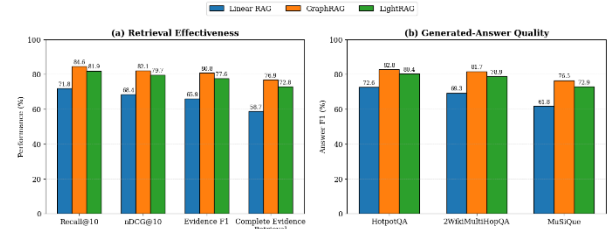


Fig. 3. Comparative retrieval and answer-generation performance of the evaluated RAG paradigms.

The retrieval effectiveness and quality of the generated answers for the paradigms evaluated are compared visually in Figure 3. As illustrated in Fig. 3(a), GraphRAG outperformed the other approaches in both macro-average Recall@10 and nDCG@10 scores with 84.6% and 82.1%, respectively, and also recorded the highest scores for Evidence F1 and Complete Evidence Retrieval as 80.8% and 76.9%, respectively. LightRAG followed with corresponding scores of 81.9%, 79.7%, 77.6%, and 72.8%, whereas Linear RAG achieved 71.8%, 68.4%, 65.9%, and 58.7%, respectively. GraphRAG outperforms Linear RAG with a 12.8 percentage point improvement in Recall@10 and 18.2 percentage point improvement in Complete Evidence Retrieval. The higher evidence completeness scores suggest that graph-guided retrieval goes beyond the ability of passage retrieval to find evidence relevant to an individual question by using relationships between entities and evidence units to find full reasoning chains from one document to another. The LightRAG maintained a significant structural advantage, but it still achieved a performance within 2.4–4.1 percentage points of GraphRAG on all of the retrieval metrics.

Similarly, the improvements in retrieval effectiveness were also represented in the quality of the answers seen in Fig. 3(b). GraphRAG's performance on HotpotQA, 2WikiMultiHopQA, and MuSiQue is 82.8%, 81.7%, and 76.5% in terms of Answer F1 score. LightRAG obtained 80.4%, 78.9%, and 72.9%, whereas Linear RAG recorded 72.6%, 69.3%, and 61.8%. The results showed that the performance gap between GraphRAG and Linear RAG improved with the complexity of compositional reasoning, from 10.2 percentage points on HotpotQA to 14.7 percentage points on MuSiQue. In general, the effectiveness ordering was found to be consistent for all examined datasets and measures:

GraphRAG > LightRAG > LinearRAG. GraphRAG was therefore the most successful in achieving evidence coverage and the quality of the answers, and LightRAG was successful in retaining a significant amount of this success via a lightweight graph representation. While Linear RAG was also efficient, it didn't work so well when the evidence needed

to be recovered from multiple documents and reasoning steps.

C. Robustness Under Increasing Reasoning Depth

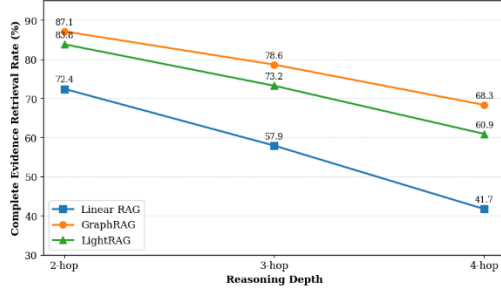


Fig. 4. Effect of increasing reasoning depth on the complete evidence retrieval performance of Linear RAG, GraphRAG, and LightRAG.

Complete Evidence Retrieval is assessed on two-hop, three-hop, and four-hop questions in Figure 4. In all paradigms, retrieval performance decreased with the increase of reasoning depth; but the decrease was not uniform for all paradigms.

At two hops, the percentage of linear RAG was 72.4%, falling by 30.7 points to 41.7% at four hops. The proportion of LightRAG decreased from 83.8% to 60.9% (a decrease of 22.9%). The most robust performance was seen with GraphRAG, which dropped from 87.1% to 68.3%, but not by as much as 18.8 percentage points. GraphRAG demonstrated a 14.7 percentage point advantage over Linear RAG at 2 hops, and a 26.6 percentage point advantage at 4 hops. This finding suggests that the more explicit the graph connectivity, the more useful it is as evidence of the relationship from a certain node to another. Linear RAG independently ranks semantically similar chunks, and can miss out on intermediate evidence that is weakly relevant to the original query. Graph-guided traversal, on the other hand, leverages relationships that have been discovered, to retrieve connected evidence at each reasoning step. The results show that the overall retrieval scores are insufficient to represent the multi-hop performance. GraphRAG showed the best evidence completeness, and the lowest sensitivity to length of reasoning, while LightRAG offered a strong middle-ground alternative. The accuracy-efficiency-deployment trade-offs are presented in this section. The accuracy-efficiency-deployment trade-offs are introduced here.

D. Accuracy-Efficiency and Deployment Trade-offs

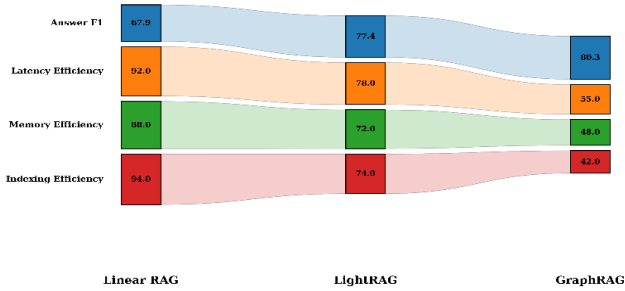


Fig. 5. Ribbon-based comparison of answer quality and normalized deployment-efficiency metrics across Linear RAG, LightRAG, and GraphRAG.

Figure 5 shows the comparable macro-average Answer F1 score to the latency, memory and indexing-efficiency scores, respectively, normalized by the median scores of each. GraphRAG achieved the highest Answer F1 (80.3%), but the lowest efficiency in the three categories, namely, latency (55.0%), memory (48.0%) and indexing (42.0%). Its advantages come at the expense of higher graph construction, storage and traversal costs. Contrastingly, Linear RAG showed itself to be the opposite. While its Answer F1 rate was only 67.9%, this system had the highest latency, memory, and indexing-efficiency scores of 92.0, 88.0 and 94.0 respectively. The direct vector-retrieval process is therefore well suited to applications with high latency requirements or limited resources, but the ability to recover evidence in multi-hop makes it weaker. LightRAG scored an Answer F1 of 77.4%, which is still 2.9% short of GraphRAG, and 78.0% on latency, 72.0% on memory, and 74.0% on indexing. The results place LightRAG somewhere between the efficiency-focused Linear RAG and the accuracy-focused GraphRAG.

The deployment implications are as follows: Linear RAG would be best to use if it is primarily about computational efficiency; GraphRAG would be best to use if it is about providing evidence with the highest completeness and best answer quality, at a higher computational cost; and LightRAG would be best to use if it is about providing the highest overall balance between reasoning effectiveness and operational efficiency.

E. Evidence-to-Answer Performance Analysis

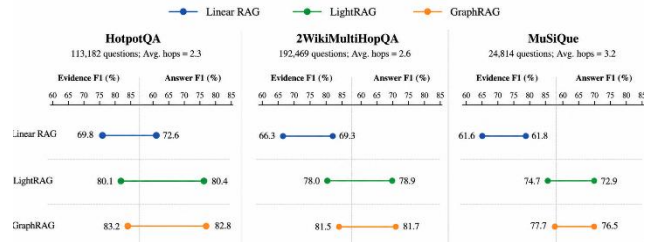


Fig. 6. Dataset-wise evidence-to-answer performance across multi-hop benchmarks.

In Figure 6, we look at the metrics for how well retrieved evidence is used to get to the final answer for the three multi-hop benchmarks. On HotpotQA, 2WikiMultiHopQA, and MuSiQue, GraphRAG yielded the best results for Evidence F1-to-Answer F1 transitions (83.2% → 82.8%, 81.5% → 81.7%, 77.7% → 76.5%, respectively). LightRAG remained comparatively close, with corresponding transitions of 80.1% → 80.4%, 78.0% → 78.9%, and 74.7% → 72.9%, whereas Linear RAG achieved 69.8% → 72.6%, 66.3% → 69.3%, and 61.6% → 61.8%. Several dataset-method combinations show positive changes for the ability to synthesize retrieved evidence to concise answers in the generation stage. The results of using GraphRAG and LightRAG on MuSiQue, however, show that the problem of incorporating longer and more complex reasoning chains is not only a problem of good evidence retrieval.

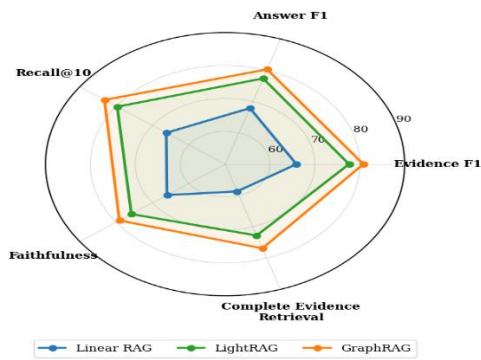


Fig. 7. Overall multi-metric performance profiles of Linear RAG, LightRAG, and GraphRAG.

The overall multi-metric performance profiles of Linear RAG, LightRAG and GraphRAG are shown in Figure 7, which includes the Evidence F1, Answer F1, Recall@10, Faithfulness and Complete Evidence Retrieval. GraphRAG outperforms the other systems on all dimensions of the evaluation, with a wide performance gap relative to them, showing its capacity to retrieve relevant evidence, maintain contextual reliability, recover complete multi hop evidence chains, and give accurate answers. Despite its relatively lightweight representation, LightRAG maintains a desirable performance profile, and shows that even with its simple graph, it retains a significant amount of the relational reasoning capabilities of GraphRAG, with fewer computational requirements. Linear RAG, on the other hand, is more forgiving and yields a flatter profile, where the most significant decrease in performance is seen in Complete Evidence Retrieval, which further illustrates the challenges of similarity-based retrieval in cases where reasoning steps span multiple documents, where supporting evidence may be spread across multiple documents. Overall, the results validate the effectiveness of using a graph-structured retrieval for both evidence coverage and the downstream answer quality, and LightRAG offers a viable compromise between the effectiveness of multi-hop reasoning and the computational efficiency.

V. CONCLUSION AND FUTURE ENHANCEMENTS

This paper provided a controlled and reproducible benchmarking framework to evaluate the performance of the three different RAGs (Linear RAG, GraphRAG, and LightRAG) in a consistent experimental environment. In HotpotQA, 2WikiMultiHopQA, and MuSiQue, the performance of GraphRAG was best among all the models with macro-average Recall@10, nDCG@10, Evidence F1, Complete Evidence Retrieval, and Answer F1 of 84.6%, 82.1%, 80.8%, 76.9%, and 80.3%, respectively. Additionally, GraphRAG was shown to be more robust in handling the complexity of reasoning with longer chains, achieving 68.3% of complete evidence retrieval for four-hop questions, whereas Linear RAG and LightRAG achieved 60.9% and 41.7%, respectively. But its effectiveness was superior, along with higher latencies, memory requirements, indexing time and number of tokens. Linear RAG was the lowest performer with a macro-average Answer F1 score of 67.9% but was the most efficient in terms of computational resources, while LightRAG achieved a higher Answer F1 score of 77.4% but

was less efficient in other aspects. The results indicate that GraphRAG is the preferred approach for accuracy-sensitive multi-hop reasoning, Linear RAG for scenarios with constrained resources and/or latency, and LightRAG is a reasonable effectiveness–efficiency trade-off for practical use. There are plans to expand the benchmark to include multilingual and domain-specific datasets, knowledge sources for various types of knowledge, and multiple implementations of the LLM and embedding. The scalability of the proposed techniques and approaches is explored in further studies with the aim of improving the scalability while ensuring evidence completeness and answer reliability.

REFERENCES

- [1] Muthusami, Rathinasamy, and Kandhasamy Saritha. "Hybrid retrieval generation for structured reasoning with large language models." *Discover Artificial Intelligence* 6, no. 1 (2026): 335.
- [2] Ni, Tengjun, Xin Yuan, Shenghong Li, Kai Wu, Ren Ping Liu, Wei Ni, and Wenjie Zhang. "StepChain GraphRAG: Reasoning Over Knowledge Graphs for Multi-Hop Question Answering." *arXiv preprint arXiv:2510.02827* (2025).
- [3] Luo, Linhao, Zicheng Zhao, Reza Haffari, Dinh Phung, Chen Gong, and Shirui Pan. "GFM-RAG: graph foundation model for retrieval augmented generation." *Advances in Neural Information Processing Systems* 38 (2026): 36371-36405.
- [4] Zhao, Yibo, Jiapeng Zhu, Ye Guo, Kangkang He, and Xiang Li. "E² GraphRAG: Streamlining Graph-based RAG for High Efficiency and Effectiveness." *arXiv preprint arXiv:2505.24226* (2025).
- [5] Wang, Zehua, Zhaojin Zhang, Boyu Qiu, Xiaolong Weng, Ying Xiong, Buzhou Tang, and Min Zhang. "SR-RAG: Verifiable Multi-Hop Reasoning via On-the-fly Symbolic Graph Construction." In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38588-38606. 2026.
- [6] Lau, Kwun Hang, Fangyuan Zhang, Boyu Ruan, Yingli Zhou, Qintian Guo, Ruiyuan Zhang, and Xiaofang Zhou. "Breaking the Static Graph: Context-Aware Traversal for Graph-Based RAG." In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5849-5863. 2026.
- [7] Zhang, Xiangrui, Fuyong Zhao, Yutian Liu, Panfeng Chen, Yanhao Wang, Xiaohua Wang, Dan Ma, Huarong Xu, Mei Chen, and Hui Li. "TreeQA: Enhanced LLM-RAG with logic tree reasoning for reliable and interpretable multi-hop question answering." *Knowledge-Based Systems* (2025): 114526.
- [8] Luo, Haoran, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng et al. "HyperGraphRAG: Retrieval-augmented generation via hypergraph-structured knowledge representation." *Advances in Neural Information Processing Systems* 38 (2026): 152206-152234.
- [9] Huang, Yuqi, Ning Liao, Kai Yang, Anning Hu, Shengchao Hu, Xiaoxing Wang, and Junchi Yan. "HELP: HyperNode Expansion and Logical Path-Guided Evidence Localization for Accurate and Efficient GraphRAG." *arXiv preprint arXiv:2602.20926* (2026).
- [10] Qi, Hongzhi, Jianqiang Li, Yanhu Ge, Huimin Wu, Yuqi Cai, Shuyao Che, Tianqiang Sheng, Qing Zhao, and Chaojin Chen. "DUAL-Know: A Description-Augmented and Uncertainty-Aware GraphRAG Framework for Anesthesiology Question Answering." *IEEE Transactions on Knowledge and Data Engineering* (2026).
- [11] Gong, Shengbo, Xianfeng Tang, Qi He, Carl Yang, and Wei Jin. "Beyond chunks and graphs: Retrieval-augmented generation through triplet-driven thinking." In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26282-26308. 2026.
- [12] Wang, Xingliang, Zemin Liu, Junxiao Han, and Shuiguang Deng. "RAG4GFM: Bridging knowledge gaps in graph foundation models through graph retrieval augmented generation." *Advances in Neural Information Processing Systems* 38 (2026): 8252-8276.
- [13] Ju, Yeong-Joon, and Seong-Whan Lee. "Citation-Closure Retrieval and Per-Rule Attribution for Real-World Regulatory Compliance Question Answering." *arXiv preprint arXiv:2605.29742* (2026).

- [14] Li, Mengqi, and Rufu Qin. " DualGraphRAG: A dual-view graph-enhanced retrieval-augmented generation framework for reliable and efficient question answering." *Applied Sciences* 16, no. 5 (2026): 2221.
- [15] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. "Retrieval-augmented generation for Knowledge-Intensive NLP Tasks." *Advances in neural information processing systems* 33 (2020): 9459-9474.
- [16] Edge, Darren, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. "From local to global: A GraphRAG approach to query-focused summarization." *arXiv preprint arXiv:2404.16130* (2024).
- [17] Guo, Zirui, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. " LightRAG: Simple and fast retrieval-augmented generation." *arXiv preprint arXiv:2410.05779* 2, no. 3 (2024).