Iterative Multi-Granular RAG with Contextual Hierarchical Graph

Yanli Hu[1*], Teng Liu[1*], Zhuangyi Zhou[1], Weixin Zeng[1†], Zhen Tan[1], Xiang Zhao[2]

[1]National Key Laboratory of Information Systems Engineering, National University of Defense Technology, China
[2]National Key Laboratory of Big Data and Decision, National University of Defense Technology, China
{huyanli, liuteng20, zhouzhuangyi, zengweixin13, tanzhen08a, xiangzhao}@nudt.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) with external knowledge retrieval, improving factual accuracy and knowledge coverage. However, existing RAG approaches face a fundamental trade-off when handling complex reasoning: while traditional iterative retrieval methods offer flexibility, their local perspective limits their ability to establish global knowledge connections. In contrast, structure-augmented RAG methods capture global relationships but incur significant construction costs. To fill in this gap, we propose MGranRAG, an innovative framework designed to integrate precise local retrieval with structured global reasoning. Our approach circumvents expensive semantic extraction by employing a lightweight contextual hierarchical graph, effectively combining the local adaptability of iterative retrieval with the global consistency of structured knowledge. The framework adopts a novel iterative optimization scheme: at the local level, the LLM identifies multi-granular contextual evidence, such as key sentences and phrases, within retrieved passages to refine retrieval. At the global level, these multi-granularity evidence nodes are then mapped and propagated within the structured hierarchical graph, enabling the diffusion of rich contextual information at different levels to introduce global semantic constraints and reorder retrieval results. This coordination between local and global iterative processes dynamically balances retrieval accuracy and contextual coherence. Experimental results on challenging multi-hop and open-domain question answering datasets show that our proposal achieves new state-of-the-art performance in both retrieval and answer accuracy.

Code — https://github.com/MoMoLT/MGranRAG

Introduction

Human intelligence excels through continuous learning and dynamic knowledge integration, enabling effective navigation of complex and evolving environments. Developing AI systems with analogous capabilities is critical in high-stakes domains, where outcomes depend on factual accuracy, verifiable evidence chains, and rigorous reasoning. Large Language Models (LLMs), despite strong task performance, suffer from static pretrained knowledge that constrains real-world applicability in dynamic, information-critical domains (Mousavi, Alghisi, and Riccardi 2025; Song et al. 2025b). Retrieval-Augmented Generation (RAG) methods (Lewis et al. 2020; Guu et al. 2020) overcome this by dynamically incorporating external knowledge during inference, making it the prevailing approach in research and industry.

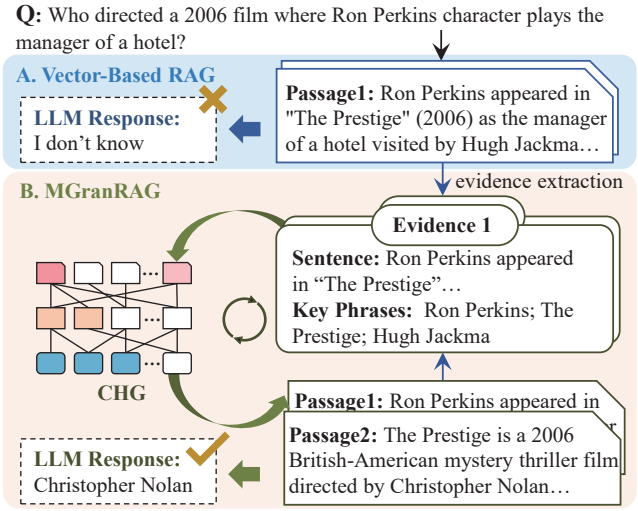

Figure 1: A case for question answering in vector-based RAG and our MGranRAG.

While RAG effectively mitigates static knowledge constraints in LLMs, its application to complex reasoning tasks reveals a fundamental limitation: evidence fragmentation. Conventional RAG systems (Karpukhin et al. 2020; Ma et al. 2023; Gao et al. 2023), typically instantiated as vector-based RAG, are optimized for local semantic similarity and retrieve topically relevant but logically isolated passages, obstructing the construction of coherent evidence chains required for multi-hop question answering (QA) (Yang et al. 2018; Trivedi et al. 2022b; Ho et al. 2020). To address evidence fragmentation, recent research has proposed adaptive RAG methods based on multi-turn iteration (Asai et al. 2023; Trivedi et al. 2022a; Zhuang et al. 2024; Guan et al.

*These authors contributed equally.
†Corresponding author

2025), which significantly improve reasoning flexibility through retrieve-generate-reflect cycles. However, these approaches remain fundamentally constrained by "local information horizons": their decision-making processes rely entirely on isolated, incrementally acquired evidence, failing to capture global associative structures within knowledge bases and are prone to suboptimal retrieval paths. To overcome local horizon limitations, structure-enhanced RAG methods (Li et al. 2024; Jimenez Gutierrez et al. 2024; Wang 2025; Liang et al. 2025) introduce explicit knowledge representations to enable global reasoning capabilities. Methods such as GraphRAG (Edge et al. 2024), HippoRAG 2 (Gutiérrez et al. 2025) and RAPTOR (Sarthi et al. 2024) leverage knowledge graphs or hierarchical structures to provide global semantic constraints for reasoning, significantly improving multi-hop reasoning performance. However, the prohibitive cost of structure construction severely limits practical deployment. Recent efforts have sought to mitigate this overhead. KET-RAG (Huang, Zhang, and Xiao 2025) adopts a hybrid indexing scheme that combines knowledge-centric skeletons with keyword-text bipartite graphs for efficient multi-hop retrieval, while E^2GraphRAG (Zhao et al. 2025) leverages lightweight entity extraction and streamlined bidirectional indexing to improve retrieval efficiency and evidence integration.

Despite progress in reducing the costs of structure construction and retrieval, existing structure-enhanced RAG approaches still struggle to balance structural expressiveness and retrieval adaptivity: adaptive methods often lack global constraints, while structured knowledge methods entail considerable overhead. To address this, we propose **M**ulti-**Gran**ular **RAG** (MGranRAG), a unified framework that integrates LLM reasoning with a lightweight **C**ontextual **H**ierarchical **G**raph (CHG) to achieve both local adaptability and global consistency, as illustrated in Figure 1. At the indexing stage, CHG efficiently encodes hierarchical relationships among paragraphs, sentences, and phrases using classical NLP techniques, creating a scalable, structured knowledge representation. Our approach introduces a collaborative refinement workflow that alternates between: **1)** Local evidence-augmented reranking by LLM-guided multi-granular evidence extraction. **2)** Global context-augmented reranking, which applies graph-based propagation over the CHG for global relevance calibration. This dual process tightly couples fine-grained local reasoning with explicit global structural modeling. Extensive experiments on multi-hop QA benchmarks confirm that MGranRAG substantially improves both reasoning consistency and answer completeness.

Our contributions are as follows:

- We introduce a Contextual Hierarchical Graph into RAG, a lightweight structured representation that captures multi-granular semantic relationships through the inherent hierarchical organization of documents.
- We propose MGranRAG, a novel RAG framework that integrates the local adaptivity of iterative RAG with the global semantic guidance of CHG.
- We conduct comprehensive experiments on six open-domain QA benchmarks, and the results demonstrate that our method achieves significant improvements on complex reasoning tasks.

Related Work

Iterative RAG for Complex Reasoning. RAG has become a pivotal approach for knowledge-intensive tasks, enhancing LLMs by integrating external knowledge sources. However, early RAG methods, which rely on static, single-step retrieval, often suffer from evidence fragmentation and a lack of global coherence, particularly in complex, multi-hop reasoning scenarios. To overcome these limitations, the field has progressed toward iterative RAG. This paradigm dynamically constructs evidence chains through multi-cycle "retrieve-then-reason" loops. Recent efforts have advanced this paradigm from several perspectives. One major thrust focuses on reasoning-guided retrieval, where methods like IRCoT (Trivedi et al. 2022a), DeepRAG (Guan et al. 2025), and ReaRAG (Lee et al. 2025) use intermediate reasoning steps to generate more precise subsequent queries. Another direction emphasizes self-correction and autonomy, with frameworks like Self-RAG (Asai et al. 2023) enabling the model to decide when to retrieve and to critique its own factual accuracy. Concurrently, approaches such as Auto-RAG (Yu, Zhang, and Feng 2024) and EfficientRAG (Zhuang et al. 2024) focus on automating and improving the efficiency of these iterative strategies.

Structure-Augmented RAG. To overcome the limited retrieval scope of purely iterative methods, a major line of work augments RAG with explicit structures to impose global semantic constraints and improve evidence integration. Prominent approaches in this area often construct explicit knowledge graphs. For instance, methods like GraphRAG (Edge et al. 2024) and KAG (Liang et al. 2025) utilize automatically built graphs of entities and relations to navigate multi-hop queries. This paradigm was further refined by HippoRAG and HippoRAG 2 (Jimenez Gutierrez et al. 2024; Gutiérrez et al. 2025), which introduced techniques such as personalized PageRank and rich paragraph-level embeddings to improve graph traversal and node representation. Concurrently, other works like GNN-RAG (Mavromatis and Karypis 2024) have employed Graph Neural Networks to fuse structural and semantic information for more relevant retrieval. A distinct strategy, exemplified by RAPTOR (Sarthi et al. 2024), adopts a hierarchical clustering and summarization approach, creating a tree structure that facilitates information retrieval across multiple granularities. Despite their efficacy in providing a global view, these methods are consistently challenged by the prohibitive computational overhead associated with constructing and maintaining their underlying structures.

To mitigate this cost, recent work has explored lightweight structured indexing. LightRAG (Guo et al. 2024) streamlines processes through multi-granularity entity matching. KET-RAG (Huang, Zhang, and Xiao 2025) introduces knowledge skeleton extraction strategies, achieving a balance between effectiveness and computational efficiency through structure simplification. E^2GraphRAG (Zhao et al.

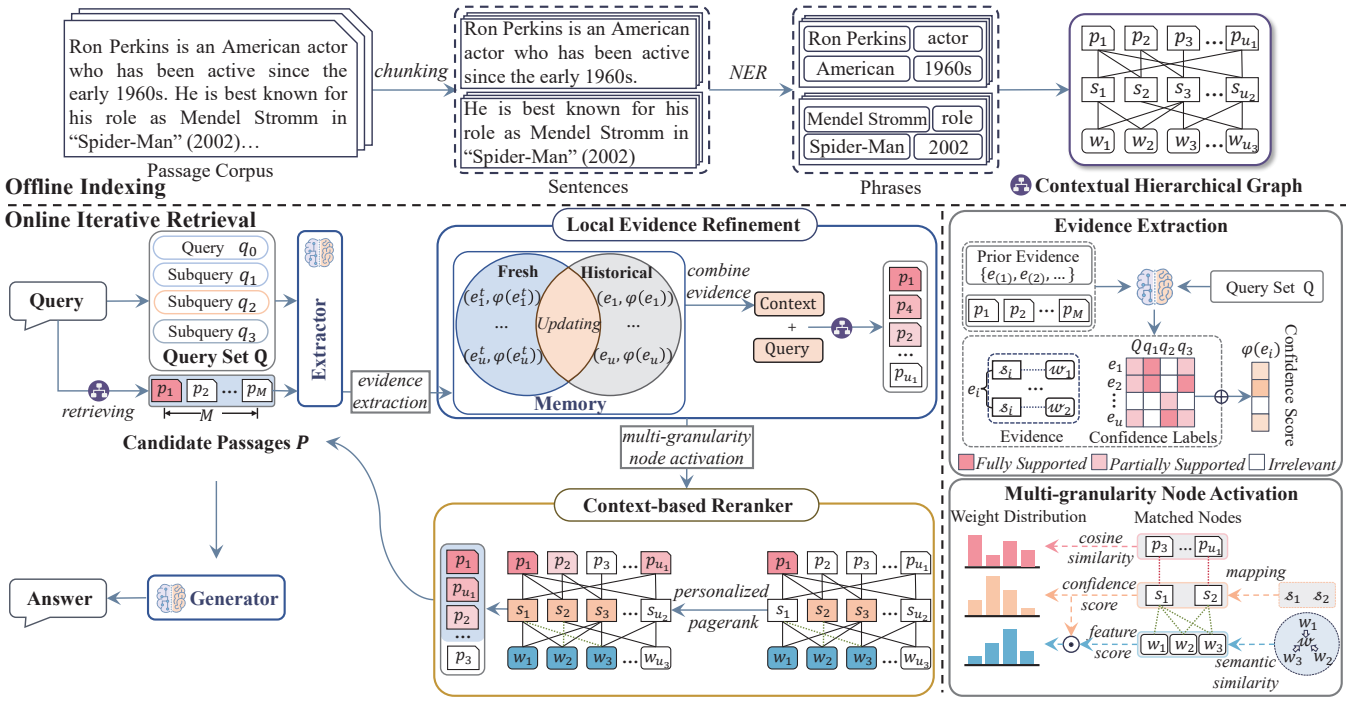

Figure 2: Overall framework of MGranRAG, which integrates multi-granular retrieval with the Contextual Hierarchical Graph.

2025) combines efficient summarization, lightweight entity extraction, and bidirectional indexing to enable faster multi-hop evidence retrieval.

The present work, MGranRAG, extends prior advances by integrating the adaptive retrieval mechanisms characteristic of iterative RAG frameworks with global semantic constraints informed by a low-cost structured knowledge representation. This approach is designed to ensure both global coherence and local optimality within the evidence aggregation process for complex reasoning tasks.

Preliminaries

This section establishes the theoretical foundation for our multi-granular graph-augmented retrieval framework. We first introduce the CHG, which captures information at multiple granularities, and then formalize the iterative evidence aggregation problem for complex reasoning tasks.

Contextual Hierarchical Graph

Given a corpus of passages $\mathcal{P}$, we construct a Contextual Hierarchical Graph (CHG), a lightweight multi-granular structure that organizes passages, sentences, and phrases from $\mathcal{P}$ into a unified graph. Formally, the CHG is defined as $G = (V, E)$, where V is the set of nodes and E is the set of edges. The node set $V = V^{(p)} \cup V^{(s)} \cup V^{(w)}$ is partitioned into three disjoint subsets, where $V^{(p)}$, $V^{(s)}$, and $V^{(w)}$ denote passage, sentence, and phrase nodes, respectively. Edges E encode hierarchical containment relations, linking passages to their sentences and sentences to their constituent phrases, thereby enabling information propagation across different granularities.

Problem Formulation

Traditional RAG frameworks aim to identify a relevant subset of passages $P_{rel} \subset \mathcal{P}$ for a given query q from a large-scale passage corpus $\mathcal{P}$, and use P_{rel} as context for an LLM to generate the final answer a. However, when handling queries requiring complex reasoning, standard RAG frameworks face a fundamental challenge: evidence fragmentation. Answers to such queries often depend on multiple evidence fragments scattered across different passages or even different sentences, with implicit logical connections between them. Standard retrieval models typically evaluate the semantic relevance of each passage to the query independently, lacking the ability to model structural relationships between passages and finer-grained information units (such as sentences and phrases), thus making it difficult to construct complete reasoning chains.

To address this challenge, we formalize the problem as a graph-based iterative evidence aggregation and ranking optimization problem. Our goal is not simply to perform one-time passage retrieval, but to design a dynamically evolving retrieval strategy capable of progressively discovering, connecting, and aggregating relevant evidence at different granularity levels.

Specifically, given a query q and a pre-constructed CHG G, our core task is to design an optimal ranking strategy Π. At iteration t, this strategy generates an updated passage ranking $P_{rank}^{(t)}$ based on the current query q, sub-queries Q, graph G, and historical evidence accumulated in the memory module $\mathcal{M}^{(t-1)}$:

$$P_{rank}^{(t)} = \Pi(q, Q, G, \mathcal{M}^{(t-1)})$$

The ranking $P_{rank}^{(t)}$ not only reflects the local relevance of individual passages but, more importantly, encapsulates global contextual information propagated through the graph structure.

Ultimately, the goal of this iterative process is to converge, within a finite number of iterations T, to an optimal candidate passage set $\mathcal{P}^{(T)}$. This set should satisfy the property of collective sufficiency; that is, the passages jointly provide all necessary evidence for generating the final answer a. In addition, structural coherence should be maintained, ensuring that the logical relationships among evidence pieces are preserved. To this end, our approach seeks to unify local information sensitivity and global semantic constraints in a cost-effective manner through the construction of a lightweight graph structure G and the design of an efficient iterative strategy Π, thereby optimizing both evidence aggregation and reasoning chain formation for complex reasoning tasks.

Methodology

To address the challenge of evidence fragmentation outlined in our problem formulation, we introduce MGranRAG, a unified framework designed for complex reasoning tasks. As shown in Figure 2, the central principle of MGranRAG is to orchestrate a dynamic, iterative synergy between the nuanced semantic understanding of an LLM and the global structural guidance of a lightweight, multi-granular graph.

Offline Indexing

The primary objective of this phase is to construct a low-cost, scalable CHG G from a large-scale corpus $\mathcal{P}$. Rather than employing computationally expensive semantic relation extraction, we focus on establishing multi-granular containment relationships. This initial graph provides a lightweight structural foundation for subsequent dynamic, LLM-driven interactions in the online phase.

The construction process, illustrated in Figure 2, utilizes the spaCy NLP toolkit. For each passage $p \in \mathcal{P}$, we first perform sentence segmentation to decompose it into a set of sentences S. Subsequently, for each sentence, we execute two parallel tasks to extract key phrases: 1) Noun Phrase Extraction using Part-of-Speech (POS) tagging and syntactic parsing rules. 2) Named Entity Recognition (NER) to identify and extract proper nouns and entities.

The union of outputs from these tasks forms our phrase set W. We retain the original text of these phrases without stemming or lemmatization to avoid potential semantic loss from over-normalization. Finally, based on the natural containment relationships, we construct the CHG. It is crucial to note that this offline-built graph has inherent structural limitations, such as restricted semantic connectivity and unresolved coreferences. These limitations are intentionally deferred to the online retrieval phase, where we employ LLM-based interactions to dynamically augment and refine the graph structure through semantic reasoning.

Online Iterative Retrieval

To realize the optimal ranking policy Π defined in Section "Problem Formulation" and fundamentally address the evidence fragmentation challenge, we design an online iterative retrieval framework. This framework, which instantiates Π, is centered on a self-enhancing refine-and-calibrate loop. This loop progressively optimizes the selection of a final passage set by reasoning deeply at the level of fine-grained evidence units, aiming to construct a context that satisfies both collective sufficiency and structural coherence.

At initialization ($t = 0$), the dense retriever $\mathcal{R}$ ranks all passages in the corpus $\mathcal{P}$ with respect to the query q_0 and selects the top M passages to form the initial candidate set $P_M^{(0)}$. A Chain-of-Thought (CoT) prompt then guides the LLM to decompose q into a set of atomic, verifiable sub-queries $Q = \{q_1, q_2, \ldots\}$. We also initialize an empty evidence memory $\mathcal{M}^{(0)}$ to store and manage discovered evidence across iterations.

In each iteration t, the framework alternates between two complementary steps that refine fine-grained evidence and update passage-level rankings.

Local Evidence Refinement. This step enhances local information sensitivity by using the LLM as a high-precision extractor of critical evidence from the current candidate set of passages $\mathcal{P}^{(t-1)}$. Given a prompt that combines the memory $\mathcal{M}^{(t-1)}$, the main query q, the sub-queries $Q = \{q_1, \ldots, q_K\}$, and the passages $\mathcal{P}^{(t-1)}$, the LLM identifies a set of new fine-grained evidence units $\mathcal{E}^{(t)} = \{\epsilon_1, \ldots, \epsilon_{N_e}\}$, where each evidence unit is represented as $\epsilon_i = (s_i, \mathcal{W}_i)$, with s_i denoting a sentence and $\mathcal{W}_i$ the associated set of key phrases extracted from s_i.

For each evidence unit ϵ_i, the LLM further assesses its semantic relation to the main query and all sub-queries, producing a discrete label $\mathcal{L}(\epsilon_i, q_k) \in \{Ts, Ps, Ns\}$, where Ts, Ps, and Ns denote fully supported, partially supported, and irrelevant evidence, respectively, where we treat $q_0 = q$ as the main query. These labels are mapped to numerical scores by a function $\mathrm{g}(\cdot)$, and aggregated into an overall confidence score:

$$\psi(\epsilon_i) = \mathrm{g}\big(\mathcal{L}(\epsilon_i, q)\big) + \frac{1}{|Q|} \sum_{q_k \in Q} \mathrm{g}\big(\mathcal{L}(\epsilon_i, q_k)\big) \qquad (1)$$

After updating the memory, we select the highest-scoring evidence from the previous memory $\mathcal{M}^{(t-1)}$ together with the newly added evidence in $\mathcal{E}^{(t)}$ to construct an evidence-augmented query $q_{aug}^{(t)}$. This query is then fed into the retriever $\mathcal{R}$ to produce an evidence-aware passage ranking $P_{rank-\epsilon}^{(t)}$.

Global Structure-Aware Calibration. In the global calibration phase, we translate locally refined evidence into a personalization vector over the CHG and perform structure-aware propagation via Personalized PageRank (PPR). After memory update at iteration t, we assign scores to nodes of different granularities to construct a multi-granular weight vector $\mathcal{W}^{(t)}$ over G.

For each passage node $v_{p_i} \in V^{(p)}$, we combine a decayed momentum term (from the previous PPR result) with the current similarity between the evidence-augmented query

	NQ	PopQA	MuSiQue	2Wiki	HotpotQA	NarrativeQA
# of Passages	8,676	9,633	11,656	6,119	9,811	4,111
# of Sentences	28,376	49,205	40,343	21,476	40,551	27,399
# of Phrases	142,899	133,741	158,595	82,388	153,416	29,493
# of Passage-to-Sentence Edges	50,887	29,869	40,905	21,551	40,726	60,951
# of Sentence-to-Phrase Edges	348,548	272,716	343,264	166,514	328,588	136,682

Table 1: Statistics of retrieval corpora and extracted nodes and edges for each dataset, with validation sets of 1,000 questions.

and the passage content:

$$\mathcal{W}^{(t)}(v_{p_i}) = \alpha_{decay} \cdot \mathcal{W}_G^{(t-1)}(v_{p_i}) + \omega_1 \cdot \mathrm{sim}\big(q_{aug}^{(t)}, c_{p_i}\big) \quad (2)$$

where α_{decay} controls the influence of previous iterations, c_{p_i} denotes the textual content of passage p_i, and $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity. This momentum term stabilizes passage rankings across iterations.

For a sentence node $v_{s_k} \in V^{(s)}$ that corresponds to an evidence unit $\epsilon_k \in \mathcal{E}^{(t)}$, we assign a weight proportional to its confidence:

$$\mathcal{W}^{(t)}(v_{s_k}) = \omega_2 \cdot \psi(\epsilon_k) \quad (3)$$

where $\psi(\epsilon_k)$ is the confidence score derived from the discrete label.

For each phrase node $v_{w_j} \in V^{(w)}$ that is obtained by aligning an LLM-extracted phrase to the CHG using semantic similarity, we aggregate evidence confidence, phrase features, and structural regularization:

$$\mathcal{W}^{(t)}(v_{w_j}) = \omega_3 \cdot \psi(\epsilon_i) \cdot f(w_j) \cdot d(v_{w_j})^{-1} \quad (4)$$

where ϵ_i is an associated evidence unit, $d(v_{w_j})$ is the node degree in G, and $f(w_j) = \lambda^\top \phi_{feat}(w_j)$ is a feature-based score capturing phrase type and origin. Here, ω_1, ω_2, and ω_3 are scaling factors that balance the contributions of passage, sentence, and phrase level signals in the multi-granular weight vector.

When the LLM discovers a new phrase or contextual relation that is not yet encoded in G, we incrementally augment the graph with the corresponding phrase nodes and edges, producing an updated graph $G^{(t)}$. This dynamic expansion allows the CHG to evolve with the reasoning process while incurring negligible additional overhead. The node weights $\mathcal{W}^{(t)}$ over $G^{(t)}$ define a personalization vector for PPR. We apply PPR to propagate local evidence signals through the graph structure:

$$\mathbf{r}^{(t)} = (1-\beta)\,\mathbf{P}^\top \mathbf{r}^{(t)} + \beta\,\mathbf{p}^{(t)} \quad (5)$$

where $\mathbf{P}$ is the transition matrix of $G^{(t)}$, $\mathbf{p}^{(t)}$ is the normalized personalization vector obtained from $\mathcal{W}^{(t)}$, and $\beta \in (0,1)$ is the restart probability. The resulting scores $\mathbf{r}^{(t)}$ induce a structure-aware passage ranking $P_{rank-G}^{(t)}$.

We select the top-M passages from $P_{rank-G}^{(t)}$ to form the next candidate set $\mathcal{P}^{(t)}$. Across iterations, this global structure-aware calibration transforms scattered local evidence into a globally coherent context. After T iterations, the final passage set $\mathcal{P}^{(T)}$ is optimized to satisfy both collective sufficiency and structural coherence, providing a robust evidence base for answer generation.

Experiments

Experimental Settings

Dataset. To assess the effectiveness of MGranRAG on knowledge-intensive QA, we conduct experiments on three multi-hop datasets—HotpotQA (Yang et al. 2018), 2WikiMultiHopQA (2Wiki) (Ho et al. 2020), and MuSiQue (Trivedi et al. 2022b)—and two single-hop benchmarks: Natural Questions (NQ) (Wang et al. 2024) and PopQA (Mallen et al. 2022). In addition, we evaluate discourse-level understanding on NarrativeQA (Kočiský et al. 2018). To ensure fair comparison, we utilize 1,000 questions extracted from each validation set and incorporate the retrieval corpus pertinent to the selected questions, meticulously crafted by the authors of HippoRAG 2. More details of the datasets are provided in Table 1.

Evaluation Metrics. Following HippoRAG 2 (Gutiérrez et al. 2025), we evaluate the retrieval performance using Recall@5 (hit rate of the top-5 retrieved passages). For QA performance, we use Exact Match and F1 score as evaluation metrics.

Baselines. We benchmark MGranRAG against a diverse set of baseline methods, broadly categorized into vector-based RAG and structure-augmented RAG. Within the vector-based paradigm, we distinguish two subgroups based on model capacity: **1)** Simple Baselines encompassing BM25 (Robertson and Walker 1994), Contriever (Izacard et al. 2021), and GTR (Ni et al. 2021). **2)** Large Embedding Models representing state-of-the-art dense retrieval architectures, specifically including Alibaba-NLP/GTE-Qwen2-7B-Instruct (Kwon et al. 2023), GritLM/GritLM-7B, and nvidia/NV-Embed-v2 (Muennighoff et al. 2024). We also evaluate iterative methods, including IRCoT (Trivedi et al. 2022a), Search-R1 (Jin et al. 2025), and R1-Searcher (Song et al. 2025a). For structure-augmented approaches, we rigorously evaluate established methods such as RAPTOR (Sarthi et al. 2024), GraphRAG (Edge et al. 2024), LightRAG (Guo et al. 2024), HippoRAG (Jimenez Gutierrez et al. 2024), and its advanced variant HippoRAG 2 (Gutiérrez et al. 2025).

Implementation Details. For our proposed MGranRAG, we perform noun phrase extraction using spaCy. We use the open-source Qwen3-8B and Qwen2.5-7B models for both query decomposition and evidence extraction. These tasks are executed in a few-shot CoT setting, and we explicitly disable the internal thinking process of the model by including the */no_think* command in the prompt. For iterative RAG methods, we follow the original papers and use their released, fine-tuned Qwen2.5-7B models during

Method	NQ			PopQA			MuSiQue			2Wiki			HotpotQA			NarrativeQA		Avg		
	R@5	EM	F1	R@5	EM	F1	R@5	EM	F1	R@5	EM	F1	R@5	EM	F1	EM	F1	R@5	EM	F1
No Retrieval	–	40.2	54.9	–	28.2	32.5	–	17.6	26.1	–	36.5	42.8	–	37.0	47.3	3.4	12.9	–	27.2	34.8
BM25	56.1	44.7	59.0	35.7	39.1	49.9	43.5	20.3	28.8	65.3	47.9	51.2	74.8	52.0	63.4	4.4	18.3	55.1	34.7	43.6
Contriever	54.6	45.0	58.9	43.2	41.6	53.1	46.6	24.0	31.3	57.5	38.1	41.9	75.3	51.3	62.3	6.5	19.7	55.4	34.4	43.1
GTR (T5-base)	63.4	45.5	59.9	49.4	43.2	56.2	49.1	25.8	34.6	67.9	49.2	52.8	73.9	50.6	62.8	6.8	19.9	60.7	36.9	46.2
GTE-Qwen2-7B-Instruct	74.3	46.6	62.0	50.6	43.5	56.3	63.6	30.6	40.9	74.8	55.1	60.0	89.1	58.6	71.0	7.9	21.3	70.5	40.4	50.3
GritLM-7B	76.6	46.8	61.3	50.1	42.8	55.8	65.9	33.6	44.8	76.0	55.8	60.6	92.4	60.7	73.3	8.2	23.9	72.2	41.3	51.6
NV-Embed-v2	75.4	47.3	61.9	51.0	42.9	55.7	69.7	34.7	45.7	76.5	57.5	61.5	94.5	62.8	75.3	**8.9**	25.7	73.4	42.4	52.6
IRCoT	–	33.1	49.2	–	36.8	48.7	–	22.3	33.3	–	45.0	55.1	–	53.7	67.1	5.8	23.0	–	32.8	44.2
Search-R1	–	**54.5**	63.2	–	**50.9**	**58.5**	–	31.8	41.0	–	50.4	57.6	–	54.7	66.2	6.5	20.7	–	41.5	49.8
R1-Searcher	–	52.8	**63.4**	–	40.2	54.7	–	**43.6**	**54.0**	–	65.3	73.0	–	63.4	76.0	7.8	16.4	–	45.5	54.7
GraphRAG	–	30.8	46.9	–	31.4	48.1	–	27.3	38.5	–	51.4	58.6	–	55.2	68.6	6.8	23.0	–	33.8	45.4
LightRAG	–	8.6	16.6	–	2.1	2.4	–	0.5	1.6	–	9.4	11.6	–	2.0	2.4	1.0	3.7	–	3.9	6.0
RAPTOR	68.3	36.9	50.7	48.7	43.1	56.2	57.8	20.7	28.9	66.2	47.3	52.1	86.9	56.8	69.5	5.1	21.4	65.6	35.0	44.8
HippoRAG	44.4	43.0	55.3	53.8	42.7	55.9	53.2	26.2	35.1	90.4	65.0	71.8	77.3	52.6	63.5	4.4	16.3	63.8	39.0	48.1
HippoRAG 2	**78.0**	48.6	63.3	51.7	42.9	56.2	74.7	37.2	48.6	90.4	65.0	71.0	96.3	62.7	75.5	**8.9**	25.9	78.2	44.2	55.0
MGranRAG (Qwen2.5-7B) T=1	76.3	45.4	60.5	68.1	47.0	57.7	76.2	40.1	51.6	91.3	62.9	75.7	97.6	**67.8**	74.8	7.9	23.6	81.9	45.2	55.6
MGranRAG (Qwen2.5-7B) T=3	76.1	46.3	60.7	68.7	47.1	58.1	**77.1**	40.8	52.3	95.0	**69.7**	**76.4**	97.5	62.3	75.2	8.2	25.3	82.9	**45.7**	**56.2**
MGranRAG (Qwen3-8B) T=1	77.3	47.0	61.5	72.9	43.6	56.0	74.4	39.5	50.9	93.7	67.7	75.3	**98.2**	63.5	**76.1**	7.9	23.9	83.3	45.0	55.5
MGranRAG (Qwen3-8B) T=3	77.8	47.3	61.8	**73.7**	44.2	56.2	75.5	40.6	52.2	**95.9**	69.8	77.1	**98.2**	63.5	**76.1**	8.3	**26.0**	**84.2**	45.3	**56.2**

Table 2: Comparison of RAG methods across six benchmarks. Performance is reported in terms of R@5 (Recall@5), EM (Exact Match), and F1 score. Best results are highlighted in bold, and secondary scores are underlined.

Method	Index		Retrieval	
	Input	Output	Input	Output
MGranRAG (T=1)	–	–	4.5M	0.7M
MGranRAG (T=3)	–	–	12.0M	1.5M
HippoRAG 2	5.8M	2.4M	3.0M	0.4M
R1-Searcher	–	–	4.0M	0.3M
Search-R1	–	–	3.8M	0.2M

Table 3: Comparison of LLM token usage in the indexing and retrieval stages on the 2Wiki dataset.

Method	PopQA	2Wiki	HotpotQA
Naive RAG	56.6	61.5	75.3
+ CHG	56.6	80.1	97.0
+ QD	58.7	84.5	97.4
+ EAR (MGranRAG)	72.9	93.7	98.2

Table 4: Ablation study of MGranRAG components on PopQA and two multi-hop QA datasets, evaluated by Recall@5.

the retrieval–reasoning process. For the structure-augmented RAG baseline, we perform knowledge triple extraction using Llama-3.3-70B-Instruct. To ensure a fair comparison, all methods use NV-Embed-v2 as the retriever and Llama-3.3-70B-Instruct for final answer generation. We use a PPR damping factor of 0.5, perform $t = 3$ iterations, and set a semantic similarity threshold of 0.9 for phrase node matching. We set the maximum number of passages for evidence extraction to $M = 10$. The final relevance score is computed as a weighted sum of scores from different granularities, with scaling factors $\omega_1 = 0.1$, $\omega_2 = 0.5$, and $\omega_3 = 0.6$ for passage, sentence, and phrase levels, respectively. All experiments were performed on a single NVIDIA A100 80GB GPU.

Retrieval Performance. As shown in Table 2, we compare the retrieval and QA performance of MGranRAG with strong baselines on six benchmarks. In the few-shot RAG setting, MGranRAG consistently outperforms prior methods on most datasets, achieving an average Recall@5 of 84.2% under the Qwen3-8B reader with the $T = 3$ configuration. Compared with the NV-Embed-v2 baseline (73.4%), our method yields large gains on challenging datasets: NQ (+2.4%), PopQA (+22.7%), MuSiQue (+5.8%), 2Wiki (+19.5%) and HotpotQA (+3.7%). It also surpasses strong structure-augmented methods such as HippoRAG 2, attaining the highest Recall@5 on four benchmarks and reaching 98.2% on HotpotQA. The iterative retrieval process is effective: even with a single round ($T = 1$), MGranRAG already outperforms all baselines on average, while increasing T to 3 brings further gains, especially on complex datasets like 2Wiki. These results strongly validate our core mechanism of leveraging global context to guide local information retrieval.

QA Performance. All QA experiments use Llama-3.3-70B-Instruct as the reader. As shown in Table 2, MGranRAG

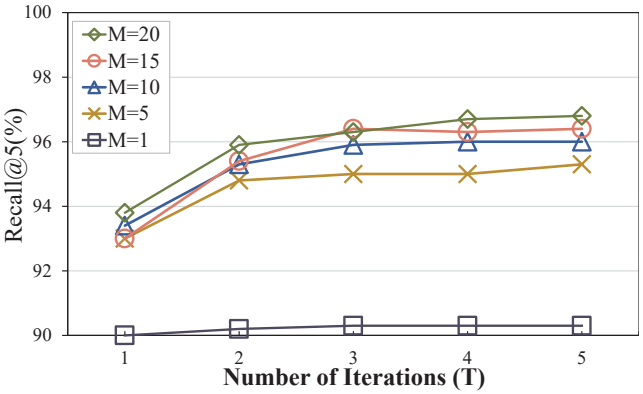

Figure 3: Impact of Iterations (T) and Candidate Passage Count (M) on Retrieval Performance. Recall@5 on the 2Wiki dataset is plotted against the number of iterations for varying candidate passage counts.

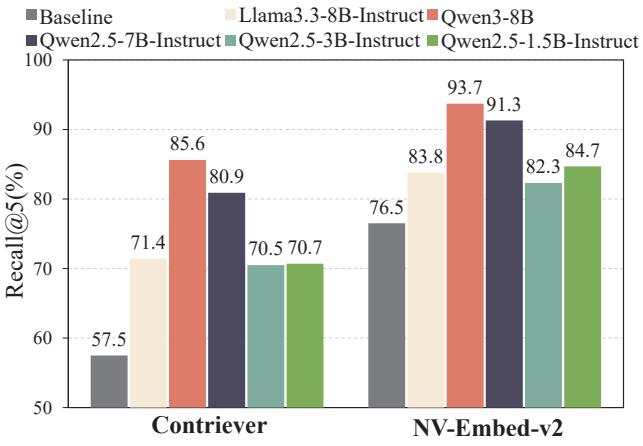

Figure 4: Comparison of Recall@5 performance on the 2Wiki dataset across different combinations of retrievers and LLM-based evidence extractors.

(Qwen3-8B, $T = 3$) achieves an average F1 score of 56.2%, outperforming all baselines, including HippoRAG 2 (55.0%). Moreover, compared with the recent R1-Searcher approach, MGranRAG, although using only a few-shot prompting configuration, achieves comparable QA performance on the three multi-hop QA benchmarks. These results indicate that MGranRAG is an effective RAG framework that substantially enhances the reasoning and generation capabilities of LLMs through iterative, multi-granular retrieval.

Ablation Study

To evaluate the contribution of the core components of MGranRAG, we performed ablation studies summarized in Table 4, using Naive RAG (NV-Embed-v2) as the baseline. All ablation experiments are conducted with the Qwen3-8B reader under a single-iteration setting ($T = 1$). Incrementally introducing major modules demonstrates their cumulative effect on retrieval performance.

Introducing the Contextual Hierarchical Graph (+CHG) leads to substantial gains on complex multi-hop datasets such as 2Wiki and HotpotQA (Recall@5 improved by 18.6% and 21.7%, respectively), underscoring the value of contextual structure for challenging queries. The query decomposition (+QD) strategy yields further improvements beyond +CHG, and proves robust even on simpler datasets like PopQA. The fully integrated MGranRAG, which incorporates evidence-augmented reranking (EAR), achieves the best results across all benchmarks. Notably, while improvements are marginal on PopQA for individual components, the complete framework still delivers a significant 16.3% increase, underscoring the synergistic effect of component integration and the importance of global context.

We further analyze the influence of the candidate passage pool size (M) on retrieval effectiveness for 2Wiki (Figure 3). Under the $T = 1$ setting, performance consistently improves as M increases up to a moderate range, after which gains saturate. At $M = 10$ and $M = 15$, the model reaches around 96% Recall@5. When M is too small, potential improvements are limited, indicating that an appropriate pool size is essential for fully leveraging multi-granular evidence in complex question answering. Considering efficiency and performance, we choose $M = 10$ and $T = 1$ as the default configuration, balancing coverage and computational cost.

Plug-and-Play Design and Computational Cost

To evaluate generalizability and plug-and-play capability, we assess MGranRAG on 2Wiki using different dense retrievers and LLM-based evidence extractors. As shown in Figure 4, NV-Embed-v2 consistently outperforms Contriever across all model pairings, yielding substantially higher Recall@5 with every backbone LLM. This confirms that retriever choice is crucial and that MGranRAG can effectively exploit stronger retrievers and LLMs in a plug-and-play manner, without changing the pipeline.

We also compare the computational cost of MGranRAG with three recent baselines in Table 3. While MGranRAG achieves the best retrieval effectiveness, it incurs higher LLM token usage, primarily due to long, carefully engineered prompts. Each evidence extraction step uses a fixed prompt of about 3k tokens. Reducing this overhead via prompt optimization or model fine-tuning is left for future work.

Conclusions and Future Work

In this work, we propose MGranRAG, a RAG framework that mitigates evidence fragmentation in complex question answering. MGranRAG combines a lightweight Contextual Hierarchical Graph with multi-granularity iterative retrieval to progressively build coherent reasoning paths. On challenging multi-hop QA benchmarks, MGranRAG significantly outperforms state-of-the-art RAG systems in both retrieval quality and answer accuracy. Although effective, the iterative process introduces additional computational overhead, and further work is needed to develop more efficient variants and extend the framework to broader complex reasoning tasks.

Acknowledgments

We sincerely thank all anonymous reviewers and meta-reviewers for their helpful comments, which have greatly improved this paper. We also acknowledge the contributions and support of all team members involved in this project. This work was partially supported by National Natural Science Foundation of China (72471237, 72371245, 62302513, 62272469).

References

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-rag: Self-reflective retrieval augmented generation. In *NeurIPS 2023 workshop on instruction tuning and instruction following*.

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R. O.; and Larson, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Guan, X.; Zeng, J.; Meng, F.; Xin, C.; Lu, Y.; Lin, H.; Han, X.; Sun, L.; and Zhou, J. 2025. DeepRAG: Thinking to Retrieve Step by Step for Large Language Models. *arXiv preprint arXiv:2502.01142*.

Guo, Z.; Xia, L.; Yu, Y.; Ao, T.; and Huang, C. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Gutiérrez, B. J.; Shu, Y.; Qi, W.; Zhou, S.; and Su, Y. 2025. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, 3929–3938. PMLR.

Ho, X.; Nguyen, A.-K. D.; Sugawara, S.; and Aizawa, A. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Huang, Y.; Zhang, S.; and Xiao, X. 2025. Ket-rag: A cost-efficient multi-granular indexing framework for graph-rag. *arXiv preprint arXiv:2502.09304*.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Jimenez Gutierrez, B.; Shu, Y.; Gu, Y.; Yasunaga, M.; and Su, Y. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37: 59532–59569.

Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arik, S.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P. S.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP (1)*, 6769–6781.

Kočiský, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6: 317–328.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, 611–626.

Lee, Z.; Cao, S.; Liu, J.; Zhang, J.; Liu, W.; Che, X.; Hou, L.; and Li, J. 2025. Rearag: Knowledge-guided reasoning enhances factuality of large reasoning models with iterative retrieval augmented generation. *arXiv preprint arXiv:2503.21729*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, S.; He, Y.; Guo, H.; Bu, X.; Bai, G.; Liu, J.; Liu, J.; Qu, X.; Li, Y.; Ouyang, W.; et al. 2024. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. *arXiv preprint arXiv:2406.14550*.

Liang, L.; Bo, Z.; Gui, Z.; Zhu, Z.; Zhong, L.; Zhao, P.; Sun, M.; Zhang, Z.; Zhou, J.; Chen, W.; et al. 2025. Kag: Boosting llms in professional domains via knowledge augmented generation. In *Companion Proceedings of the ACM on Web Conference 2025*, 334–343.

Ma, X.; Gong, Y.; He, P.; Zhao, H.; and Duan, N. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5303–5315.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Mavromatis, C.; and Karypis, G. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.

Mousavi, S. M.; Alghisi, S.; and Riccardi, G. 2025. Llms as repositories of factual knowledge: Limitations and solutions. *arXiv preprint arXiv:2501.12774*.

Muennighoff, N.; Hongjin, S.; Wang, L.; Yang, N.; Wei, F.; Yu, T.; Singh, A.; and Kiela, D. 2024. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

Ni, J.; Qu, C.; Lu, J.; Dai, Z.; Abrego, G. H.; Ma, J.; Zhao, V. Y.; Luan, Y.; Hall, K. B.; Chang, M.-W.; et al. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

Robertson, S. E.; and Walker, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic

weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, 232–241. Springer.

Sarthi, P.; Abdullah, S.; Tuli, A.; Khanna, S.; Goldie, A.; and Manning, C. D. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Song, H.; Jiang, J.; Min, Y.; Chen, J.; Chen, Z.; Zhao, W. X.; Fang, L.; and Wen, J.-R. 2025a. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Song, Z.; Yan, B.; Liu, Y.; Fang, M.; Li, M.; Yan, R.; and Chen, X. 2025b. Injecting domain-specific knowledge into large language models: a comprehensive survey. *arXiv preprint arXiv:2502.10708*.

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

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022b. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10: 539–554.

Wang, J. 2025. PropRAG: Guiding Retrieval with Beam Search over Proposition Paths. *arXiv preprint arXiv:2504.18070*.

Wang, Y.; Ren, R.; Li, J.; Zhao, W. X.; Liu, J.; and Wen, J.-R. 2024. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering. *arXiv preprint arXiv:2402.17497*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Yu, T.; Zhang, S.; and Feng, Y. 2024. Auto-rag: Autonomous retrieval-augmented generation for large language models. *arXiv preprint arXiv:2411.19443*.

Zhao, Y.; Zhu, J.; Guo, Y.; He, K.; and Li, X. 2025. Eˆ 2GraphRAG: Streamlining Graph-based RAG for High Efficiency and Effectiveness. *arXiv preprint arXiv:2505.24226*.

Zhuang, Z.; Zhang, Z.; Cheng, S.; Yang, F.; Liu, J.; Huang, S.; Lin, Q.; Rajmohan, S.; Zhang, D.; and Zhang, Q. 2024. Efficientrag: Efficient retriever for multi-hop question answering. *arXiv preprint arXiv:2408.04259*.