

HEAR-RAG: A Hypergraph-Based Evidence-Aware Retrieval-Augmented Generation Framework for Multi-Hop Question Answering

Changbin Lv¹ Weihua Wang^{1,2,3*} Lianlong Tao¹ Jiale Xing¹ Feilong Bao^{1,2,3}

¹ College of Computer Science, Inner Mongolia University, Hohhot, China

² National and Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Hohhot, China

³ Inner Mongolia Key Laboratory of Multilingual Artificial Intelligence Technology, Hohhot, China

32409206@mail.imu.edu.cn wangwh@imu.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) improves model factuality, but multi-hop question answering still suffers from incomplete and disconnected evidence retrieval across documents or chunks. Although graph-enhanced RAG methods organize evidence with graphs or hypergraphs, existing structures still struggle to balance the completeness of higher-order factual structures with fine-grained evidence connectivity. Graph-based methods mainly rely on pairwise edges and may fail to preserve complex multi-entity facts, while hypergraph-based methods can capture higher-order contexts but may underuse low-order relations for step-by-step evidence tracing. Both types of methods may further suffer from evidence fragmentation when semantically equivalent entity mentions are not effectively organized across chunks. To address these issues, we propose HEAR-RAG, a hypergraph-based evidence-aware retrieval framework for multi-hop question answering. HEAR-RAG organizes distributed evidence into a unified factual evidence hypergraph that jointly preserves higher-order factual contexts, low-order relational clues, and source chunks. It further introduces relation-aware entity mention clustering to reduce evidence fragmentation, and uses hybrid structural retrieval to construct connected evidence paths for answer generation. Experiments on multiple multi-hop QA benchmarks demonstrate its effectiveness.

1 Introduction

Retrieval-Augmented Generation (RAG) improves the factuality and domain adaptability of large language models (LLMs) by incorporating external documents into generation (Lewis et al., 2020; Gao et al., 2024b). In practice, these documents are often split into text chunks, which are retrieved

and re-ranked through lexical or semantic similarity (Jiang et al., 2023; Robertson and Zaragoza, 2009; Karpukhin et al., 2020; Shao et al., 2023; Huang et al., 2025). However, similarity-based retrieval mainly ranks chunks by their individual relevance to the query, rather than by their roles in a connected reasoning process. This makes conventional RAG insufficient for multi-hop questions, where the answer depends on composing evidence distributed across multiple documents or chunks (Zhang et al., 2025).

Structure-enhanced RAG systems move beyond independent chunk retrieval by organizing entities and relations into structured indexes. Representative methods include graph-based and hierarchy-based RAG systems (Edge et al., 2025; Guo et al., 2025; Wang et al., 2024; Gutiérrez et al., 2024, 2025). However, ordinary graphs mainly rely on pairwise edges, which makes them vulnerable to fragmented entity mentions across chunks and insufficient for preserving higher-order factual contexts involving multiple entities. Hypergraphs address the latter limitation by allowing a hyperedge to connect multiple nodes simultaneously (Antelmi et al., 2023; Feng et al., 2019), and recent hypergraph-based RAG methods use hyperedges to model higher-order factual structures (Luo et al., 2025; Hu et al., 2026). Nevertheless, higher-order hyperedges alone cannot fully support stable multi-hop retrieval. Although these methods can preserve local multi-entity contexts, they often underuse low-order relations that provide local connectivity cues for step-by-step evidence tracing, path expansion, and noise-controlled retrieval. As a result, retrieval may suffer from insufficient depth, weak cross-chunk connectivity, and redundant expansion through weakly relevant hyperedges. This motivates us to jointly model higher-order factual contexts, low-order relational clues, and cross-chunk mention associations within a unified retrieval structure.

*Corresponding author.

Our code is publicly available at <https://github.com/LeoIR21/hear-rag>.

In this paper, we propose HEAR-RAG, a retrieval-augmented framework for multi-hop question answering based on a factual evidence hypergraph. HEAR-RAG jointly organizes low-order relations and factual hyperedges within a unified evidence structure. Factual hyperedges preserve source-grounded multi-entity contexts, while low-order relations provide fine-grained transition signals for local traversal, multi-hop path expansion, and noise-controlled evidence selection. To reduce evidence fragmentation caused by text chunking, HEAR-RAG further introduces relation-aware entity mention clustering, which connects semantically equivalent mentions across chunks while retaining their original forms and source chunks. During retrieval, HEAR-RAG constructs generation-oriented evidence paths over the factual evidence hypergraph.

The main contributions of this paper are as follows:

- We propose a factual evidence hypergraph-based retrieval framework that jointly organizes low-order relations, higher-order factual structures, and source chunks for multi-hop question answering.
- We introduce relation-aware entity mention clustering and factual hyperedges to improve cross-chunk evidence connectivity and preserve multi-entity factual contexts.
- We design a hybrid structural retrieval strategy for constructing generation-oriented evidence paths, and experiments on multiple multi-hop QA benchmarks demonstrate the effectiveness of HEAR-RAG.

2 Related Work

RAG for Multi-hop QA. Retrieval-Augmented Generation (RAG) improves the factuality and adaptability of large language models by introducing external non-parametric knowledge. In multi-hop QA, answers often rely on evidence distributed across multiple documents or chunks, requiring retrieval systems to support both evidence recall and reasoning.

Beyond early retrieve-then-read methods based on lexical or semantic similarity, recent studies couple retrieval with reasoning. IRCOT interleaves retrieval with chain-of-thought reasoning (Trivedi et al., 2023); EfficientRAG reduces repeated LLM calls during iterative retrieval (Zhuang et al., 2024);

Q-DREAM optimizes the semantic space of sub-questions for dynamic multi-hop retrieval (Ye et al., 2025); and ChainRAG mitigates key-entity loss by constructing progressive retrieval chains (Zhu et al., 2025). These works suggest that multi-hop QA requires reasoning-aware retrieval rather than one-shot passage matching.

Structure-Enhanced and Hypergraph-Based RAG. Another line of work improves RAG by organizing corpora into structured indexes. Instead of treating documents as independent chunks, these methods build graphs, hierarchies, or memory-like structures to support long-range evidence discovery. GraphRAG constructs entity graphs and community summaries from text for global and local retrieval (Edge et al., 2025); LightRAG integrates graph structures into text indexing and combines low-level and high-level retrieval (Guo et al., 2025); HippoRAG2 enhances associative retrieval with knowledge-graph-based memory and graph search (Gutiérrez et al., 2025); RAPTOR organizes documents through recursive tree-structured summarization for hierarchical retrieval (Sarthi et al., 2024); and LeanRAG builds hierarchical knowledge graphs through semantic aggregation for structure-guided retrieval (Zhang et al., 2026). These studies show that structured indexes can better organize distributed evidence and support complex reasoning.

Beyond ordinary graph or hierarchical structures, hypergraph-based retrieval uses hyperedges to represent higher-order relations and multi-entity facts (Luo et al., 2025; Hu et al., 2026). Such structures are suitable for preserving complex factual contexts, but they still require fine-grained relational paths and effective cross-chunk entity connections to support stable multi-hop retrieval. Compared with hypergraph retrieval methods that primarily emphasize higher-order structures, our work unifies low-order relational paths, factual hyperedges, and source chunks to support traceable path expansion and source-grounded multi-hop evidence retrieval.

Entity Mention Clustering. Entity mention organization has been studied in entity linking, entity disambiguation, and cross-document coreference resolution. Recent entity linking methods use large language models to enrich mention contexts and improve entity disambiguation (Vollmers et al., 2025; Ye and Mitchell, 2025). Meanwhile, cross-document coreference studies cluster related mentions across documents with LLM-based collabora-

tion or document-level structural information (Min et al., 2024; Gao et al., 2024a). Aligned-entity embedding methods further learn fusion representations to connect corresponding entities across hetero-field knowledge graphs (Xiao et al., 2025). Unlike these tasks, we build retrieval-oriented entity mention clusters directly from the given corpus rather than relying on external entity identifiers. These clusters connect semantically equivalent mentions across chunks while preserving original mentions, aliases, and source chunks for multi-hop retrieval.

3 Method

3.1 Overview of HEAR-RAG

Given a document corpus, we first split the documents into a set of source text chunks $\mathcal{C}$. Conventional RAG typically retrieves the top- k chunks according to the semantic similarity between the query q and each chunk. However, in multi-hop question answering, the answer often depends on a chain of evidence distributed across multiple chunks, and similarity-based retrieval alone may fail to cover the complete reasoning path.

To address this issue, we construct a factual evidence hypergraph:

$$\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{H}, \mathcal{C}), \quad (1)$$

where $\mathcal{V}$ denotes the set of mention-preserving entity clusters, $\mathcal{R}$ denotes the set of low-order relation instances, $\mathcal{H}$ denotes the set of factual hyperedges, and $\mathcal{C}$ denotes the set of source text chunks. Low-order relation instances provide fine-grained connectivity, factual hyperedges preserve higher-order factual structures, and source chunks provide the evidence context required for generation. As illustrated in Figure 1, HEAR-RAG consists of two stages: offline factual evidence hypergraph construction and online hybrid structural retrieval.

3.2 Offline Factual Evidence Hypergraph Construction

The offline stage converts the source chunk set $\mathcal{C}$ into a retrievable factual evidence hypergraph. This process consists of local low-order structure extraction, factual hyperedge construction, and relation-aware entity mention clustering. The first two steps preserve local factual structures within each source chunk, while the third connects semantically equivalent entity mentions across different chunks for subsequent retrieval.

Local Low-order Structure Extraction. For each source chunk $c_i \in \mathcal{C}$, the system first extracts entity mentions and then extracts relations under the constraint of the extracted entity mention set:

$$E_i = f_{\text{NER}}(c_i), \quad R_i = f_{\text{REL}}(c_i, E_i), \quad (2)$$

where E_i denotes the set of entity mentions extracted from chunk c_i , and R_i denotes the corresponding set of low-order relation instances. Each relation instance is represented as a triple $r = (s, p, o)$, where s and o denote the subject and object entity mentions, respectively, and p denotes the relation predicate.

To improve local structural consistency, we retain relation instances supported by the extracted entity mentions and associate them with their corresponding evidence sentences and source chunks. These local structures are then used for factual hyperedge construction.

Factual Hyperedge Construction. Low-order relation instances provide fine-grained relational clues but may not by themselves preserve coherent multi-entity factual contexts. Therefore, we use LLM-based evidence induction to group related relation instances and their associated evidence sentences into factual evidence units. The induction operates only on the extracted relation instances and does not introduce new entities, relations, or unsupported facts. For each source chunk c_i , a factual hyperedge is defined as:

$$h = (E_h, R_h, z_h, c_i), \quad (3)$$

where $E_h \subseteq E_i$ denotes the entity mentions involved in the hyperedge, $R_h \subseteq R_i$ denotes the grouped low-order relation instances, z_h denotes the associated evidence text, and c_i denotes the source chunk identifier.

The generated evidence units are subsequently validated according to local entity connectivity and evidence-sentence continuity before being converted into factual hyperedges. The resulting hyperedges preserve the grouped relation instances, associated entity mentions, evidence spans, and source chunk identifiers. Relation instances that cannot be assigned to a valid evidence unit are retained as fallback factual hyperedges to avoid losing potentially useful evidence for subsequent path expansion and evidence recovery.

Relation-aware Entity Mention Clustering. During global factual evidence hypergraph construction, entity mentions extracted independently

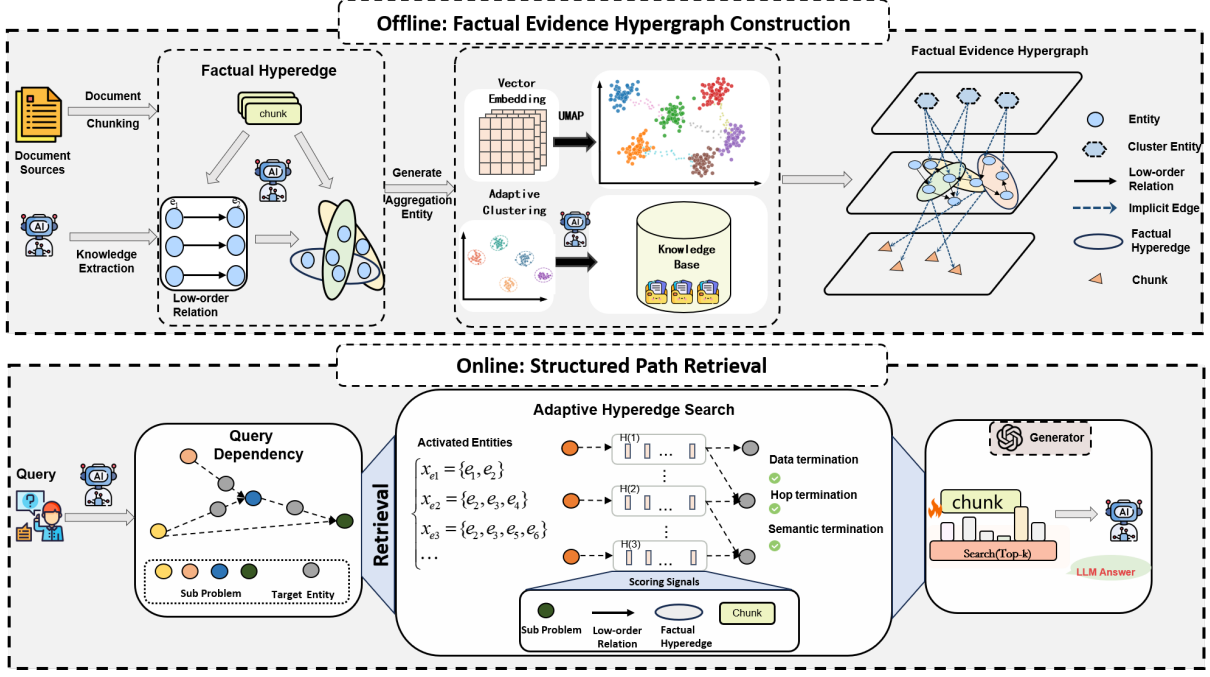


Figure 1: Overview of the HEAR-RAG framework.

from different source chunks may express semantically equivalent entities or concepts using different surface forms. Treating these mentions as independent nodes can scatter related evidence and fragment cross-chunk reasoning paths, whereas direct normalization may remove useful mention-level distinctions. To address this issue, we organize semantically equivalent mentions into mention-preserving entity clusters, enabling cross-chunk connectivity while preserving the original mention and source-chunk associations.

Let $\mathcal{E} = \bigcup_i E_i$ denote the set of entity mentions extracted from all source chunks. We first generate high-recall candidate mention groups using name- and profile-based representations. Entity-centered evidence from the factual hypergraph, including relevant low-order relation instances, hyperedge descriptions, and local neighboring entities, is collected for subsequent verification.

For an oversized candidate group $\mathcal{B}$, we apply UMAP and HDBSCAN to its profile embeddings:

$$\mathcal{K} = \text{HDBSCAN}(\text{UMAP}(\Phi(\mathcal{B}))), \quad (4)$$

where $\mathcal{K} = \{\mathcal{K}_j\}_{j=1}^J$ denotes the resulting candidate blocks, and $\Phi(\mathcal{B})$ denotes the encoded name- and profile-based representations of the candidate group. UMAP reduces the dimensionality of the candidate group, while HDBSCAN identifies locally dense candidate blocks using adaptive parameters. Candidate groups below the predefined size

limit are retained directly as candidate blocks.

Subsequently, a large language model verifies selected candidate blocks using their entity profiles and local evidence from the factual hypergraph. For a candidate block $\mathcal{K}_j \in \mathcal{K}$, the model identifies mentions that refer to the same entity or concept:

$$\mathcal{S}_j = \text{LLM}_{\text{verify}}(P_{\text{verify}}, \mathcal{K}_j), \quad (5)$$

where $\mathcal{S}_j$ denotes the accepted mention groups. The prompt P_{verify} requires the model to establish links only when the mentions are supported by their names, types, descriptions, and local hypergraph evidence.

The resulting mention-preserving entity clusters form $\mathcal{V}$. The original low-order relation instances and factual hyperedges remain associated with their mention-level endpoints and source chunks, while the resulting clusters provide additional cross-chunk connectivity during retrieval. Implementation details and hyperparameter settings of this module are provided in Appendix A.4.

3.3 Hybrid Structural Retrieval

In the online stage, HEAR-RAG retrieves multi-hop evidence paths from the factual evidence hypergraph and maps the selected paths back to source chunks for answer generation. The retrieval process combines dependency-aware query decomposition, entity-cluster localization, factual hyperedge selection, and iterative path expansion.

Query-guided Initialization. Given a query q , an LLM first generates a dependency-aware retrieval plan represented as a directed acyclic graph (DAG):

$$\mathcal{D}_q = \text{LLM}_{\text{plan}}(P_{\text{plan}}, q), \quad (6)$$

where $\mathcal{D}_q$ represents the dependency structure among the subqueries. The plan is then compiled into an ordered retrieval sequence:

$$(q_0, Q_q) = \text{Compile}(\mathcal{D}_q), \quad Q_q = (q_1, \dots, q_T), \quad (7)$$

where q_0 denotes the initial query and q_t denotes the query used at the t -th retrieval step. The resulting sequence respects the dependency relations specified by the query plan.

The initial query q_0 is used to identify relevant mention-preserving entity clusters and their associated factual hyperedges. These hyperedges form the initial path set P_0 . Each candidate path records the factual hyperedges and entity clusters visited during retrieval, together with the corresponding low-order relation instances and source chunks.

Hyperedge Path Expansion. For each hop query q_j , HEAR-RAG expands the current candidate paths through structurally connected entity clusters, relation instances, and factual hyperedges:

$$\tilde{P}_j = \text{Expand}(P_{j-1}, q_j). \quad (8)$$

The expanded paths are ranked and pruned using beam search:

$$P_j = \text{BeamTopK}_{K_p}(\tilde{P}_j; s_{\text{RRF}}(\cdot, q_j)), \quad (9)$$

where K_p denotes the beam size.

To combine heterogeneous ranking signals, we use reciprocal rank fusion:

$$s_{\text{RRF}}(p, q_j) = \sum_{m \in \mathcal{M}} \frac{1}{\eta + \text{rank}_m^j(p)}, \quad (10)$$

where $p \in \tilde{P}_j$, $\mathcal{M}$ denotes the set of semantic and structural ranking signals, $\text{rank}_m^j(p)$ is the rank of path p under signal m for query q_j , and η is the RRF smoothing constant. These signals consider the relevance of candidate evidence to the current hop query and its structural continuity with the existing path.

The expansion proceeds until all hop queries in the retrieval plan have been processed or no further valid expansion is available. The resulting candidate path set is used for final evidence selection.

Answer Generation. HEAR-RAG maps the selected evidence paths to their associated source chunks using the source identifiers stored in the factual hyperedges:

$$\mathcal{C}_q = \text{SelectTopK}_{K_c}(\text{Chunks}(P_q)), \quad (11)$$

where $\text{Chunks}(P_q)$ returns the source chunks associated with the selected paths, and K_c denotes the final evidence budget. Chunks are ranked according to the scores of their associated evidence paths, while duplicate chunks are removed during selection.

Finally, the original query and the selected evidence chunks are provided to the generation model:

$$a = \text{LLM}_{\text{gen}}(q, \mathcal{C}_q). \quad (12)$$

The hyperparameter settings of HEAR-RAG are provided in Appendix A.5.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on three multi-hop question answering datasets: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (2WikiMQA) (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). For each dataset, we sample 1,000 questions from the validation set and aggregate their supporting documents and distractor documents into a unified retrieval corpus, simulating a multi-hop QA setting with both relevant evidence and irrelevant information. Details of data preprocessing and the constructed factual evidence hypergraph statistics are provided in Appendix A.1.

Baselines. We compare HEAR-RAG with three categories of baselines. The first category includes zero-shot LLM reasoning methods, namely Llama-3.3-70B-Instruct (AI@Meta, 2024) and GPT-4o-mini (OpenAI et al., 2024), which evaluate the multi-hop reasoning capability of language models without external retrieval. The second category includes vanilla RAG methods, including BM25 (Robertson and Zaragoza, 2009) and BGE-M3 (Chen et al., 2024). The third category includes structure-enhanced RAG methods, including KGP (Wang et al., 2024), RAPTOR (Sarathi et al., 2024), GraphRAG (Edge et al., 2025), LightRAG (Guo et al., 2025), HippoRAG2 (Gutiérrez et al., 2025), and HyperGraphRAG (Luo et al., 2025).

Evaluation Metrics. We evaluate both answer quality and retrieval quality. For answer evaluation, we report Exact Match (EM) and Answer F1 (AF1). For retrieval evaluation, we report Precision@5 (P@5), Recall@5 (R@5), Retrieval F1@5 (RF1@5), All-Support Recall@5 (AR@5), and All-Support Recall@2 (AR@2). Here, P@5, R@5, and RF1@5 measure the quality of the top-5 retrieved evidence units, while AR@k measures the percentage of questions for which all gold supporting evidence units are retrieved within the top- k results. Detailed metric definitions are provided in Appendix A.2.

Implementation Details. We use BGE-M3 to generate semantic embeddings for queries and text chunks, with an embedding dimension of 1024. For baseline methods, we follow their original pipelines and use GPT-4o-mini for required knowledge extraction, summary generation, or query parsing steps. Unless otherwise stated, all retrieval-based methods use GPT-4o-mini as the answer generator, with temperature set to 0, and share the same answer-generation prompt. We set the final retrieval budget to $k = 5$ for all methods, following the retrieval-unit definition of each method.

We additionally evaluate generator generalization using Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct. The detailed setup and results are presented in Section 4.5. Further method-specific parameter settings are provided in Appendix A.3.

4.2 Main Results

Table 1 reports the overall performance on three multi-hop QA datasets. Overall, HEAR-RAG achieves the best performance across all datasets, consistently outperforming both vanilla RAG and structure-enhanced RAG baselines.

Zero-shot LLMs show limited performance, indicating that parametric knowledge alone is insufficient for reliable multi-hop reasoning. Vanilla RAG methods improve performance by introducing external retrieval, but still lag behind structure-enhanced methods, suggesting the importance of structured evidence organization.

Among structure-enhanced baselines, performance varies across datasets, reflecting the influence of graph construction, entity mention alignment, and path connectivity. In contrast, HEAR-RAG maintains consistent advantages, benefiting from factual hyperedges and relation-aware entity mention clustering for more effective multi-hop

evidence construction.

4.3 Retrieval Quality Evaluation

Answer-level EM and F1 do not directly reveal whether a method retrieves complete supporting evidence. We therefore additionally evaluate retrieval quality using supporting-passage precision, recall, and F1, together with all-support recall.

Table 2 shows that HEAR-RAG achieves the best P@5, R@5, RF1@5, AR@5, and AR@2 across all three datasets. In particular, the improvements in R@5 and all-support recall indicate that HEAR-RAG more effectively retrieves the complete evidence required for multi-hop reasoning. These results provide direct retrieval-side support for the proposed factual evidence structure.

4.4 Ablation Study

Table 3 reports the ablation results of key components in HEAR-RAG, including relation-aware entity mention clustering, factual hyperedges, and DAG-based query decomposition. Removing any of these components leads to performance degradation across the evaluated datasets, indicating that they contribute to multi-hop evidence organization from complementary perspectives.

Effect of Entity Mention Clustering. Removing entity mention clustering degrades performance on all datasets. This indicates that relation-aware entity mention clustering helps reduce graph fragmentation caused by semantically equivalent entity mentions with different surface forms, allowing evidence distributed across chunks to be connected through mention-preserving entity clusters. The effect is especially important for datasets where evidence is more dispersed and compositional.

Effect of Factual Hyperedges. Removing factual hyperedges also results in consistent performance degradation. This suggests that factual hyperedges provide useful higher-order evidence units beyond entity mentions and low-order relations, helping preserve the local integrity of multi-entity facts during path expansion.

Effect of DAG-based Query Decomposition. For this ablation, we remove the DAG-based query decomposition and directly use the original question as the retrieval query. Therefore, no dependency DAG is constructed and no ordered subqueries are generated. Without the DAG structure,

Table 1: Overall results on three multi-hop question answering datasets. Avg denotes the average score across datasets, and bold numbers indicate the best results in each column.

Category	Method	HotpotQA		2WikiMQA		MuSiQue		Avg	
	Model	EM	F1	EM	F1	EM	F1	EM	F1
Zero-shot LLM	GPT-4o-mini	37.5	46.5	31.5	38.9	12.2	21.2	27.1	35.5
	Llama-3.3-70B-Instruct	39.6	48.9	33.1	41.3	15.6	24.8	29.4	38.3
Vanilla RAG	BM25	37.6	45.9	24.9	29.9	11.2	18.9	24.6	31.6
	BGE-M3	43.7	54.4	36.2	45.6	15.9	26.5	31.9	42.2
Structure-RAG	KGP	45.5	57.9	35.5	49.1	19.6	29.8	33.5	45.6
	RAPTOR	46.6	58.3	36.2	48.8	22.7	32.4	35.2	46.5
	GraphRAG	42.0	55.8	31.5	45.9	18.8	27.6	30.8	43.1
	LightRAG	45.5	57.4	26.2	39.0	15.0	25.4	28.9	40.6
	HippoRAG2	48.0	59.6	42.8	52.6	24.3	36.5	38.4	49.6
	HyperGraphRAG	44.0	56.1	29.4	40.6	17.3	26.9	30.2	41.2
Ours	HEAR-RAG	50.7	61.5	45.0	55.5	28.2	39.6	41.3	52.2

Table 2: Retrieval-quality evaluation results. All values are percentages. P@5, R@5, and RF1@5 denote supporting-passage Precision@5, Recall@5, and Retrieval F1@5, respectively. AR@5 and AR@2 denote All-Support Recall@5 and All-Support Recall@2.

Method		P@5	R@5	RF1@5	AR@5	AR@2
HotpotQA	BGE-M3	29.5	73.9	42.2	48.2	21.1
	HippoRAG2	30.1	75.4	43.0	67.4	46.5
	HEAR-RAG	34.3	86.0	49.0	72.2	51.5
2WikiMQA	BGE-M3	35.1	74.1	47.6	46.5	33.9
	HippoRAG2	42.1	86.0	56.5	65.1	40.3
	HEAR-RAG	45.4	89.5	60.2	70.2	45.6
MuSiQue	BGE-M3	25.5	57.7	35.4	24.8	11.0
	HippoRAG2	27.7	63.7	38.6	41.6	19.5
	HEAR-RAG	31.0	68.0	42.6	45.6	23.7

Table 3: Ablation results in terms of answer F1 on HotpotQA, 2WikiMQA, and MuSiQue.

Variant	HotpotQA	2WikiMQA	MuSiQue
HEAR-RAG	61.5	55.5	39.6
w/o Entity Mention Clustering	59.3	53.6	36.7
w/o Factual Hyperedges	59.6	53.9	37.8
w/o Query Decomposition	58.8	53.2	36.4

the retriever cannot explicitly organize intermediate subproblems according to their reasoning dependencies. This results in a clear performance drop across all three datasets, demonstrating the importance of DAG-based query decomposition for multi-hop evidence retrieval.

Overall, entity mention clustering improves cross-chunk evidence connectivity, factual hyperedges preserve coherent higher-order factual contexts, and DAG-based query decomposition organizes reasoning dependencies for multi-step retrieval. These contributions jointly support the construction of structured multi-hop evidence paths.

4.5 Generator Generalization

To examine whether the observed improvements depend on a particular answer generator, we additionally evaluate Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct alongside GPT-4o-mini. For each method and dataset, the offline construction and retrieval outputs are kept fixed, while only the answer generator is changed. The same retrieved evidence and evaluation protocol are used across the different generators.

Table 4 compares the three answer generators under identical retrieval outputs. R@5 denotes supporting-passage Recall@5, while the remaining columns report answer F1.

As shown in Table 4, HEAR-RAG achieves the highest answer F1 under all three generators across all three datasets. Because the retrieved evidence is held fixed within each method, these results indicate that the performance improvement is not specific to GPT-4o-mini as the answer generator.

Table 4: Generator generalization results. For each method and dataset, the retrieval output is fixed while the answer generator is varied. R@5 denotes supporting-passage Recall@5. GPT-4o-mini, Llama-8B, and Llama-70B denote answer F1 obtained with GPT-4o-mini, Llama-3.1-8B-Instruct, and Llama-3.3-70B-Instruct, respectively. All values are percentages.

Method	HotpotQA				2WikiMQA				MuSiQue			
	R@5	GPT-4o-mini	Llama-8B	Llama-70B	R@5	GPT-4o-mini	Llama-8B	Llama-70B	R@5	GPT-4o-mini	Llama-8B	Llama-70B
BGE-M3	73.9	54.4	52.5	54.8	74.1	45.6	37.0	47.6	57.7	26.5	22.1	28.2
HippoRAG2	75.4	59.6	55.6	60.5	86.0	52.6	46.1	53.4	63.7	36.5	29.9	36.5
HEAR-RAG	86.0	61.5	57.8	61.7	89.5	55.5	50.4	56.3	68.0	39.6	33.1	40.2

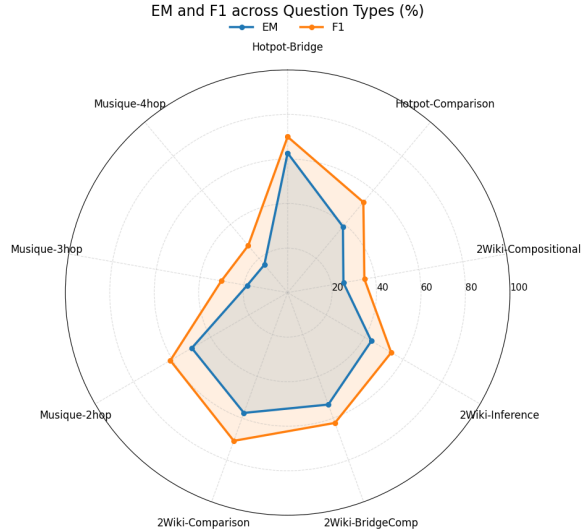


Figure 2: Performance breakdown of HEAR-RAG across different question types.

4.6 Breakdown by Question Type

Figure 2 shows the EM and F1 score of HEAR-RAG across different question types. The model achieves relatively strong performance on Hotpot-Bridge, 2WikiMQA-BridgeComp, 2WikiMQA-Comparison, and MuSiQue-2hop, where the reasoning process mainly depends on explicit entity connections or a small number of factual fragments. This indicates that the proposed factual evidence hypergraph can effectively support entity-centered evidence discovery and path construction. However, performance decreases on more complex reasoning types, especially MuSiQue-3hop and MuSiQue-4hop. As the reasoning chain becomes longer, retrieval paths are more likely to be incomplete or affected by irrelevant evidence. The relatively lower performance on 2Wiki-Compositional also suggests that precise relation selection remains challenging. These results show that HEAR-RAG is effective for structured multi-hop evidence retrieval, while long-chain reasoning and fine-grained relation filtering still leave room for improvement.

Table 5: Online retrieval efficiency and answer performance. Tok. denotes the total token usage during online retrieval and generation, Time denotes average latency in seconds, and F1 denotes the average answer F1 across three datasets.

Method	HotpotQA		2WikiMQA		MuSiQue		F1 ↑
	Tok. ↓	Time ↓	Tok. ↓	Time ↓	Tok. ↓	Time ↓	
KGP	13,462	38.85	15,953	39.74	14,313	35.14	45.6
RAPTOR	1,096	2.69	1,104	3.15	1,153	2.78	46.5
GraphRAG	4,803	2.78	4,514	2.50	4,228	2.53	43.1
LightRAG	5,540	10.45	5,666	9.78	5,456	9.11	40.6
HippoRAG2	4,155	4.47	4,532	4.30	4,267	4.21	49.6
HyperGraphRAG	6,616	2.35	6,983	2.22	6,517	2.55	41.2
HEAR-RAG	3,766	4.17	3,345	4.02	3,656	4.12	52.2

4.7 Query Efficiency Comparison

Table 5 reports the online query efficiency of different methods in terms of token usage, latency, and average F1, excluding offline index or graph construction.

The results show a trade-off between efficiency and answer quality. RAPTOR achieves the lowest token usage and latency, but its average F1 is lower than stronger structure-enhanced methods. GraphRAG and HyperGraphRAG have relatively low latency but limited answer performance, while KGP and LightRAG incur higher online costs. In comparison, HEAR-RAG achieves the highest average F1 with moderate token usage and latency. Compared with HippoRAG2, HEAR-RAG uses fewer tokens on all datasets and improves the average F1 from 49.6 to 52.2, indicating a better balance between retrieval efficiency and multi-hop answer quality.

4.8 Distractor Subset Robustness Experiment

To evaluate robustness in noisy multi-hop contexts, we conduct a distractor subset robustness experiment on 2WikiMQA using the original distractor passages provided by the dataset. For each question, all gold supporting passages are kept fixed, while different proportions of distractor passages are retained in the input context. The distractor

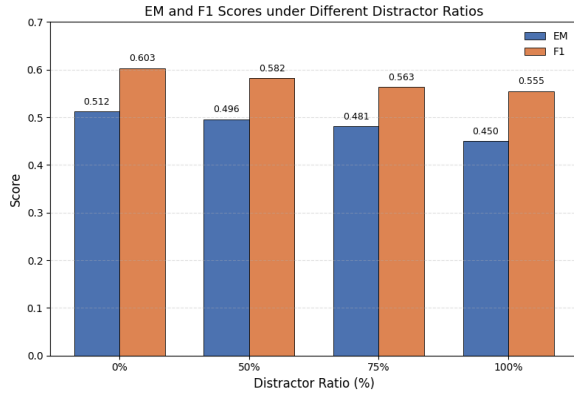


Figure 3: EM and F1 scores under different distractor ratios on 2WikiMQA.

ratios are set to 0%, 50%, 75%, and 100%.

As shown in Figure 3, both EM and F1 gradually decrease as more distractor passages are included. This suggests that distractors make evidence selection and multi-hop reasoning more challenging. Nevertheless, our method shows a relatively stable performance trend, indicating its robustness to noisy candidate contexts.

5 Conclusion

In this paper, we propose HEAR-RAG, a hypergraph-based evidence-aware retrieval framework for multi-hop question answering. HEAR-RAG organizes distributed evidence with a factual evidence hypergraph, preserving both higher-order factual contexts and fine-grained relational clues. It further uses relation-aware entity mention clustering to reduce evidence fragmentation and hybrid structural retrieval to construct generation-oriented evidence paths. Experiments on multiple multi-hop QA benchmarks demonstrate that HEAR-RAG improves both answer accuracy and structured evidence retrieval over vanilla and structure-enhanced RAG baselines.

Limitations

Although HEAR-RAG improves multi-hop evidence organization, several limitations remain. First, constructing the factual evidence hypergraph relies on LLM-based extraction and normalization, which introduces additional offline cost and may propagate extraction errors into the retrieval structure. This also makes efficient updates to evolving corpora an important direction for future work. Second, relation-aware entity mention clustering may still be affected by ambiguous or

context-dependent mentions, where over-clustering or under-clustering can influence evidence path construction. Third, further improving robustness on highly complex questions with long-range evidence dependencies remains an important direction for future work.

Acknowledgments

This work is supported by Inner Mongolia Natural Science Foundation (No.2024MS06013); National Natural Science Foundation of China(No. 62566045); Inner Mongolia Autonomous Region Science and Technology Programme Project (Nos.2026KJTW0003, 2025YFDZ0017, 2025 KYPT0041, 2025 KYPT0064, 2023YFSW0001); the special research project of Inner Mongolia First-class Discipline (YLXKZX-ND-036). We sincerely thank all reviewers for their valuable feedback.

References

- AI@Meta. 2024. [Llama 3 model card](#).
- Alessia Antelmi, Gennaro Cordasco, Mirko Polato, Vittorio Scarano, Carmine Spagnuolo, and Dingqi Yang. 2023. [A survey on hypergraph representation learning](#). *ACM Comput. Surv.*, 56(1).
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitanansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. [From local to global: A graph rag approach to query-focused summarization](#). *Preprint*, arXiv:2404.16130.
- Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. [Hypergraph neural networks](#). In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI’19/IAAI’19/EAAI’19. AAAI Press.
- Qiang Gao, Bobo Li, Zixiang Meng, Yunlong Li, Jun Zhou, Fei Li, Chong Teng, and Donghong Ji. 2024a. [Enhancing cross-document event coreference resolution by discourse structure and semantic information](#). In *Proceedings of the 2024 Joint International*

- Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5907–5921, Torino, Italia. ELRA and ICCL.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024b. [Retrieval-augmented generation for large language models: A survey](#). *Preprint*, arXiv:2312.10997.
- Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. [LightRAG: Simple and fast retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10746–10761, Suzhou, China. Association for Computational Linguistics.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. [Hipporag: Neurobiologically inspired long-term memory for large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 59532–59569. Curran Associates, Inc.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. [From RAG to memory: Non-parametric continual learning for large language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 21497–21515. PMLR.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Hao Hu, Yifan Feng, Ruoxue Li, Rundong Xue, Xingliang Hou, Zhiqiang Tian, Yue Gao, and Shaoyi Du. 2026. [Cog-rag: cognitive-inspired dual-hypergraph with theme alignment retrieval-augmented generation](#). In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’26/IAAI’26/EAAI’26*. AAAI Press.
- Zhisheng Huang, Xudong Jia, Tao Chen, and Zhongwei Zhang. 2025. [A general scalable approach for knowledge selection based on iterative hybrid encoding and re-ranking](#). *Data Intell.*, 7(1):124–142.
- Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [Active retrieval augmented generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. 2025. [Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 152206–152234. Curran Associates, Inc.
- Qingkai Min, Qipeng Guo, Xiangkun Hu, Songfang Huang, Zheng Zhang, and Yue Zhang. 2024. [Synergetic event understanding: A collaborative approach to cross-document event coreference resolution with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2985–3002, Bangkok, Thailand. Association for Computational Linguistics.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Foundations and Trends in Information Retrieval*, 4(1-2):1–174.
- Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. [RAPTOR: recursive abstractive processing for tree-organized retrieval](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. [Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. [MuSiQue: Multi-hop questions via single-hop question composition](#).

Transactions of the Association for Computational Linguistics, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. [Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Daniel Vollmers, Hamada Zahera, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2025. [Contextual augmentation for entity linking using large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8535–8545, Abu Dhabi, UAE. Association for Computational Linguistics.

Yu Wang, Nedim Lipka, Ryan A. Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024. [Knowledge graph prompting for multi-document question answering](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’24/IAAI’24/EAAI’24. AAAI Press.

Peng Xiao, Chao Liu, Wei Jia, and Lijun Dong. 2025. [Aligned-entities-based fusion embedding on hetero-field knowledge graphs](#). *Data Intell.*, 7(3):618–635.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Christophe Ye and Cassie S. Mitchell. 2025. [LLM as entity disambiguator for biomedical entity-linking](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–312, Vienna, Austria. Association for Computational Linguistics.

Linhao Ye, Lang Yu, Zhikai Lei, Qin Chen, Jie Zhou, and Liang He. 2025. [Optimizing question semantic space for dynamic retrieval-augmented multi-hop question answering](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17814–17824, Vienna, Austria. Association for Computational Linguistics.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, Yi Chang, and Xiao Huang. 2025. [A survey of graph retrieval-augmented generation for customized large language models](#). *Preprint*, arXiv:2501.13958.

Yaoze Zhang, Rong Wu, Pinlong Cai, Xiaoman Wang, Guohang Yan, Song Mao, Ding Wang, and Botian Shi. 2026. [Leanrag: Knowledge-graph-based generation with semantic aggregation and hierarchical retrieval](#). In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI 2026, Singapore, January 20-27, 2026, pages 34862–34869. AAAI Press.

Rongzhi Zhu, Xiangyu Liu, Zequn Sun, Yiwei Wang, and Wei Hu. 2025. [Mitigating lost-in-retrieval problems in retrieval augmented multi-hop question answering](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22362–22375, Vienna, Austria. Association for Computational Linguistics.

Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2024. [EfficientRAG: Efficient retriever for multi-hop question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3392–3411, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 Datasets

Our experimental evaluation is conducted on three established multi-hop question answering benchmark datasets: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). These datasets cover different types of multi-hop reasoning scenarios, including bridge reasoning, comparison reasoning, compositional reasoning, inference reasoning, and multi-step reasoning with different hop numbers. Below, we provide a concise overview of each dataset and summarize the statistics of the sampled evaluation data in Table 6.

HotpotQA is a widely used multi-hop question answering benchmark designed to evaluate cross-document reasoning ability. Each question requires the model to integrate information from multiple documents and identify supporting evidence for answer prediction. In our experiments, we select two representative question types from HotpotQA: *Bridge* and *Comparison*. Bridge questions require the model to connect intermediate entities across documents, while Comparison questions require the model to compare attributes or facts of different entities.

2WikiMultiHopQA is a multi-hop reasoning benchmark constructed from Wikipedia articles. It requires models to synthesize evidence from

multiple related documents and perform structured cross-document reasoning. We consider four reasoning types in this dataset: *Compositional*, *Inference*, *Comparison*, and *Bridge-Comparison*. These types evaluate the model’s ability to handle compositional reasoning, factual inference, comparative reasoning, and hybrid bridge-comparison reasoning.

MuSiQue is a challenging multi-hop QA benchmark that emphasizes coherent multi-step reasoning. Compared with datasets that mainly focus on two-hop reasoning, MuSiQue includes questions requiring different numbers of reasoning steps. In our experiments, we divide MuSiQue questions according to their reasoning hops, including *2-hop*, *3-hop*, and *4-hop* questions, which correspond to different levels of reasoning complexity.

As shown in Table 6, we sample 1,000 validation questions from each dataset and construct a unified retrieval corpus by aggregating the corresponding supporting and distractor passages. In our experiments, each original supporting passage or distractor passage is retained as a single source chunk and used as one retrieval unit. Therefore, the terms “supporting passage” and “source chunk” refer to the same evaluation unit throughout the experiments. The sampled data cover different reasoning types, including bridge, comparison, compositional, inference, and multi-hop reasoning with different hop numbers. The table also reports the scale of the constructed factual evidence hypergraphs, including the numbers of passages, entities, triples, and hyperedges, providing an overview of the retrieval corpora and structural indexes used in our experiments.

A.2 Evaluation Metric Definitions

For each question q , let G_q denote the set of gold supporting evidence units, and let R_q^k denote the top- k retrieved evidence units. Retrieved units are matched with gold units using their source passage or source chunk identifiers. Duplicate retrieved units are removed before evaluation.

The retrieval precision, recall, and F1 at rank k are defined as:

$$P@k = \frac{1}{N} \sum_{q=1}^N \frac{|R_q^k \cap G_q|}{|R_q^k|}, \quad (13)$$

$$R@k = \frac{1}{N} \sum_{q=1}^N \frac{|R_q^k \cap G_q|}{|G_q|}, \quad (14)$$

$$RF1@k = \frac{2 \cdot P@k \cdot R@k}{P@k + R@k}. \quad (15)$$

All-Support Recall@ k measures whether all gold supporting evidence units for a question are retrieved within the top- k results:

$$AR@k = \frac{1}{N} \sum_{q=1}^N \mathbf{1} [G_q \subseteq R_q^k], \quad (16)$$

where $\mathbf{1}[\cdot]$ equals 1 when the condition is satisfied and 0 otherwise. We report $P@5$, $R@5$, $RF1@5$, $AR@5$, and $AR@2$.

For answer evaluation, EM measures exact string match after standard answer normalization, while AF1 denotes token-level F1 between the predicted answer and the reference answer. All metrics are averaged over questions and reported as percentages.

A.3 Baseline Details

In the performance comparison experiments, we select representative baseline methods and divide them into three categories according to their retrieval and reasoning mechanisms: zero-shot LLM reasoning methods, text retrieval-augmented generation methods, and structured RAG methods. These baselines are used to comprehensively evaluate the retrieval, reasoning, and structured relation modeling capabilities of our method on multi-hop question answering tasks.

Zero-shot LLM Reasoning Methods. This category does not use external retrieval modules and directly relies on the parametric knowledge and reasoning ability of large language models to generate answers. We select GPT-4o-mini (OpenAI et al., 2024) and Llama-3.3-70B-Instruct (AI@Meta, 2024) as zero-shot reasoning baselines.

Text Retrieval-Augmented Generation Methods. This category first retrieves relevant text passages from the corpus and then uses them as context for the large language model to generate answers. We select BM25 (Robertson and Zaragoza, 2009) and BGE-M3 (Chen et al., 2024) as text RAG baselines, where BM25 represents sparse keyword-based retrieval and BGE-M3 represents dense vector-based retrieval.

Structured RAG Methods. This category introduces structured representations, such as knowledge graphs, hierarchical trees, hypergraphs, or

Table 6: Statistics of the evaluated datasets and constructed factual evidence hypergraphs.

Dataset	Type	# Questions	Total	# Passages	# Entities	# Triples	# Hyperedges
HotpotQA	Bridge	500	1000	9,459	110,982	93,099	29,621
	Comparison	500					
2WikiMQA	Compositional	250	1000	6,467	65,331	53,511	18,233
	Inference	250					
	Comparison	250					
	Bridge-Comparison	250					
MuSiQue	2-hop	800	1000	11,316	121,322	98,940	31,347
	3-hop	100					
	4-hop	100					

community structures, during indexing or retrieval to support evidence localization and multi-hop reasoning. We select KGP (Wang et al., 2024), RAPTOR (Sarathi et al., 2024), LightRAG (Guo et al., 2025), HippoRAG2 (Gutiérrez et al., 2025), GraphRAG (Edge et al., 2025), and HyperGraphRAG (Luo et al., 2025) as structured RAG baselines. Specifically, KGP constructs a knowledge graph with large language models and collects supporting passages through graph traversal; RAPTOR builds a hierarchical tree via clustering and summarization; LightRAG combines fine-grained entity-relation structures with coarse-grained thematic structures; HippoRAG2 improves multi-hop retrieval by enhancing paragraph integration, contextualization, and PageRank-based retrieval mechanisms; GraphRAG constructs community reports based on entity-relation graphs and Leiden community detection, and we adopt its Local Search mode in our experiments; HyperGraphRAG represents knowledge with hypergraph structures to capture higher-order relations and performs hybrid retrieval over entities and hyperedges.

Overall, these baselines cover three representative paradigms: retrieval-free reasoning, text RAG, and structured RAG. To preserve the evidence annotations provided by the benchmarks, we retain the original passage or paragraph boundaries as retrieval units. The remaining method-specific hyperparameter settings of the structured RAG baselines are summarized in Table 7.

A.4 Relation-aware Entity Mention Clustering Details

We first construct high-recall candidate mention groups using name- and profile-based entity representations. UMAP and HDBSCAN are applied only to oversized candidate groups to obtain lo-

Table 7: Hyperparameter settings of structured RAG baselines.

Method	Parameter	Value
RAPTOR	Summary length limit	100
	Max tokens per cluster	3500
	Target cluster size	10
	Max tree depth	5
	Retrieved vector candidates	5
KGP	Entity linking threshold	0.8
	Max summary tokens	500
	Candidate size	30
	Neighbors per seed entity	3
GraphRAG	Retrieval mode	Local
	Max summary tokens	500
	Max community size	10
	Max report input tokens	2000
	Retrieved seed entities	5
	Maximum total context budget	12000
LightRAG	Retrieval mode	Hybrid
	Max retrieved entities/relations	5
	Max summary tokens	500
	Text unit context budget	4000
	Local context budget	4000
	Global context budget	4000
HippoRAG2	Similarity threshold for synonymy edges	0.8
	Weight assigned to passage nodes	0.05
	Number of linked candidate nodes	5
	Damping factor for personalized PageRank	0.5
HyperGraphRAG	Retrieval mode	Hybrid
	Top- k vector search results	5
	Max tokens for entity summary	500
	Text unit context budget	4000
	Global context budget	4000
	Local context budget	4000

cally dense candidate blocks, while smaller groups are retained directly. An LLM then verifies these blocks using entity profiles and local factual evidence. The resulting mention-preserving entity clusters remain linked to their original relation endpoints and source chunks. The default settings are listed in Table 8.

A.5 Hyperparameter Settings and Beam Size Sensitivity

Table 9 summarizes the retrieval settings used by HEAR-RAG. The number of hop queries is deter-

Table 8: Default hyperparameter settings of relation-aware entity mention clustering.

Module	Setting	Value
UMAP	Target reduced dimension	20
	Number of nearest neighbors	5
HDBSCAN	Minimum cluster size	2
	Minimum samples	3
	Cluster selection epsilon	0.1
	Cluster selection method	eom

Table 9: Retrieval settings of HEAR-RAG.

Module	Setting	Value
Retrieval	Beam size K_p	5
	Final evidence budget K_c	5
	RRF smoothing constant η	60

mined by the dependency-aware query plan, while queries that cannot be effectively decomposed fall back to single-step retrieval.

The beam size K_p controls the number of candidate evidence paths retained after each expansion step. The final evidence budget K_c controls the number of source chunks passed to the generation model. The RRF smoothing constant η is fixed for candidate-path ranking.

We further evaluate the effect of the beam size K_p on retrieval quality. The final evidence budget is fixed to $K_c = 5$, while $K_p \in \{1, 3, 5\}$. All other retrieval settings remain unchanged. Each cell reports Recall@5 (R@5) and All-Support Recall@5 (AR@5).

The corresponding retrieval results are visualized in Figure 4.

Increasing K_p from 1 to 3 substantially improves retrieval quality. The performance then largely saturates at $K_p = 5$, which is used as the default setting in the main experiments.

A.6 Case Study

We present a case study to illustrate how HEAR-RAG connects semantically equivalent entity mentions, filters distracting evidence, and constructs a multi-hop evidence path. The question is: “Which U.S. state is home to the city where Jennifer Doudna’s affiliated institution is located?” The answer is *California*. Answering this question requires connecting three pieces of evidence: Jennifer Doudna is affiliated with the University of California, Berkeley, the institution is located in Berkeley, and Berkeley is located in California.

Dependency-aware Retrieval Process. Given the question, HEAR-RAG generates a dependency-aware retrieval plan represented as a directed acyclic graph. The plan produces the following hop queries:

- q_1 : Which institution is Jennifer Doudna affiliated with?
- q_2 : Where is the institution associated with Jennifer Doudna located?
- q_3 : Which U.S. state contains the city identified from the institution’s location evidence?

The initial query identifies the entity cluster associated with *Jennifer Doudna*. HEAR-RAG then retrieves relevant factual hyperedges and expands the candidate path according to the dependency order of the retrieval plan.

For q_1 , the first source chunk identifies Jennifer Doudna’s affiliated institution as the *University of California, Berkeley*. For q_2 , a second source chunk uses the mention *UC Berkeley* and identifies *Berkeley* as the institution’s city. The two institution mentions are assigned to the same mention-preserving entity cluster, allowing the evidence path to continue across the source chunks. For q_3 , a third source chunk provides the evidence that *Berkeley* is located in *California*.

Role of Entity Mention Clustering. This case illustrates how entity mention clustering connects different surface forms referring to the same institution. In particular, *University of California, Berkeley* and *UC Berkeley* appear in different source chunks but refer to the same institution. The clustering module assigns these mentions to the same mention-preserving entity cluster based on their names, entity profiles, and local factual evidence.

The original mentions remain associated with their respective source chunks and factual hyperedges. The entity cluster provides an additional cross-chunk connection without replacing the original mention-level information. Consequently, the first hyperedge containing *University of California, Berkeley* can be connected to the second hyperedge containing *UC Berkeley*. This connection allows the retrieval path to continue from the institutional affiliation to the city and then to the state evidence.

Without this semantic-equivalence link, the two institution mentions could be treated as independent local nodes, causing the retrieval path to stop between the corresponding source chunks. With

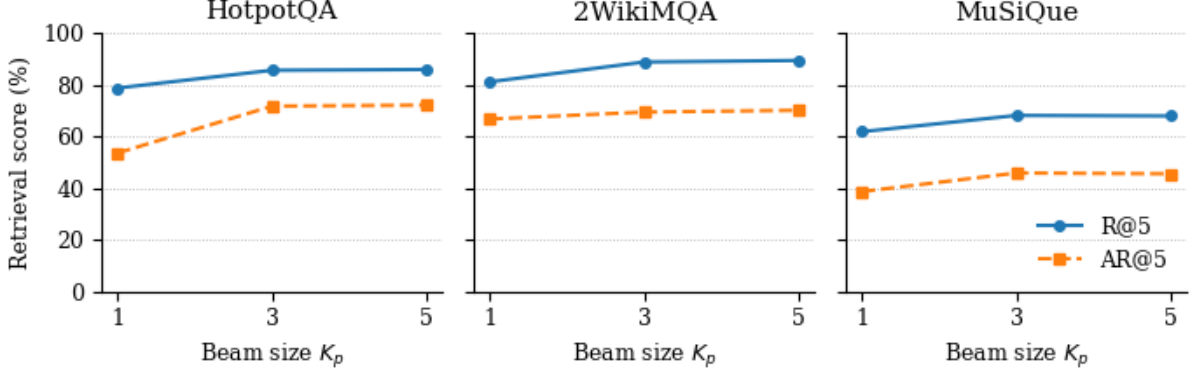


Figure 4: Sensitivity of retrieval quality to beam size K_p . The curves report Recall@5 (R@5) and All-Support Recall@5 (AR@5).

the link, HEAR-RAG can recover the complete multi-hop evidence path.

Factual Evidence Path. At the hyperedge level, the evidence path is represented as a sequence of factual hyperedges:

$$p = (h_1, h_2, h_3). \quad (17)$$

The first factual hyperedge h_1 contains the affiliation evidence between *Jennifer Doudna* and *University of California, Berkeley*. The second factual hyperedge h_2 contains the location evidence connecting *UC Berkeley* with *Berkeley*. The third factual hyperedge h_3 contains the geographic evidence connecting *Berkeley* with *California*.

Thus, HEAR-RAG organizes the retrieved evidence as a connected sequence of factual hyperedges. The path contains both local factual structures and a cross-chunk entity-cluster connection between *University of California, Berkeley* and *UC Berkeley*.

Distractor Evidence. The retrieved candidate set also contains chunks with partial lexical or semantic overlap but without the relations required to answer the question. A chunk about the *University of California, Los Angeles* shares the expression *University of California*, but it does not identify the institution affiliated with Jennifer Doudna. Another chunk about Berkeley tourism mentions *Berkeley* and *California*, but it does not provide the institutional relation connecting Jennifer Doudna to the city.

These chunks are relevant to individual entities in the question but do not form a complete evidence path. By combining semantic relevance with structural continuity through entity clusters, low-order

Prompt for Entity Extraction
PROMPTS["ner_extraction"] = ""You are an information extraction system specialized in identifying entities from text. Task: Extract all meaningful entities from the given text as exhaustively as possible. Guidelines: 1. Extract named entities, important concepts, events, organizations, locations, persons, works, dates, and domain-specific terms when they are relevant to the text. 2. Preserve the original surface form of each entity as it appears in the text. 3. Do not infer entities that are not explicitly mentioned. 4. Provide a concise type and description for each entity. 5. If no meaningful entities are found, return an empty list. Output Format: <pre>{ "entities": [{ "name": "...", "type": "...", "description": "..." }] }</pre> Text: {input_text} Output: ""

Figure 5: Prompt for entity extraction.

relation instances, and factual hyperedges, HEAR-RAG assigns higher priority to the connected path from Jennifer Doudna to California.

A.7 Prompts for Local Structure Extraction and Factual Hyperedge Construction

To improve reproducibility, we provide the prompts used for entity extraction, relation extraction, and factual hyperedge construction in Figures 5, 6, and 7. The entity extraction prompt identifies meaningful entities from each source chunk. The relation extraction prompt extracts low-order relation triples from the text. The factual hyperedge construction prompt groups related low-order relation instances and their evidence sentences into coherent factual evidence units.

Table 10: Case study illustrating entity mention clustering, factual evidence path construction, and distractor filtering.

Item	Content
Question	Which U.S. state is home to the city where Jennifer Doudna’s affiliated institution is located?
Reference Answer	California
Initial Entity	Jennifer Doudna
Dependency-aware Hop Queries	q_1 : Which institution is Jennifer Doudna affiliated with? q_2 : Where is the institution associated with Jennifer Doudna located? q_3 : Which U.S. state contains the city identified from the institution’s location evidence?
Mention-preserving Entity Cluster	v_{UCB} : { <i>University of California, Berkeley</i> in the first source chunk, <i>UC Berkeley</i> in the second source chunk }. The two mentions have different surface forms but refer to the same institution.
Factual Hyperedge h_1	Contains <i>Jennifer Doudna</i> and <i>University of California, Berkeley</i> , providing the institutional affiliation evidence.
Factual Hyperedge h_2	Contains <i>UC Berkeley</i> and <i>Berkeley</i> , providing the institution-location evidence. The path expands from h_1 to h_2 through v_{UCB} .
Factual Hyperedge h_3	Contains <i>Berkeley</i> and <i>California</i> , providing the city-state evidence. The path expands from h_2 to h_3 through the <i>Berkeley</i> entity cluster.
Selected Evidence Path	$p = (h_1, h_2, h_3)$, where adjacent factual hyperedges are connected through mention-preserving entity clusters and compatible relation endpoints.
Retrieved Distractor 1	A chunk about the <i>University of California, Los Angeles</i> shares the expression <i>University of California</i> , but it is not connected to Jennifer Doudna or the target evidence path.
Retrieved Distractor 2	A chunk about Berkeley tourism mentions <i>Berkeley</i> and <i>California</i> , but it does not provide the institutional relation required by the question.
Distractor Filtering	The distractors have partial lexical or semantic overlap but do not form a valid path from Jennifer Doudna to California. Structural ranking therefore favors the complete evidence path.
Observation	This case shows that factual hyperedges preserve local evidence, while mention-preserving entity clusters connect equivalent mentions across source chunks and enable complete multi-hop evidence retrieval.

Prompt for Relation Extraction
PROMPTS["relation_extraction"] = ""You are an information extraction system. Task: Extract meaningful relations among the entities mentioned in the text. Requirements: 1. Extract relations that are explicitly supported by the text. 2. Preserve entity names in their original surface forms whenever possible. 3. Do not introduce unsupported entities, relations, or facts. 4. Do not normalize or rewrite entity names unnecessarily. Output Format: {"triples": [["subject", "predicate", "object"]]]} Entities: {entity_list} Text: {input_text} Output: ""

Figure 6: Prompt for relation extraction.

Prompt for Factual Hyperedge Construction
PROMPTS["hyperedge_construction"] = ""You are an evidence graph construction assistant. Task: Group the provided low-order relation instances into coherent local factual evidence units. Requirements: 1. Group relation instances that share entities and occur in the same or adjacent evidence sentences. 2. Use only the provided relation instances and evidence. 3. Do not create new entities, relations, or unsupported facts. 4. Provide a concise claim for each factual evidence unit. Output Format: {"evidence_frames": [{"claim_text": "...", "member_relation_ids": ["..."]}], "unassigned_relation_ids": ["..."]} ""

Figure 7: Prompt for factual hyperedge construction.