Cog-RAG: Cognitive-Inspired Dual-Hypergraph with Theme Alignment Retrieval-Augmented Generation

Hao Hu[1], Yifan Feng[2], Ruoxue Li[3], Rundong Xue[1], Xingliang Hou[4], Zhiqiang Tian[4], Yue Gao[2], Shaoyi Du[1]*,

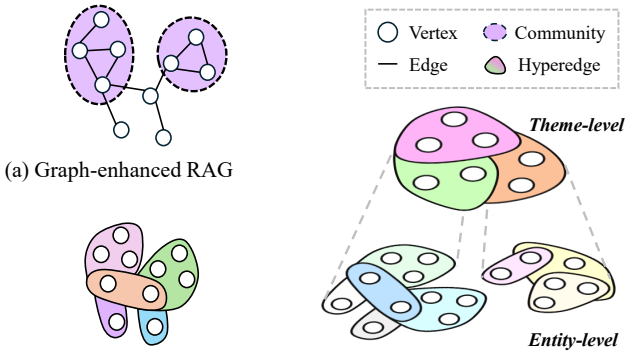
[1]State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]BNRist, THUIBCS, BLBCI, School of Software, Tsinghua University
[3]School of Artificial Intelligence, Xidian University
[4]School of Software Engineering, Xi'an Jiaotong University
huhao@stu.xjtu.edu.cn, dushaoyi@xjtu.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) enhances the response quality and domain-specific performance of large language models (LLMs) by incorporating external knowledge to combat hallucinations. In recent research, graph structures have been integrated into RAG to enhance the capture of semantic relations between entities. However, it primarily focuses on low-order pairwise entity relations, limiting the high-order associations among multiple entities. Hypergraph-enhanced approaches address this limitation by modeling multi-entity interactions via hyperedges, but they are typically constrained to inter-chunk entity-level representations, overlooking the global thematic organization and alignment across chunks. Drawing inspiration from the top-down cognitive process of human reasoning, we propose a theme-aligned dual-hypergraph RAG framework (Cog-RAG) that uses a theme hypergraph to capture inter-chunk thematic structure and an entity hypergraph to model high-order semantic relations. Furthermore, we design a cognitive-inspired two-stage retrieval strategy that first activates query-relevant thematic content from the theme hypergraph, and then guides fine-grained recall and diffusion in the entity hypergraph, achieving semantic alignment and consistent generation from global themes to local details. Our extensive experiments demonstrate that Cog-RAG significantly outperforms existing state-of-the-art baseline approaches.

Introduction

Retrieval-Augmented Generation (RAG) has recently gained increasing attention for enhancing the performance of large language models (LLMs) on knowledge-intensive tasks (Lewis et al. 2020; Gao et al. 2023; Li et al. 2024). It combats LLMs' hallucination by incorporating external knowledge, thereby enhancing response quality and reliability (Ayala and Bechard 2024; Xia et al. 2025). Moreover, it enables integration with private or domain-specific knowledge bases, thereby increasing the model's adaptability to vertical domains. With these advantages, RAG has emerged as a fundamental component in question answering, document understanding, and intelligent assistants (Fan et al. 2024; Dong et al. 2025).



(a) Graph-enhanced RAG (b) Hypergraph-enhanced RAG (c) Theme-aligned Hypergraph RAG

Figure 1: Knowledge modeling of graph, hypergraph, and our theme-enhanced RAG.

Despite the notable potential of RAG in enhancing LLMs' response quality, current methods mostly rely on a flattened chunk-based retrieval that matches queries to document chunks via vector similarity (Asai et al. 2023; Yang et al. 2024). However, this fails to capture inter-chunk dependencies and semantic hierarchies, resulting in fragmented and weakly connected retrieval content, which weakens the model's structured understanding of the entire knowledge.

To address this, recent studies have attempted to introduce graph structures into RAG framework, aiming to construct corpus-wide knowledge graphs that capture the structural semantic relations between entities (Peng et al. 2024; Zhang et al. 2025; Wang et al. 2025). For instance, GraphRAG (Edge et al. 2024) and LightRAG (Guo et al. 2024) utilize graph structures to strengthen entity-level indexing and retrieval, explicitly capturing semantic relations to improve information organization. Hyper-RAG (Feng et al. 2025a) uses hypergraphs to model complex relations between multiple entities. Nevertheless, these approaches primarily concentrate on entity-level structural modeling and lack a unified organization of knowledge themes and semantic-driven reasoning, making it difficult to support hierarchical integration

*Corresponding Authors.

of information from macro comprehension to micro details.

It is worth noting that humans tend to follow a top-down information processing path when handling complex tasks (Cheng et al. 2025; Gutiérrez et al. 2024). They begin by identifying the core themes of the problem and constructing a global semantic scaffold. Based on this, they recall and integrate relevant details to form a coherent and structured response. This "theme-driven, detail-recall" cognitive pattern reflects the inherent hierarchical organization and semantic coherence in human information processing.

Inspired by this cognitive insight, we propose a dual-hypergraph with theme alignment RAG framework (Cog-RAG). Figure 1 shows its difference with other methods in knowledge modeling. Our method leverages a dual-hypergraph structure to model the global theme structure and fine-grained high-order semantic relations. In addition, it introduces a cognitive-inspired two-stage retrieval strategy that simulates the human top-down information comprehension process, thereby enhancing the semantic consistency and structural expressiveness of generated responses. The main contributions are summarized as follows:

- We propose Cog-RAG that simulates the human top-down information processing path, enabling hierarchical generation modeling from macro-level semantic comprehension to micro-level information integration.
- We design a dual-hypergraph semantic indexing scheme to separately model global inter-chunk theme structure and intra-chunk fine-grained high-order semantic relations, overcoming the limitations of prior graph-enhanced RAG models that focus only on pairwise relations and lack unified thematic organization.
- We develop a cognitive-inspired two-stage retrieval strategy that first activates relevant context in the theme hypergraph and then triggers detail recall and diffusion in the entity hypergraph. This "theme-driven, detail-recall" process enables semantic alignment across granularity and significantly improves the coherence and quality of the response.

Related Work

RAG with Knowledge Graph

Most text-based RAG methods (Asai et al. 2023; Zhang et al. 2024; Xia et al. 2025; Yang et al. 2024) rely on a flattened paragraph structure, which makes it difficult to model semantic associations and contextual dependencies across text chunks, thereby limiting the accuracy and completeness of generated responses. To address this issue, recent studies (Sarmah et al. 2024; Peng et al. 2024) have explored knowledge graphs within the RAG framework to structurally represent entities and relations, aiming to enhance the organization and semantic expressiveness of retrieved content.

Some recent studies (Gutiérrez et al. 2024; Li and Du 2023; Cheng et al. 2025) attempt to automatically extract knowledge graph triples from the corpus and retrieve relevant subgraphs to improve content relevance and interpretability. However, these methods typically construct sparse graph structures, making it difficult to capture the full semantic space and contextual dependencies. To address the semantic sparsity issue and better model the semantic structure of documents, graph-enhanced RAG approaches (Edge et al. 2024; Guo et al. 2024; Chen et al. 2025) extract entities and their relations, and directly build document-level graph databases enriched with contextual information, thereby reducing information loss during the text-to-graph conversion process. GraphRAG and LightRAG employ LLMs to extract entities and relations from texts as vertices and edges of the graph. Nevertheless, existing methods primarily focus on low-order pairwise relations between entities, neglecting high-order group associations and global topic modeling, which limits the semantic coverage and structural expressiveness of the generated content.

Hypergraph

Hypergraphs connect multiple vertices via hyperedges, effectively modeling complex high-order relationships among entities and overcoming the limitation of conventional graphs, which support only binary relations (Gao et al. 2022; Feng et al. 2025b). These strong modeling capabilities have led to significant progress in fields such as recommender systems, social network analysis, and brain network modeling (Ji et al. 2020; Sun et al. 2023; Han et al. 2025). However, in the RAG framework, existing research is constrained to graph structures, primarily focusing on the pairwise relationships between entities. To model multiple entity group semantic associations, GraphRAG generates community reports through the semantic clustering of entities, while HiRAG (Huang et al. 2025) incorporates hierarchical graph knowledge via multi-level clustering. While effective in capturing local relationships, these methods rely on discrete category divisions and fail to model higher-order dependencies, resulting in information loss.

In contrast, hypergraphs naturally connect multiple entities through hyperedges, allowing them to interact with multiple hyperedges at once. This enables the capture of higher-order dependencies in a unified framework. It avoids the fragmentation and loss of information typical in clustering approaches and maximizes the retention of semantic information during text-to-graph conversion (Feng et al. 2025a; Luo et al. 2025). The hypergraph structure enhances semantic associations both within and across documents, thereby improving the RAG system's ability to understand context and ensure consistency in generated responses.

Preliminary

In this section, we provide a general expression for RAG and graph-enhanced RAG, referring to the definitions in (Edge et al. 2024; Guo et al. 2024).

An RAG system $\mathcal{M}$ generally includes LLM, retriever, and corpora, which can be defined as follows:

$$\mathcal{M} = (LLM, \mathcal{R}(q, \mathcal{D})). \tag{1}$$

Given a query q, the retriever $\mathcal{R}$ selects relevant contexts from the corpora $\mathcal{D}$, which are then used by the LLM to generate a response.

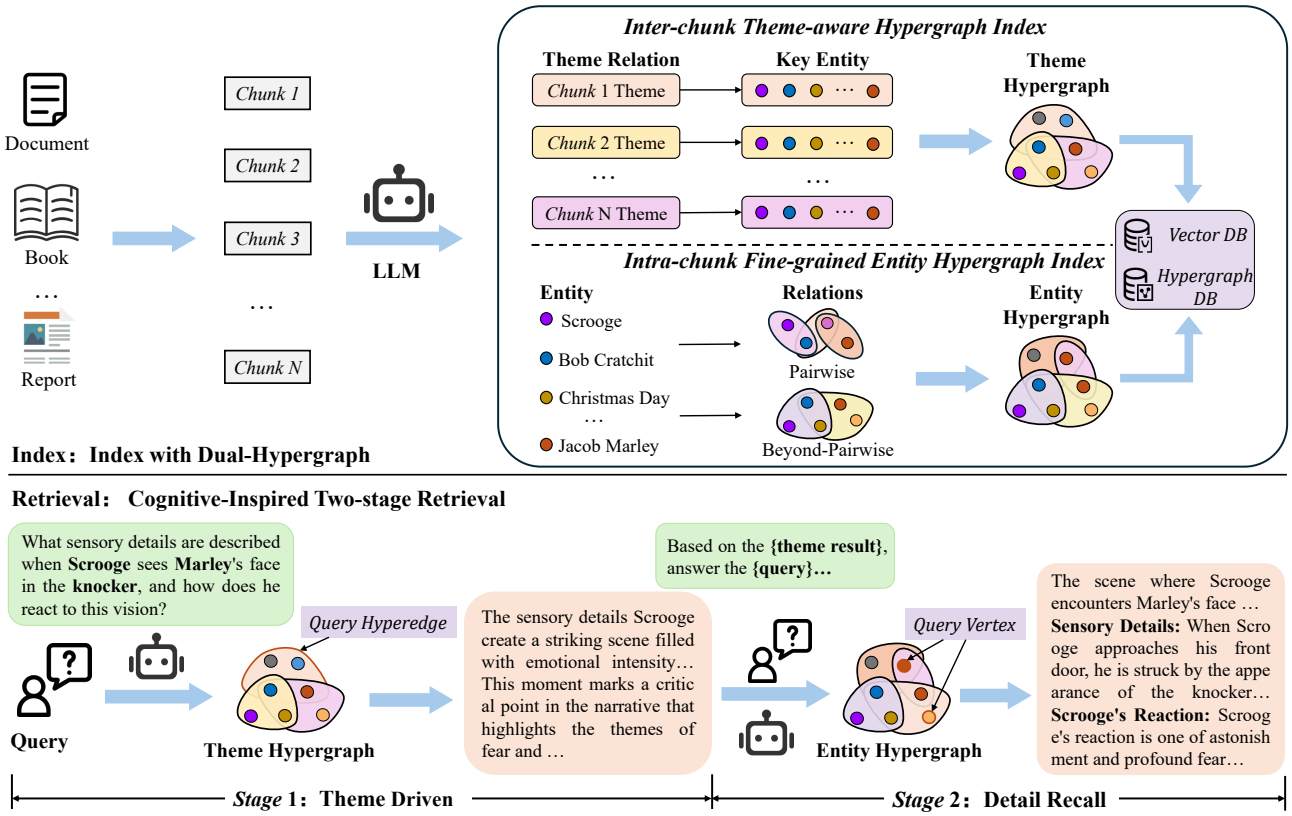


Figure 2: The overall framework of Cog-RAG.

For the graph-enhanced RAG, the corpus is organized into a graph structure, where vertices represent entities and edges represent the relations. It can be formally defined as follows:

$$\mathcal{M} = (LLM, \mathcal{R}(q, \mathcal{D} = \{\mathcal{V}, \mathcal{E}\})). \quad (2)$$

The query q retrieves relevant vertices or edges from the graph-structured corpus $\mathcal{D} = \{\mathcal{V}, \mathcal{E}\}$, enabling the LLM to respond.

Method

Overview

As illustrated in Figure 2, Cog-RAG comprises two main components: dual-hypergraph indexing and cognitive-inspired two-stage retrieval. We construct the dual-hypergraph with complementary semantic granularity: the theme hypergraph captures semantic theme associations between chunks (such as storyline, narrative outline, and summary), providing global semantic theme organization; the entity hypergraph models fine-grained high-order relations among entities (such as persons, concepts, and events), supporting local semantic relations. In the retrieval stage, mimicking the human "top-down" reasoning pattern, Cog-RAG first activates relevant themes in the theme hypergraph as global semantic anchors. Guided by these anchors, it then retrieves related entities and relations information from the entity hypergraph. The final response is generated via LLMs, utilizing theme-driven, detail-recall knowledge as evidence.

Dual-Hypergraph Indexing

To more effectively model complex high-order associations among multiple entities in corpora and avoid the information loss by graph structure, we introduce hypergraphs for modeling. The general formulation is defined as follows:

$$\mathcal{M} = (LLM, \mathcal{R}(q, \mathcal{D} = \{\mathcal{V}, \mathcal{E}_{\text{low}}, \mathcal{E}_{\text{high}}\})), \quad (3)$$

where hyperedges are used to represent relations. $\mathcal{E}_{\text{low}}$ denotes low-order pairwise entity relations, while $\mathcal{E}_{\text{high}}$ refers to high-order beyond pairwise multiple entities associations.

Theme-Aware Hypergraph Index The theme hypergraph is designed to model the semantic storyline structure of a document, establishing a narrative outline that provides cognitive guidance for subsequent detail retrieval.

Given a corpus $\mathcal{D}$, such as books, reports, or manuals, we first segment it into a set of chunks using a fixed-length sliding window with partial overlap to maintain semantic integrity, denoted as:

$$\mathcal{D} = \{D_1, D_2, ..., D_N\}, \quad (4)$$

where D_i denotes the i-th document chunk, serving as the basic unit for subsequent analysis.

Then, we perform semantic parsing on each chunk using LLMs to automatically extract its latent theme and associated key entities, thereby constructing a theme hypergraph. Specifically, we first employ predefined theme-level

extraction prompts $\mathcal{P}_{\text{ext\_theme}}$, $\mathcal{P}_{\text{ext\_key}}$ (detailed in Appendix) to guide the LLM in performing semantic parsing for each chunk D_i and outputting the corresponding theme. Then, further extract the key entities related to the theme. The calculation process is as follows:

$$\begin{cases} \mathcal{E}_{\text{theme}} = LLM(\mathcal{P}_{\text{ext\_theme}}(D_i)) \\ \mathcal{V}_{\text{key}} = LLM(\mathcal{P}_{\text{ext\_key}}(D_i, \mathcal{E}_{\text{theme}})) \end{cases} \quad \text{for} \quad D_i \in \mathcal{D}. \tag{5}$$

Based on the extracted themes and entities, we can construct the theme hypergraph $\mathcal{G}_{\text{theme}}$, denoted as:

$$\mathcal{G}_{\text{theme}} = \{\mathcal{V}_{\text{key}}, \mathcal{E}_{\text{theme}}\}, \tag{6}$$

where each hyperedge $\mathcal{E}_{\text{theme}}$ represents the narrative theme of the chunk, while the vertices $\mathcal{V}_{\text{key}}$ are the key entities.

Fine-Grained Entity Hypergraph Index After constructing the theme hypergraph, we obtain a global thematic structure among chunks. To further capture fine-grained multi-entity relations, we construct an entity hypergraph within each chunk to model high-order relations among entities, supporting subsequent fine-grained retrieval.

For each chunk D_i, we first extract entities (such as person, event, organization, etc.) and their descriptions using LLMs, which serve as the vertex set for the fine-grained entity hypergraph. Based on the semantic relations among these entities, we then construct two types of hyperedges: low-order hyperedges $\mathcal{E}_{\text{low}}$ capture basic pairwise relations, while high-order hyperedges $\mathcal{E}_{\text{high}}$ model more complex semantic associations among multiple entities, such as co-occurrence in events or causal links. The extraction process is represented as follows:

$$\begin{cases} \mathcal{V} = LLM(\mathcal{P}_{\text{ext\_entity}}(D_i)) \\ \mathcal{E}_{\text{low}} = LLM(\mathcal{P}_{\text{ext\_low}}(D_i, \mathcal{V})) \\ \mathcal{E}_{\text{high}} = LLM(\mathcal{P}_{\text{ext\_high}}(D_i, \mathcal{V})) \end{cases} \quad \text{for} \quad D_i \in \mathcal{D}, \tag{7}$$

where $\mathcal{P}_{\text{ext\_entity}}$ refers to the prompt designed for entity extraction from the text. $\mathcal{P}_{\text{ext\_low}}$ and $\mathcal{P}_{\text{ext\_high}}$ (detailed in Appendix) represent the extraction of paired and group relations from the obtained entities, respectively.

Finally, all extracted entities, along with their low-order and high-order relations, are organized into a fine-grained entity hypergraph $\mathcal{G}_{\text{entity}}$ and stored in a hypergraph database.

$$\mathcal{G}_{\text{entity}} = \{\mathcal{V}, \mathcal{E}_{\text{low}}, \mathcal{E}_{\text{high}}\}. \tag{8}$$

Cognitive-Inspired Two-Stage Retrieval

Motivated by the top-down information processing pattern observed in human memory retrieval, we design a cognitive-inspired two-stage retrieval strategy. Specifically, it first identify theme threads in the theme hypergraph related to the query. These threads then serve as cues to guide the retrieval of fine-grained information from the entity hypergraph.

For a given user query q, we first extract theme keywords (overarching concepts or themes), as follows:

$$\mathcal{X}_{\text{theme}} = LLM(\mathcal{P}_{\text{keyword}}(q)), \tag{9}$$

where $\mathcal{X}_* = \{x_1, x_2, ...\}$, $\mathcal{P}_{\text{keyword}}$ is the prompt for extracting theme keywords from the query, detailed in Appendix.

Theme-Aware Hypergraph Retrieval Subsequently, we perform structured retrieval over the hypergraph database. It is worth noting that theme keywords reflect abstract semantic relations among multiple entities and are therefore used to retrieve relevant hyperedges.

Therefore, in the first stage of retrieval, the extracted theme keywords are used to perform semantic matching within the theme hypergraph and select the top-k relevant theme hyperedges.

$$\mathcal{E}_{\text{rel}} = \{\mathcal{R}(x_i, \mathcal{E}_{\text{theme}}) | x_i \in \mathcal{X}_{\text{theme}}\}, \tag{10}$$

where $\mathcal{E}_{\text{rel}}$ represents the relevant hyperedges retrieved from the vector database. Then, we perform a diffusion process over the hypergraph database to retrieve their neighboring vertices, providing additional context awareness for the retrieved theme.

$$\mathcal{V}_{\text{dif}} = \{\mathcal{N}(e_i, \mathcal{G}_{\text{theme}}) | e_i \in \mathcal{E}_{\text{rel}}\}, \tag{11}$$

where $\mathcal{N}$ denotes the function of obtaining the corresponding neighbors from the hypergraph. $\mathcal{V}_{\text{dif}}$ is the diffusion vertices. Then, both the $\mathcal{E}_{\text{rel}}$ and $\mathcal{V}_{\text{dif}}$, along with the corresponding textual contexts, are fed into LLMs as prior knowledge to generate an initial theme-aware answer as follows:

$$\mathcal{A}_{\text{theme}} = LLM(q, \mathcal{E}_{\text{rel}}, \mathcal{V}_{\text{dif}}, \mathcal{C}_{\text{e\_rel}}, \mathcal{C}_{\text{v\_dif}}), \tag{12}$$

where $\mathcal{A}_{\text{theme}}$ denotes the output of query q after retrieving from $\mathcal{G}_{\text{theme}}$, $\mathcal{C}_*$ is the corresponding context.

Theme-aligned Entity Hypergraph Retrieval After completing the initial theme-based retrieval, we further perform fine-grained information retrieval within the entity hypergraph. Guided by the retrieved themes, this section supplements entity-level semantic details, enabling effective alignment between local information and global themes.

Based on the theme response, we further extract the aligned entity keywords (specific entities or details) from query q, which is as follows:

$$\mathcal{X}_{\text{entity}} = LLM(\mathcal{P}_{\text{align}}(q, \mathcal{A}_{\text{theme}})), \tag{13}$$

where $\mathcal{P}_{\text{align}}$ is the prompt (detailed in Appendix) used for extracting entity keywords aligned with the theme. $\mathcal{X}_{\text{entity}}$ primarily describe concrete individual information and are thus matched to vertices. The natural combination of two types of keywords and hypergraph structure enhances both retrieval specificity and structural compatibility.

Unlike the theme retrieval stage which targets hyperedges, this stage focuses on retrieving top-k vertices within the entity hypergraph by entity keywords, thereby achieving fine-grained semantic supplement and structured alignment.

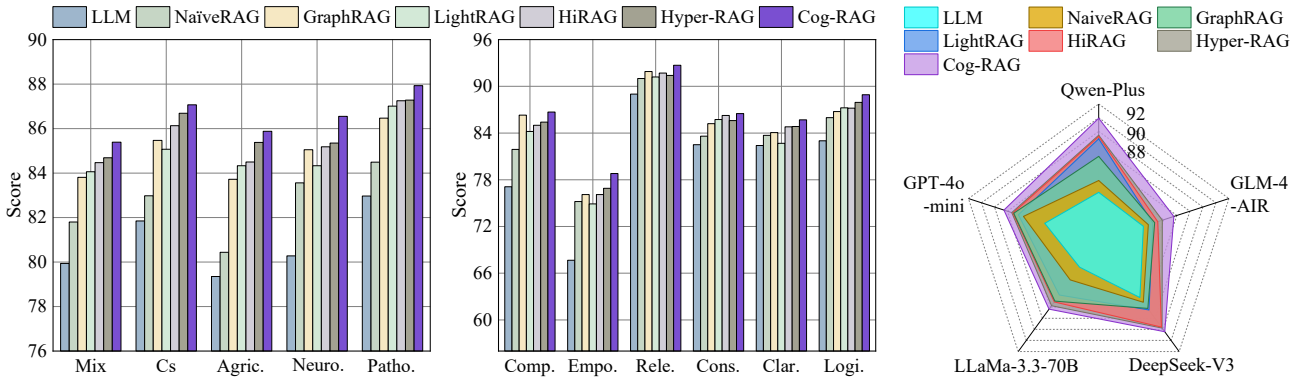
$$\mathcal{V}_{\text{rel}} = \{\mathcal{R}((x_i, \mathcal{V}_{\text{entity}}) | x_i \in \mathcal{X}_{\text{entity}}\}, \tag{14}$$

where $\mathcal{V}_{\text{rel}}$ refers the retrieved relevant entities. Then perform a hypergraph structure diffusion as follows:

$$\mathcal{E}_{\text{dif}} = \{\mathcal{N}(v_i, \mathcal{G}_{\text{entity}}) | v_i \in \mathcal{V}_{\text{rel}}\}. \tag{15}$$

Finally, the retrieved $\mathcal{V}_{\text{rel}}$, diffusion $\mathcal{E}_{\text{dif}}$, and their corresponding contexts, integrated with the previous theme information $\mathcal{A}_{\text{theme}}$, to form a structured input for LLMs to generate the final answer $\mathcal{A}$ for query q, thereby achieving a comprehensive semantic generation process from theme guidance to detailed support.

$$\mathcal{A} = LLM(q, \mathcal{A}_{\text{theme}}, \mathcal{V}_{\text{rel}}, \mathcal{E}_{\text{dif}}, \mathcal{C}_{\text{v\_rel}}, \mathcal{C}_{\text{e\_dif}}). \tag{16}$$



(a) Comparison results on different datasets (b) Comparison results on different metrics (c) Comparison results on different LLMs

Figure 3: Test results by scoring. (a) is the comparison results on five datasets; (b) is the results of the neurology dataset on six dimensions; (c) shows the evaluation results on different LLMs.

Experiments

Experimental Setup

Datasets To systematically evaluate our method across diverse application scenarios, we adopt five datasets from two benchmarks: Mix, CS, and Agriculture from the UltraDomain benchmark (Qian et al. 2024), and Neurology and Pathology from the MIRAGE benchmark (Xiong et al. 2024). UltraDomain covers typical RAG applications across different domains, while MIRAGE focuses on medical question answering and domain-specific knowledge coverage. The statistical information is given in the Appendix.

Based on domain consistency and semantic correlation within the texts, we categorize the datasets into three types to enable a comprehensive analysis of the model's adaptability: ***Cross-domain Sparse*** (Mix): Fragmented passages from unrelated domains with weak semantic coherence. ***Intra-domain Sparse*** (CS, Agriculture): Domain-specific documents with weak inter-passage context. ***Intra-domain Dense*** (Neurology, Pathology): Highly structured medical textbooks with strong semantic continuity from MIRAGE. Additionally, we follow the data processing and query procedure of LightRAG, utilizing GPT-4o to generate complex, document-related queries.

Baselines We compared our approach with the state-of-the-art and popular RAG methods. Including text-base RAG: NaiveRAG (Gao et al. 2023), graph-enhanced RAG approaches: GraphRAG (Edge et al. 2024), LightRAG (Guo et al. 2024), HiRAG (Huang et al. 2025), hypergraph-enhanced methods: Hyper-RAG (Feng et al. 2025a). The baseline details are provided in the Appendix.

Implementation Details To ensure fairness and consistency for both the baseline and proposed methods, we validate on five different LLMs for information extraction, question answering, including GPT-4o-mini (Achiam et al. 2023), Qwen-Plus (Yang et al. 2025), GLM-4-Air (GLM et al. 2024), DeepSeek-V3 (Liu et al. 2024), and LLaMa-3.3-70B (Dubey et al. 2024). The result evaluation is default on GPT-4o-mini, as well as the text-embedding-3-small embedding model for vector encoding and retrieval tasks. Unless otherwise specified, all reported results are based on GPT-4o-mini.

Evaluation Metrics Following the recent works, we adopt two evaluation strategies: Selection-based (Guo et al. 2024; Huang et al. 2025) and Score-based (Wang et al. 2024; Feng et al. 2025a), providing both relative and absolute perspectives on model performance. ***The Selection-based*** evaluation uses LLMs to reports win rates of answer quality between two methods. ***The Score-based*** evaluation employs LLMs to score responses for different methods. Both strategies assess models from six dimensions: Comprehensiveness, Empowerment, Relevance, Consistency, Clarity, and Logical. We report both per-dimension and overall average scores. Detailed evaluation descriptions are in the Appendix.

Main Results

Our primary results are presented in Table 1 and Figure 3, and more results are provided in the Appendix. Cog-RAG consistently outperforms all baselines across multiple dimensions. Additionally, we have several key insights:

1) **Knowledge graphs can enhance RAG to model a broader scope of information.** Graph-enhanced methods, represented by GraphRAG and LightRAG, demonstrate significant advantages over the conventional NaiveRAG, primarily due to the modeling of graph structures. In contrast, NaiveRAG relies solely on vector similarity and fails to account for these structured semantic relations. Hypergraph-enhanced approaches, such as Hyper-RAG and Cog-RAG, offer a more comprehensive modeling of knowledge structures that extend beyond pairwise relations, demonstrating superior potential in knowledge representation.

2) **Cog-RAG outperforms the baselines across all kinds of evaluation datasets and LLMs.** For *Selection-based results* in Table 1, we can see that in cross-domain sparse settings, both Hyper-RAG and Cog-RAG utilize hypergraphs to capture high-order relations, resulting in an average improvement of over 10.0% compared to graph-based methods. In intra-domain sparse datasets, Cog-RAG outper-

	Mix		CS		Agriculture		Neurology		Pathology	
	NaiveRAG	**Cog-RAG**	NaiveRAG	**Cog-RAG**	NaiveRAG	**Cog-RAG**	NaiveRAG	**Cog-RAG**	NaiveRAG	**Cog-RAG**
Comp.	12.0%	**88.0%**	4.0%	**96.0%**	1.0%	**99.0%**	3.0%	**97.0%**	6.0%	**94.0%**
Empo.	10.0%	**90.0%**	3.0%	**97.0%**	2.0%	**98.0%**	1.0%	**99.0%**	4.0%	**96.0%**
Rele.	27.0%	**73.0%**	18.0%	**82.0%**	6.0%	**94.0%**	11.0%	**89.0%**	8.0%	**92.0%**
Cons.	10.0%	**90.0%**	4.0%	**96.0%**	1.0%	**99.0%**	2.0%	**98.0%**	4.0%	**96.0%**
Clar.	23.0%	**77.0%**	11.0%	**89.0%**	6.0%	**94.0%**	6.0%	**94.0%**	8.0%	**92.0%**
Logi.	11.0%	**89.0%**	5.0%	**95.0%**	1.0%	**99.0%**	1.0%	**99.0%**	5.0%	**95.0%**
Overall	15.5%	**84.5%**	7.5%	**92.5%**	2.8%	**97.2%**	3.2%	**96.0%**	5.8%	**94.2%**
	GraphRAG	**Cog-RAG**	GraphRAG	**Cog-RAG**	GraphRAG	**Cog-RAG**	GraphRAG	**Cog-RAG**	GraphRAG	**Cog-RAG**
Comp.	40.0%	**60.0%**	36.0%	**64.0%**	32.0%	**68.0%**	34.0%	**66.0%**	32.0%	**68.0%**
Empo.	36.0%	**64.0%**	35.0%	**65.0%**	26.0%	**74.0%**	27.0%	**73.0%**	23.0%	**77.0%**
Rele.	45.0%	**55.0%**	39.0%	**61.0%**	35.0%	**65.0%**	37.0%	**63.0%**	31.0%	**69.0%**
Cons.	40.0%	**60.0%**	35.0%	**65.0%**	29.0%	**71.0%**	31.0%	**69.0%**	31.0%	**69.0%**
Clar.	46.0%	**54.0%**	36.0%	**64.0%**	38.0%	**62.0%**	36.0%	**64.0%**	30.0%	**70.0%**
Logi.	39.0%	**61.0%**	37.0%	**63.0%**	27.0%	**73.0%**	33.0%	**67.0%**	29.0%	**71.0%**
Overall	41.0%	**59.0%**	36.3%	**63.7%**	31.2%	**68.8%**	33.0%	**67.0%**	29.5%	**70.5%**
	LightRAG	**Cog-RAG**	LightRAG	**Cog-RAG**	LightRAG	**Cog-RAG**	LightRAG	**Cog-RAG**	LightRAG	**Cog-RAG**
Comp.	38.0%	**62.0%**	30.0%	**70.0%**	23.0%	**77.0%**	28.0%	**72.0%**	30.0%	**70.0%**
Empo.	30.0%	**70.0%**	26.0%	**74.0%**	20.0%	**80.0%**	22.0%	**78.0%**	25.0%	**75.0%**
Rele.	36.0%	**64.0%**	27.0%	**73.0%**	25.0%	**75.0%**	28.0%	**72.0%**	32.0%	**68.0%**
Cons.	34.0%	**66.0%**	29.0%	**71.0%**	21.0%	**79.0%**	25.0%	**75.0%**	27.0%	**73.0%**
Clar.	38.0%	**62.0%**	24.0%	**76.0%**	22.0%	**78.0%**	26.0%	**74.0%**	26.0%	**74.0%**
Logi.	35.0%	**65.0%**	29.0%	**71.0%**	23.0%	**77.0%**	26.0%	**74.0%**	26.0%	**74.0%**
Overall	35.2%	**64.8%**	27.5%	**72.5%**	22.3%	**77.7%**	25.8%	**74.2%**	27.7%	**72.3%**
	HiRAG	**Cog-RAG**	HiRAG	**Cog-RAG**	HiRAG	**Cog-RAG**	HiRAG	**Cog-RAG**	HiRAG	**Cog-RAG**
Comp.	44.0%	**56.0%**	40.0%	**60.0%**	41.0%	**59.0%**	35.0%	**65.0%**	40.0%	**60.0%**
Empo.	39.0%	**61.0%**	36.0%	**64.0%**	36.0%	**64.0%**	31.0%	**69.0%**	37.0%	**63.0%**
Rele.	45.0%	**55.0%**	47.0%	**53.0%**	44.0%	**56.0%**	35.0%	**65.0%**	41.0%	**59.0%**
Cons.	39.0%	**61.0%**	40.0%	**60.0%**	37.0%	**63.0%**	32.0%	**68.0%**	37.0%	**63.0%**
Clar.	45.0%	**54.0%**	50.0%	**50.0%**	44.0%	**56.0%**	31.0%	**69.0%**	40.0%	**60.0%**
Logi.	40.0%	**60.0%**	40.0%	**60.0%**	38.0%	**62.0%**	31.0%	**69.0%**	36.0%	**64.0%**
Overall	42.0%	**58.0%**	42.2%	**57.8%**	40.0%	**60.0%**	32.5%	**67.5%**	38.5%	**61.5%**
	Hyper-RAG	**Cog-RAG**	Hyper-RAG	**Cog-RAG**	Hyper-RAG	**Cog-RAG**	Hyper-RAG	**Cog-RAG**	Hyper-RAG	**Cog-RAG**
Comp.	43.0%	**57.0%**	45.0%	**55.0%**	49.0%	**51.0%**	40.0%	**60.0%**	42.0%	**58.0%**
Empo.	42.0%	**58.0%**	43.0%	**57.0%**	40.0%	**60.0%**	37.0%	**63.0%**	37.0%	**63.0%**
Rele.	**53.0%**	47.0%	47.0%	**53.0%**	45.0%	**55.0%**	46.0%	**54.0%**	37.0%	**63.0%**
Cons.	43.0%	**57.0%**	44.0%	**56.0%**	44.0%	**56.0%**	38.0%	**62.0%**	36.0%	**64.0%**
Clar.	**56.0%**	44.0%	48.0%	**52.0%**	42.0%	**58.0%**	41.0%	**59.0%**	32.0%	**68.0%**
Logi.	44.0%	**56.0%**	46.0%	**54.0%**	43.0%	**57.0%**	35.0%	**65.0%**	37.0%	**63.0%**
Overall	46.8%	**53.2%**	45.5%	**54.5%**	43.8%	**56.2%**	39.5%	**60.5%**	36.8%	**63.2%**

Table 1: Average win rates of six evaluation metrics across five datasets. The comparison is made between baselines and Cog-RAG. Among them, we refer to the six metrics as Comp. (Comprehensiveness), Empo. (Empowerment), Rele. (Relevance), Cons. (Consistency), Clar. (Clarity), and Logi. (Logical).

forms HiRAG by 15.6% and 20.0%, benefiting from multi-hyperedge propagation that uncovers latent themes and entity relations. In intra-domain dense medical corpora, Cog-RAG achieves the most significant gains. Through dual hypergraph modeling and cognitive-inspired retrieval, enhancing the alignment and aggregation of theme and fine-grained details. Compared to Hyper-RAG, it improves by 21.0% and 26.4%, respectively. For *Score-based results*, Figure 3 objectively presents the evaluation results across six dimensions on five LLMs. The results demonstrate that Cog-RAG achieves consistent and significant improvements over baseline methods in all dimensions. Moreover, when applying different LLMs for indexing and answering, it still exhibits clear advantages, highlighting its structural effectiveness.

3) **Dual-hypergraph alignment enhances knowledge representation and semantic consistency.** Inspired by human top-down cognitive pathways, Cog-RAG utilizes a dual hypergraph structure to align macro to micro knowledge. Specifically, for *Selection-based results* in Intra-domain Dense scenario, Cog-RAG improves by 35.0% and 23.0% compared to the entity-level hierarchical method HiRAG. For *Score-based results*, Cog-RAG outperforms HiRAG and

Models	Mix	CS	Neurology
	(Overall)	(Overall)	(Overall)
COG-RAG	85.39	87.07	86.55
w./o. Entity Hypergraph	76.58	84.58	84.49
w./o. Theme Hypergraph	84.82	85.88	85.41
w./o. Two-Stage Retrieval	84.88	86.41	86.18

Table 2: Ablation study on different datasets by scoring, where w./o. indicates without the part of the method.

Hyper-RAG by 1.37 and 1.20 on neurology datasets, significantly enhancing the model's ability to handle knowledge-intensive domains and ensuring semantic consistency.

Ablation Study

This section conducts ablation studies to evaluate the contribution of each core component in Cog-RAG: the theme, entity hypergraph, and two-stage retrieval strategy. Table 2 shows the results by Scoring-based evaluation. The results from three types of datasets are summarized below.

1) **Effectiveness of the Entity Hypergraph.** Removing the entity hypergraph leads to a significant decrease in performance on all three types of datasets, especially on the Mix dataset. This indicates its critical role in capturing fine-grained semantic relations within chunks. This effect is consistently observed across domains, confirming that intra-chunk entity-level modeling can enhance the representation of local knowledge.

2) **Effectiveness of the Theme Hypergraph.** Excluding the theme hypergraph causes a moderate decrease (drop 1.19 on CS and 1.14 on Neurology), highlighting its role in modeling global theme structures across chunks. The benefit is particularly noticeable in intra-domain tasks, where maintaining coherent theme alignment helps with cross-chunk reasoning and retrieval. However, on the Mix dataset, using only the theme hypergraph leads to performance degradation (from 85.39 to 76.58), indicating that in cross-domain sparse and weakly structured scenarios, theme relations may introduce noise that interferes with retrieval and answering.

3) **Effectiveness of the Two-Stage Retrieval.** Bypassing this component (by directly concatenating information from both the theme and entity hypergraphs and inputting it into LLMs) leads to consistent performance drops. This highlights the importance of the two-stage retrieval, especially in knowledge-intensive scenarios where global semantic guidance followed by entity-level refinement enables more accurate and coherent retrieval.

Hypergraph Visualization

In the Neurology dataset, Figure 4 visualized the relations of Sleep Apnea in the entity hypergraph. It illustrates the complex relations between Sleep Apnea and multiple related entities such as Chronic Lung Disease, Headache, and Respiratory Centers. It captures not only pairwise relations but also reveals multi-entity dependencies beyond pairs. As observed, the complex hypergraph among Hypertension, Sleep Apnea, Kyphoscoliosis, and Muscular Dystrophy illustrates

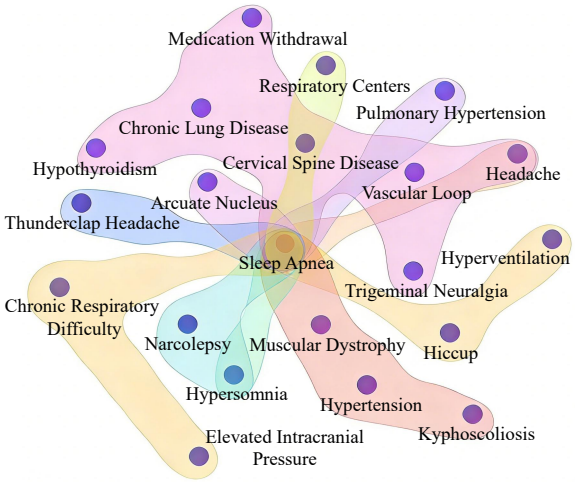


Figure 4: Entity Hypergraph Visualization.

various health risks and respiratory challenges connected to sleep quality and disorders, affecting overall wellness.

Why is Cog-RAG Effective?

Theme-Aligned vs. Graph / Hypergraph Index

Graph and hypergraph-enhanced RAG mainly focus on modeling local entity-level relations within document chunks, making them less effective for tasks that require global semantic reasoning. In contrast, Cog-RAG introduces a dual-hypergraph structure that supports alignment from global themes to fine-grained entities, leading to improved contextual grounding and response consistency. Notably, our analysis reveals that the theme hypergraph is particularly beneficial in structured, domain-specific settings, while it may introduce noise in loosely structured, open-domain scenarios. This suggests further opportunities for dynamic filtering and graph construction.

Cognitive-Inspired vs. Conventional Retrieval

Conventional RAG systems rely on single-stage retrieval, which merges all retrieved content into LLMs. This design often leads to incomplete or noisy evidence aggregation for complex knowledge-intensive tasks. The cognitive-inspired two-stage retrieval strategy enables top-down semantic alignment and aggregation, providing more accurate knowledge support and reducing redundant information.

Conclusion

Inspired by human cognitive pathways, this paper introduces Cog-RAG, which enhances LLM responses by integrating dual-hypergraph structures and a cognitive-inspired two-stage retrieval mechanism. Cog-RAG enables hierarchical knowledge modeling and semantic alignment at both macro-thematic and micro-entity levels, addressing issues of information loss and semantic gaps inherent in graph-based methods. Experimental results show that Cog-RAG significantly outperforms state-of-the-art methods across various types of datasets on knowledge-intensive tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62088102 and U24A20252, the Key Research and Development Program of Shaanxi Province of China under Grant Nos. 2024PT-ZCK-66 and 2024CY2-GJHX-48.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. arXiv:2303.08774.

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Ayala, O.; and Bechard, P. 2024. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, 228–238.

Chen, B.; Guo, Z.; Yang, Z.; Chen, Y.; Chen, J.; Liu, Z.; Shi, C.; and Yang, C. 2025. Pathrag: Pruning graph-based retrieval augmented generation with relational paths. arXiv:2502.14902.

Cheng, Y.; Zhao, Y.; Zhu, J.; Liu, Y.; Sun, X.; and Li, X. 2025. Human Cognition Inspired RAG with Knowledge Graph for Complex Problem Solving. arXiv:2503.06567.

Dong, G.; Song, X.; Zhu, Y.; Qiao, R.; Dou, Z.; and Wen, J.-R. 2025. Toward verifiable instruction-following alignment for retrieval augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23796–23804. Philadelphia, Pennsylvania, USA.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. arXiv:2407.21783.

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:2404.16130.

Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 6491–6501. New York, NY, USA.

Feng, Y.; Hu, H.; Hou, X.; Liu, S.; Ying, S.; Du, S.; Hu, H.; and Gao, Y. 2025a. Hyper-RAG: Combating LLM Hallucinations using Hypergraph-Driven Retrieval-Augmented Generation. arXiv:2504.08758.

Feng, Y.; Yang, C.; Hou, X.; Du, S.; Ying, S.; Wu, Z.; and Gao, Y. 2025b. Beyond Graphs: Can Large Language Models Comprehend Hypergraphs? In *The Thirteenth International Conference on Learning Representations*, 42468–42495. Singapore.

Gao, Y.; Feng, Y.; Ji, S.; and Ji, R. 2022. HGNN$^+$: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3181–3199.

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:2312.10997.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv:2406.12793.

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

Gutiérrez, B. J.; Shu, Y.; Gu, Y.; Yasunaga, M.; and Su, Y. 2024. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *Advances in Neural Information Processing Systems*, 59532–59569. Red Hook, NY, USA.

Han, X.; Xue, R.; Feng, J.; Feng, Y.; Du, S.; Shi, J.; and Gao, Y. 2025. Hypergraph foundation model for brain disease diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Huang, H.; Huang, Y.; Yang, J.; Pan, Z.; Chen, Y.; Ma, K.; Chen, H.; and Cheng, J. 2025. Retrieval-Augmented Generation with Hierarchical Knowledge. arXiv:2503.10150.

Ji, S.; Feng, Y.; Ji, R.; Zhao, X.; Tang, W.; and Gao, Y. 2020. Dual channel hypergraph collaborative filtering. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020–2029. New York, NY, USA.

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, R.; and Du, X. 2023. Leveraging structured information for explainable multi-hop question answering and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6779–6789. Singapore.

Li, Z.; Chen, X.; Yu, H.; Lin, H.; Lu, Y.; Tang, Q.; Huang, F.; Han, X.; Sun, L.; and Li, Y. 2024. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. arXiv:2410.08815.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. arXiv:2412.19437.

Luo, H.; Chen, G.; Zheng, Y.; Wu, X.; Guo, Y.; Lin, Q.; Feng, Y.; Kuang, Z.; Song, M.; Zhu, Y.; et al. 2025. HyperGraphRAG: Retrieval-Augmented Generation via Hypergraph-Structured Knowledge Representation. arXiv:2503.21322.

Peng, B.; Zhu, Y.; Liu, Y.; Bo, X.; Shi, H.; Hong, C.; Zhang, Y.; and Tang, S. 2024. Graph retrieval-augmented generation: A survey. arXiv:2408.08921.

Qian, H.; Zhang, P.; Liu, Z.; Mao, K.; and Dou, Z. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. arXiv:2409.05591.

Sarmah, B.; Mehta, D.; Hall, B.; Rao, R.; Patel, S.; and Pasquali, S. 2024. Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 608–616. New York, NY, USA.

Sun, X.; Cheng, H.; Liu, B.; Li, J.; Chen, H.; Xu, G.; and Yin, H. 2023. Self-supervised hypergraph representation learning for sociological analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35(11): 11860–11871.

Wang, M.; Chen, L.; Fu, C.; Liao, S.; Zhang, X.; Wu, B.; Yu, H.; Xu, N.; Zhang, L.; Luo, R.; et al. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 5627–5646. Miami, Florida, USA.

Wang, S.; Fang, Y.; Zhou, Y.; Liu, X.; and Ma, Y. 2025. ArchRAG: Attributed Community-based Hierarchical Retrieval-Augmented Generation. arXiv:2502.09891.

Xia, Y.; Zhou, J.; Shi, Z.; Chen, J.; and Huang, H. 2025. Improving retrieval augmented language model with self-reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, 25534–25542.

Xiong, G.; Jin, Q.; Lu, Z.; and Zhang, A. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, 6233–6251. Bangkok, Thailand.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. arXiv:2505.09388.

Yang, X.; Sun, K.; Xin, H.; Sun, Y.; Bhalla, N.; Chen, X.; Choudhary, S.; Gui, R.; Jiang, Z.; Jiang, Z.; et al. 2024. Crag-comprehensive rag benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 10470–10490. Red Hook, NY, USA.

Zhang, L.; Yu, Y.; Wang, K.; and Zhang, C. 2024. ARL2: Aligning Retrievers with Black-box Large Language Models via Self-guided Adaptive Relevance Labeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 3708–3719. Bangkok, Thailand.

Zhang, Q.; Chen, S.; Bei, Y.; Yuan, Z.; Zhou, H.; Hong, Z.; Dong, J.; Chen, H.; Chang, Y.; and Huang, X. 2025. A survey of graph retrieval-augmented generation for customized large language models. arXiv:2501.13958.