Relink: Constructing Query-Driven Evidence Graph On-the-Fly for GraphRAG

Manzong Huang[1], Chenyang Bu[1*], Yi He[2], Xingrui Zhuo[1], Xindong Wu [1*]

[1]Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China), School of Computer Science and Information Engineering, Hefei University of Technology, China
[2]Department of Data Science, College of William and Mary, USA
{manzonghuang, zxr}@mail.hfut.edu.cn, yihe@wm.edu, {chenyangbu, xwu}@hfut.edu.cn

Abstract

Graph-based Retrieval-Augmented Generation (GraphRAG) mitigates hallucinations in Large Language Models (LLMs) by grounding them in structured knowledge. However, current GraphRAG methods are constrained by a prevailing *build-then-reason* paradigm, which relies on a static, pre-constructed Knowledge Graph (KG). This paradigm faces two critical challenges. First, the KG's inherent incompleteness often breaks reasoning paths. Second, the graph's low signal-to-noise ratio introduces distractor facts, presenting query-relevant but misleading knowledge that disrupts the reasoning process. To address these challenges, we argue for a *reason-and-construct* paradigm and propose Relink, a framework that dynamically builds a query-specific evidence graph. To tackle incompleteness, **Relink** instantiates required facts from a latent relation pool derived from the original text corpus, repairing broken paths on the fly. To handle misleading or distractor facts, Relink employs a unified, query-aware evaluation strategy that jointly considers candidates from both the KG and latent relations, selecting those most useful for answering the query rather than relying on their pre-existence. This empowers Relink to actively discard distractor facts and construct the most faithful and precise evidence path for each query. Extensive experiments on five Open-Domain Question Answering benchmarks show that Relink achieves significant average improvements of 5.4% in EM and 5.2% in F1 over leading GraphRAG baselines, demonstrating the superiority of our proposed framework.

Code — https://github.com/DMiC-Lab-HFUT/Relink

Introduction

Despite the impressive performance in open-domain question answering (ODQA) (Patel et al. 2023; Kamalloo et al. 2023), large language models (LLMs) are prone to factual errors—known as hallucinations (Pan et al. 2024; Huang et al. 2025a). Such errors often arise from their over-reliance on internal parametric knowledge. To mitigate this, Retrieval-Augmented Generation (RAG) grounds LLMs in external knowledge (Gupta, Ranjan, and Singh 2024). GraphRAG (Pan et al. 2024; Peng et al. 2024) further advances this approach by utilizing Knowledge Graph (KG) structures to improve multi-hop query resolution through explicit relational reasoning.

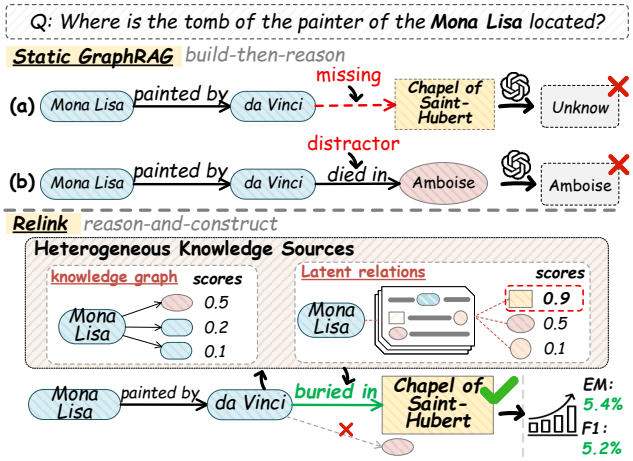

Figure 1: Static GraphRAG failures vs. Relink's Dynamic Construction. Pre-built knowledge graphs cause two critical failures in GraphRAG: (a) missing links breaking reasoning paths, and (b) distractor facts (query-relevant but goal-misaligned). In contrast, our *reason-and-construct* approach, Relink, addresses both by discarding distractor facts and dynamically instantiating missing ones from the latent relations derived from the original text corpus.

However, this progress has also exposed a fundamental constraint shared by all current GraphRAG methods: the dominant ***build-then-reason*** paradigm (Huang et al. 2025b; Sun et al. 2024). This paradigm, which relies on a pre-constructed, static KG, faces two critical challenges.

The first challenge is *the inherent incompleteness of KGs*. Static KGs inherently suffer from incomplete coverage due to evolving knowledge and extraction errors (Zhong et al. 2024; Xu et al. 2024; Zhao et al. 2022). To address this, approaches such as Knowledge Graph Completion (KGC) (Guo et al. 2023; Sun et al. 2025) and LLM-based KG construction (Mo et al. 2025; Chen et al. 2024a; Chen and Bertozzi 2023) attempt to densify the graph in advance. However, this "global completion" strategy often fails to provide the necessary "local" facts for a given query, causing the reasoning chains to remain fragile.

* Corresponding authors.

The second challenge is the *low signal-to-noise ratio*, characterized by an abundance of query-relevant yet distracting facts. General-purpose KGs contain numerous facts that may topically relate to queries but lack precise answer utility. As illustrated in Figure 1(b), the relation *died in* (versus *buried in*) exemplifies this issue: highly query-relevant yet functionally distracting. Existing methods, such as retrieval refinement and textual KG supplementation (Ma et al. 2025; Huang et al. 2025b), remain fundamentally static KG-dependent. This leaves them vulnerable to error propagation, where misleading facts are amplified during reasoning.

These challenges reveal a fundamental flaw in the prevailing ***build-then-reason*** paradigm, which relies on a one-graph-fits-all approach. Current methods are constrained by static KGs rather than adaptively serving query-specific user queries. To overcome this, we advocate a paradigm shift to ***reason-and-construct*** (Wu, Huang, and Bu 2025; Bu et al. 2025), dynamically constructing compact and query-aligned evidence graph that ensures precise reasoning path alignment.

To realize this new *reason-and-construct* paradigm, we propose **Relink**, a framework designed to address both aforementioned challenges via complementary mechanisms: (1) For KG incompleteness, Relink dynamically instantiates missing relations from the latent relations derived from the original text corpus. A high-precision KG serves as a skeletal backbone, providing a reliable foundation that inherently minimizes the presence of distractor facts. To complement its limited coverage, a high-recall latent relation pool built from entity co-occurrences in the text corpus supplies additional candidate links. This enables Relink to dynamically repair broken paths by constructing the missing facts required to answer a query. (2) For distractor noise, Relink adopts a unified evaluation strategy. At each step, a query-aware ranker assesses a pool of competing candidates, drawn from both existing KG facts and potential relations. The ranker's choice is based on a candidate's utility for answering the query, not its pre-existence. This enables Relink to actively discard misleading paths (the `"died in"` relation in Figure 1(b)) and instead construct the most relevant ones (the `"buried in"` relation in Figure 1), ensuring that the evidence graph remains precise and free of noise from the outset.

Extensive experiments on five ODQA benchmarks show that Relink outperforms leading GraphRAG baselines, achieving average gains of 5.4% in EM and 5.2% in F1. This provides compelling evidence for the superiority of the proposed paradigm.

The specific contributions of this paper are as follows:

1) We systematically analyze the prevailing *build-then-reason* paradigm, identifying its core failures in handling *KG incompleteness*, and more critically in navigating the distractor facts that stem from *low signal-to-noise ratios*.
2) Following the idea of reason-and-construct, we propose the framework Relink to dynamically construct evidence graphs through unified evaluation of explicit and latent knowledge, achieving on-the-fly path repair and distractor filtering.
3) Extensive experiments on five ODQA benchmarks show that Relink outperforms leading GraphRAG baselines by an average of 5.4% in EM and 5.2% in F1.

Related Work

GraphRAG enhances LLM reasoning by grounding it in KGs (Pan et al. 2024; Zhang et al. 2025; Zhuo et al. 2025b), proving particularly effective for complex, multi-hop question answering (Zhou et al. 2025). However, the prevailing methods are built upon a *build-then-reason* paradigm, where reasoning is performed over a pre-constructed, static KG. This approach faces two challenges: the KG's inherent incompleteness, which breaks potential reasoning paths, and the presence of misleading distractor facts, which disrupt the reasoning process (Biswas, Sack, and Alam 2024; He et al. 2024a; Min et al. 2013; Xu et al. 2024).

Existing methods attempt to address these issues within the constraints of this paradigm. One line of work focuses on alleviating KG incompleteness ahead of time through "global completion", either via KGC techniques (Guo et al. 2023; Sun et al. 2025; Zhong et al. 2024; Zhuo et al. 2024) or LLM-based KG construction (Mo et al. 2025; Chen et al. 2024a; Chen and Bertozzi 2023). However, these query-agnostic strategies densify the graph indiscriminately and often fail to satisfy the "local" facts needed for the specific query, leaving reasoning paths fragile when key links are missing. Another line of work aims to improve information relevance by refining retrieval (Guo et al. 2024; He et al. 2024b; Edge et al. 2024; Jiang et al. 2023; Chen et al. 2024b; Zhuo et al. 2025a) or supplementing evidence graphs with additional text (Huang et al. 2025b; Panda et al. 2024; Edge et al. 2024). While these approaches enhance retrieval or ranking to reduce noise, they remain fundamentally tied to the initial graph. Consequently, they struggle to establish new reasoning paths when required links are absent and are still vulnerable to being misled by distractor facts.

In contrast, Relink embodies the *reason-and-construct* paradigm, eschewing reliance on a static graph by dynamically constructing a compact, query-specific evidence graph. Rather than simply traversing a pre-built structure, Relink adopts a unified evaluation strategy at each step, assessing candidates from both the KG and a latent relation pool derived from co-occurrence patterns in the corpus. This enables Relink to instantiate essential links while actively discarding distractors. By doing so, it addresses both KG incompleteness and distractors in KG, ensuring that the resulting reasoning path is both robust and highly relevant.

Proposed Framework

We propose Relink, a framework that embodies the *reason-and-construct* paradigm, with its overall architecture illustrated in Figure 2. Rather than reasoning over a static KG, Relink dynamically constructs a compact, query-specific evidence graph. This approach addresses the challenges of KG incompleteness and distractor facts through two core designs: a Heterogeneous Knowledge Source that integrates complementary sources of candidate facts, leveraging their combined coverage to mitigate incompleteness, and a

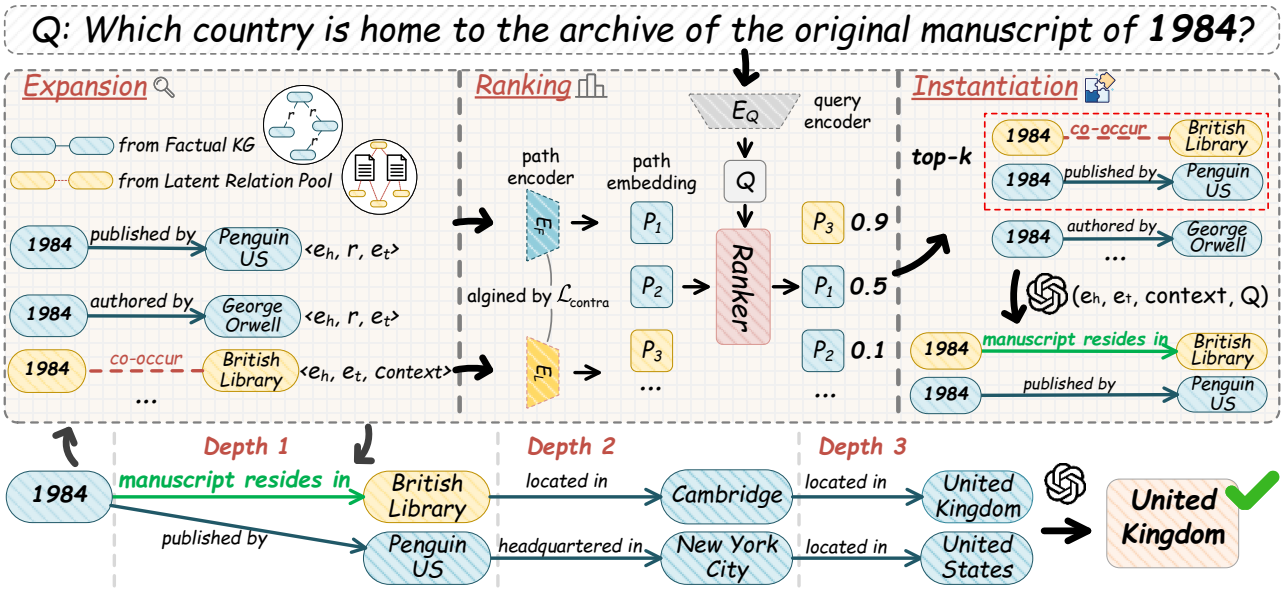

Figure 2: Relink's dynamic evidence graph construction. Relink iteratively builds reasoning paths by leveraging candidates from both the explicit KG ($\mathcal{G}_b$), and the latent co-occurrence relation pool ($\mathcal{R}_c$) derived from the corpus. Encoders E_L and E_F project these candidates into a unified semantic space where a query-driven ranker evaluates their relevance. When latent relations are selected, an LLM instantiates them into factual relations (e.g., "`manuscript resides in`") using source context, dynamically repairing missing path segments during construction.

Query-Driven Dynamic Path Exploration module that selectively navigates these sources to filter out distractors.

Heterogeneous Knowledge Source Construction

To address **KG incompleteness**, Relink's ability to repair broken paths stems from its heterogeneous knowledge source. This source is designed to balance the precision of a factual KG with the high recall of latent textual relations.

High-Precision Factual KG ($\mathcal{G}_b$): This is a standard KG, $\mathcal{G}_b = (\mathcal{E}, \mathcal{R}_b)$, where $\mathcal{E}$ is the set of entities and $\mathcal{R}_b$ is a set of high-confidence factual relations. This graph is constructed using an LLM-based extractor from the text corpus. As a result, it serves as a reliable yet inherently incomplete backbone for reasoning.

High-Recall Latent Relation Pool ($\mathcal{R}_c$): To address the limitations of incompleteness in $\mathcal{G}_b$, we introduce a high-recall pool of potential relations, $\mathcal{R}_c$, derived from textual entity co-occurrences. This pool serves as the raw material for path repair. Specifically, we identify co-occurring entity pairs (e_i, e_j) within the corpus and filter for meaningful associations using a Pointwise Mutual Information (PMI) threshold:

$$\text{PMI}(e_i, e_j) = \log \frac{p(e_i, e_j)}{p(e_i)p(e_j)} > \tau_c$$

PMI is chosen for its ability to capture non-linear co-occurrence patterns while reducing bias toward high-frequency entities. For each valid pair, we encode its context sentence c_{ij} using a pretrained language model, Encoder_L, to generate a dense representation for the latent relation, $\mathbf{r}_{ij} \in \mathbb{R}^d$. Following standard practice (Yang et al. 2024; Genest et al. 2022), we use the last hidden state of the `[MASK]` token from a formatted input s_{ij} = `[CLS]` c_{ij} `[SEP]` e_i `[MASK]` e_j `[SEP]`:

$$\mathbf{r}_{ij} = \text{Encoder}_L(s_{ij})_{[\text{MASK}]}$$

This collection, $\mathcal{R}_c$, serves as a repository of candidate connections that can be leveraged to dynamically repair and enrich reasoning paths.

Query-Driven Dynamic Path Exploration

This module lies at the heart of Relink, designed to efficiently navigate the search space and tackle the challenge posed by **distractor facts**. Given a query q, it iteratively constructs a set of evidence paths, $\mathcal{P}^*$, by intelligently selecting steps from the heterogeneous knowledge source.

Unified Semantic Space. A crucial initial step is to jointly reason over explicit triples from $\mathcal{G}_b$ and latent relations from $\mathcal{R}_c$ by projecting both into a unified semantic space $\mathbb{R}^d$. This unification is essential, as it enables a single ranker to evaluate the facts from diverse sources on a common basis.

An **explicit factual triple** $(h, p, t) \in \mathcal{G}_b$ is linearized into a sequence s = `[CLS]` $h\,p\,t$ `[SEP]`, then encoded by Encoder_F. The representation of the `[CLS]` token represents the triple: $\mathbf{v}_f = \text{Encoder}_F(s)_{[\text{CLS}]}$.

A **latent relation**'s pre-computed representation $\mathbf{r}_{ij}$ is used directly as its vector $\mathbf{v}_l$.

This unification yields a common representation $\mathbf{v}_{\text{edge}}$ for each candidate edge, enabling seamless joint ranking. The semantic coherence of this unified space is maintained during training through a contrastive alignment loss, as described in a subsequent section.

Iterative Path Expansion and Ranking. We employ a beam search algorithm, initiating from the topic entities present in the query q to systematically explore the reasoning space. At each step, we prioritize not only "relevant" but also truly "precise" knowledge from the heterogeneous source, focusing on facts that directly contribute to answering the question while effectively filtering out distractors. The search unfolds in three stages:

Candidate Expansion: For each partial path P_{k-1} in the beam, we expand a set of candidate extensions by identifying all one-hop neighbors of its last entity. These neighbors are drawn from *both* the explicit graph $\mathcal{G}_b$ and the latent pool $\mathcal{R}_c$.

Query-Driven Ranking: To efficiently identify the most promising extensions, we employ a **coarse-to-fine ranking strategy**:
Coarse Ranking: A lightweight, trainable Ranker model first scores all candidate extensions. This ranker is optimized to predict an extension's relevance to the query q. This step rapidly filters out a large number of irrelevant or noisy paths, retaining a small set of high-potential candidates.
Fine-grained Re-ranking: The top candidates from the coarse stage are then re-evaluated by an LLM. Using a structured prompt, the LLM assesses the semantic contribution of the edge toward answering the query and provides a relevance score increment ΔS.
The final average score for a path P_k is updated recursively:

$$\bar{S}(P_k|q) = \frac{(k-1)\cdot \bar{S}(P_{k-1}|q) + \Delta S(e_{\text{new}} \mid P_{k-1}, q)}{k}$$

At each step, the top-K paths with the highest average scores are retained for the next iteration.

Dynamic Instantiation: This step serves a dual purpose. Primarily, it repairs paths disrupted by KG incompleteness. At the same time, it acts as a final, precise filter against distractor facts. When a top-ranked path requires the instantiation of a latent relation $\mathbf{r}_{ij}$, we prompt the LLM with both the source context c_{ij} and the original query q. This query-aware generation encourages the LLM to produce a factual triple (h, p, t) that is closely aligned with the user's intent, rather than generating a more generic or potentially misleading relation that may also appear in the source text.

$$(h, p, t) = \text{LLM}_{\text{instantiate}}(e_i, e_j, c_{ij}, q)$$

This ensures that the newly constructed link is not only a repair, but also the most relevant possible, effectively filtering out less pertinent alternatives.

The search process terminates when either a maximum path length is reached or the LLM deems a path complete.

Evidence-Grounded Answer Generation

The output of the exploration phase is a compact, query-specific evidence graph. Each triple in this graph is linked to a single source sentence, either as provenance from $\mathcal{G}_b$ or as the specific sentence c_{ij} used for instantiated facts. This graph, along with all associated source sentences, is then provided to a generator LLM. The LLM is prompted with both the original query q and the structured evidence. By grounding the generation process in the logical reasoning structure and the precise sentence-level sources for each triple, the model produces answers that are not only accurate but also faithful and verifiable.

Joint Training Objective

The framework's learnable components, including the Ranker and the encoders (Encoder_F, Encoder_L), are optimized jointly via a multi-task objective designed to foster both ranking precision and semantic alignment.

Ranking Loss ($\mathcal{L}_{\text{rank}}$). The Ranker is trained to distinguish beneficial reasoning steps from unhelpful ones using a pairwise ranking loss. We construct a dataset of preference tuples (q, P^+, P^-), where path P^+ represents a more direct step toward the correct answer for query q compared to path P^-. The Ranker is then optimized to assign a higher score to P^+ than to P^-:

$$\mathcal{L}_{\text{rank}} = \mathbb{E}_{(q,P^+,P^-)}\left[\max(0, m - S(P^+|q) + S(P^-|q))\right]$$

where m is a margin hyperparameter. This loss directly optimizes the model's ability to navigate the search space efficiently.

Contrastive Alignment Loss ($\mathcal{L}_{\text{contra}}$). To ensure coherence within the unified semantic space, we align the representations of explicit facts ($\mathbf{v}_f$) and their corresponding latent relations ($\mathbf{v}_l$) using a contrastive objective. This alignment is crucial for enabling the Ranker to make meaningful comparisons. Specifically, we employ the InfoNCE loss to draw the embeddings of a factual triple and its corresponding latent relation closer together, while simultaneously pushing them apart from N in-batch negative samples:

$$\mathcal{L}_{\text{contra}} = -\mathbb{E}\left[\log \frac{\exp(s(\mathbf{v}_f, \mathbf{v}_l)/\tau)}{\exp(s(\mathbf{v}_f, \mathbf{v}_l)/\tau) + \sum_{j=1}^{N} \exp(s(\mathbf{v}_f, \mathbf{v}_j^-)/\tau)}\right]$$

where $s(\cdot,\cdot)$ is cosine similarity, $\mathbf{v}_j^-$ are negative samples, and τ is a temperature parameter.

Staged Optimization Strategy. We adopt a staged training procedure to stably optimize the two objectives, decoupling the challenges of learning semantic representations and ranking logic. Training alternates between two stages:

- **Ranker Training:** The encoders are frozen while the Ranker is trained for one epoch using $\mathcal{L}_{\text{rank}}$.
- **Encoder Alignment:** The Ranker is frozen while the encoders ($\text{Encoder}_F, \text{Encoder}_L$) are trained for one epoch using $\mathcal{L}_{\text{contra}}$.

This cycle repeats until convergence on the validation set, enabling each component to learn its respective task without interference.

Method Type	Method	2WikiMultiHopQA		HotpotQA		ConcurrentQA		MuSiQue-Ans		MuSiQue-Full	
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
LLM only	**Deepseek-v3**	0.312	0.365	0.294	0.402	0.114	0.170	0.066	0.167	0.078	0.164
	GPT-4o	0.292	0.358	0.330	0.424	0.086	0.158	0.106	0.213	0.106	0.214
Text-based	**Vanilla RAG**	0.264	0.295	0.376	0.494	0.31	0.389	0.118	0.178	0.08	0.145
	RAPTOR	0.467	0.548	0.472	0.631	0.400	0.476	0.258	0.371	0.188	0.286
Graph-based	**TOG**	0.328	0.373	0.241	0.292	0.166	0.185	0.095	0.127	0.053	0.070
	G-Retriever	0.362	0.455	0.369	0.493	0.162	0.209	0.166	0.271	0.092	0.169
Hybrid	**LightRAG**	0.300	0.364	0.384	0.498	0.308	0.389	0.124	0.183	0.104	0.171
	GraphRAG	0.318	0.379	0.45	0.569	0.398	0.475	0.203	0.358	0.138	0.206
	HippoRAG	0.578	0.684	0.498	0.647	0.458	0.536	0.254	0.381	0.190	0.298
Proposed	**Relink**	**0.628**	**0.722**	**0.558**	**0.704**	**0.505**	**0.596**	**0.304**	**0.413**	**0.252**	**0.370**

Table 1: Main performance comparison. Relink consistently outperforms all baseline methods across all datasets, demonstrating the effectiveness of its dynamic, query-driven path repair mechanism. Best results are in **bold**; second-best are underlined.

Experiments

To validate the effectiveness of the proposed **Relink** framework, we performed a comprehensive evaluation designed to assess its multi-hop reasoning capabilities against leading GraphRAG baseline methods.

Experimental Setup

Datasets and Evaluation Metrics. We evaluate on five standard multi-hop QA benchmarks: 2WikiMultiHopQA (Ho et al. 2020), HotpotQA (Yang et al. 2018), ConcurrentQA (Arora et al. 2023), MuSiQue-Ans, and MuSiQue-Full (Trivedi et al. 2022). Performance is measured by EM and F1, consistent with prior work (Panda et al. 2024; Huang et al. 2025b). Since LLMs often generate answers accompanied by explanations, all model outputs undergo a uniform post-processing step in which an LLM extracts the standardized answer string before scoring.

Baseline Methods. Relink was evaluated against a diverse set of representative baselines spanning the major paradigms in multi-hop QA. For consistency, all graph-based methods were implemented using the unified framework (Zhou et al. 2025). The baselines are grouped as follows: **(1) LLM Baselines:** Methods using only LLMs, including `deepseek-v3-0324` and `gpt-4o-2024-07-06`. **(2) Text-based RAG:** Approaches that retrieve chunks from corpora. This includes **Vanilla RAG** (via LangChain) and **RAPTOR** (Sarthi et al. 2024), which builds a tree-structured index. **(3) Graph-based RAG:** Methods reasoning over pre-constructed KGs, including **ToG** (Sun et al. 2024) and **G-Retriever** (He et al. 2024b). **(4) Hybrid RAG:** Leading methods combining graph and text retrieval, represented by **GraphRAG** (Edge et al. 2024), **LightRAG** (Guo et al. 2024) and **HippoRAG** (Gutierrez et al. 2024).

Implementation Details. All RAG variants, including Relink, employ `deepseek-v3-0324` as the backbone LLM to ensure comparability. Following prior works (Panda et al. 2024; Huang et al. 2025b), each method is evaluated on 500 randomly sampled questions from the test split of every dataset to reduce computational costs.

Main Performance Comparison

Table 1 presents the comparative results, which allow us to address the following Research Question (RQ):

RQ1 ***How does Relink, which follows the "reason-and-construct" paradigm, perform compared to leading GraphRAG methods on ODQA benchmarks?***

The experimental results offer empirical evidence for the proposed method. Relink consistently outperforms all baselines on five widely used ODQA benchmarks, demonstrating its effectiveness in addressing the incompleteness of knowledge graphs and the presence of distractor facts.

Compared to the LLM-only and Text-based RAG baselines, Relink delivers substantial improvements. For example, on 2WikiMultiHopQA, it achieves an EM score of 0.628, representing a 115.1% relative improvement over `GPT-4o` (0.292). It also outperforms the strong `RAPTOR` baseline, with a relative EM gain of 18.2% on HotpotQA (0.558 vs. 0.472) and 34.5% on 2WikiMultiHopQA (0.628 vs. 0.467). Experimental results show that relying solely on parameterized knowledge or unstructured text falls short for multi-hop reasoning. Multi-hop QA requires not just facts, but also clear relationships and reasoning chains between them. Relink addresses this by constructing structured evidence graphs to explicitly organize information and relationships, significantly improving the accuracy and traceability of complex reasoning.

When compared to Graph-based and Hybrid methods, Relink demonstrates clear advantages. While existing GraphRAG approaches rely on static knowledge graphs, they are often limited by the incompleteness of pre-constructed graphs and the presence of distracting facts. Relink, in contrast, adopts a dynamic ***reason-and-construct*** paradigm, which enables it to construct query-specific evidence graphs on the fly. Empirical results confirm the effectiveness of this approach. For instance, Relink surpasses

all graph-based and hybrid baselines, including the leading HippoRAG. On HotpotQA, Relink achieves a 12.0% relative EM improvement over HippoRAG. The advantage is even more pronounced on challenging datasets such as MuSiQue-Full, where Relink attains a 32.6% higher EM score (0.252 vs. 0.190). These gains highlight that dynamically constructing evidence graphs enables Relink to more effectively select relevant facts, repair incomplete reasoning chains, and filter out misleading information.

Overall, these results validate our core hypothesis that dynamic, query-aware graph construction is a more robust and effective strategy for complex multi-hop reasoning than static graph-based methods.

Ablation Study

We conducted a comprehensive ablation study to deconstruct the architecture of **Relink**. The results presented in Table 2 provide an empirical answer to RQ2.

RQ2 ***How effectively do Relink's core components address the challenges of KG incompleteness and distractor facts?***

Method	2WikiMultiHopQA		HotpotQA	
	EM	F1	EM	F1
Relink (Full Model)	**0.628**	**0.722**	**0.558**	**0.704**
w/o Explicit Graph ($\mathcal{G}_b$)	0.582	0.672	0.486	0.636
w/o Dynamic Repair ($\mathcal{R}_c$)	0.616	0.714	0.526	0.680
w/o Query-Driven Ranker	0.552	0.649	0.450	0.600
w/o $\mathcal{L}_{\text{contra}}$	0.603	0.695	0.518	0.675

Table 2: Ablation study of Relink on the 2WikiMultiHopQA and HotpotQA datasets. The removal of any component leads to a notable performance drop, highlighting their individual contributions and synergistic importance.

Heterogeneous Knowledge for Completeness. Our dual-source approach is designed to combat KG incompleteness. The results confirm its necessity. Removing the latent relation pool ($\mathcal{R}_c$) causes a 5.7% relative drop in EM on HotpotQA, while removing the explicit graph backbone ($\mathcal{G}_b$) leads to a more severe **12.9%** drop. This demonstrates that a reliable factual backbone is essential to ground the reasoning process, while the latent pool is critical for dynamically repairing incomplete reasoning paths. Neither source is sufficient on its own, and their synergy is vital for robust performance.

Query-Driven Ranker for Precision. The query-driven ranker is our primary mechanism for tackling the low signal-to-noise ratio and filtering distractor facts. We replaced the ranker component with a general method that computes cosine similarity over embeddings from OpenAI's `text-embedding-3-small` model to validate its contribution. This replacement results in the most significant performance degradation, causing a **19.4%** relative EM drop on HotpotQA. This result highlights that generic semantic similarity captures topical relevance but fails to identify facts that are truly useful for reasoning. In contrast, our ranker is explicitly trained to select evidence that supports the specific reasoning needs of each query. This ability to distinguish "useful" from merely "related" facts is essential for precise, goal-directed evidence graph construction.

Unified Alignment for Synergy. A unified semantic space is essential for reasoning over heterogeneous knowledge sources. Removing the contrastive alignment loss ($\mathcal{L}_{\text{contra}}$) leads to a 7.2% relative EM drop on HotpotQA, underscoring the challenge of integrating structured and unstructured evidence. Without alignment, the ranker cannot meaningfully compare facts from different sources, reducing its effectiveness. By aligning representations, the model enables the ranker to evaluate all evidence on a common basis, regardless of origin. This unified alignment is thus crucial for fully leveraging heterogeneous knowledge and achieving effective multi-hop reasoning.

The above ablation studies demonstrate that Relink's components work in concert to balance reasoning coverage and precision. The latent pool enhances completeness by bridging knowledge gaps. The query-driven ranker delivers precision by filtering this noise and selecting relevant facts, but it relies on a diverse candidate set. The explicit graph ensures reliability by grounding reasoning in verified knowledge. This interplay among components is key to the effectiveness of our *reason-and-construct* approach.

Robustness under Knowledge Sparsity

We assess Relink's performance under increasingly incomplete knowledge to address RQ3:

RQ3 ***How does the resilience of the "reason-and-construct" paradigm to knowledge sparsity compare against the inherent brittleness of static reasoning approaches?***

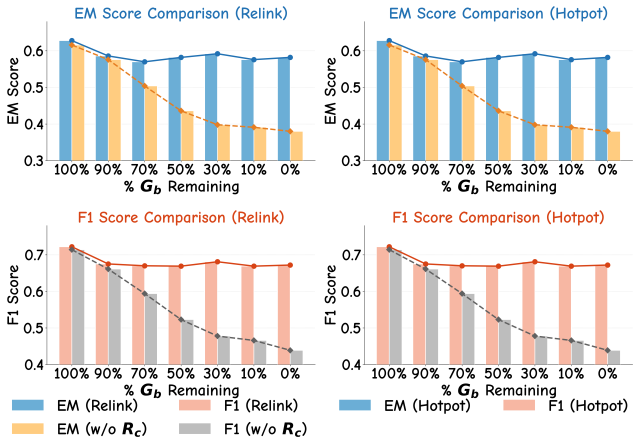

Figure 3: Performance trend as the factual graph is reduced. Relink exhibits remarkable robustness to knowledge sparsity, whereas the baseline's performance collapses.

To simulate increasing knowledge sparsity, we incrementally remove edges from the explicit graph ($\mathcal{G}_b$) and compare the full Relink with the variant without dynamic repair (*w/o* $\mathcal{R}_c$). As shown in Figure 3, Relink consistently outperforms the variant, highlighting the effectiveness of dynamic repair.

The Inherent Brittleness of Static GraphRAG Our experiments demonstrate that static reasoning is fundamentally fragile in the presence of incomplete knowledge. The *w/o* $\mathcal{R}_c$ variant, which represents this approach, relies entirely on the integrity of the existing graph structure. As shown in Figure 3, the F1 score on 2WikiMultiHopQA drops by 34.7% when 90% of the edges are removed. This significant decline highlights a key limitation that static reasoning treats paths as static, so the removal of a single crucial edge can cause the entire reasoning process for a query to fail. These results reveal a major weakness in methods that depend solely on static knowledge graphs.

Dynamic Repair Provides Remarkable Resilience. In contrast, Relink demonstrates strong adaptability through its *reason-and-construct* paradigm. Even when 90% of the explicit graph is missing, Relink maintains a high F1 score of 0.669 with only a slight decrease. This robustness is achieved by actively constructing reasoning paths rather than relying on pre-defined routes. Relink leverages the explicit graph as a set of reliable connections, but when these are unavailable, it dynamically explores latent relations from $\mathcal{R}_c$ to bridge the gaps. By flexibly building paths based on available information, Relink overcomes the limitations of static approaches and achieves greater resilience to sparsity in knowledge graphs.

These findings call for a rethinking of robustness in reasoning systems. Instead of assuming a complete knowledge base, robust systems should be built to function effectively with incomplete information. As embodied by the *reason-and-construct* paradigm, dynamic reasoning and adaptive path construction offer a practical and resilient approach for real-world reasoning tasks.

Case Study

To provide a concrete, qualitative illustration of our Relink's mechanism, we analyze its behavior on a specific multi-hop query, addressing RQ4:

RQ4 ***How does Relink's reason-and-construct process operate in practice to overcome knowledge gaps and highly relevant distractor facts where static methods fail?***

The process, depicted in Figure 4, provides a compelling illustration of the fundamental limitations inherent in the dominant **build-then-reason** paradigm. The baseline model (*w/o* $\mathcal{R}_c$) is fundamentally constrained by what the pre-existing graph **has**, not what the query **needs**. Faced with a missing `composer of` link, it is forced to piece together a low-confidence, circuitous path from available but suboptimal edges. Subsequently, when confronted with multiple facts about the correct entity, it falls into a distractor fact by selecting the highly relevant but incorrect `resides in` distractor. This failure is a direct consequence of a "one-graph-fits-all" approach, where a static, noisy graph dictates a brittle reasoning process.

In contrast, Relink exemplifies the power of our proposed *reason-and-construct* paradigm, which is guided by what the query **needs**. It treats the initial KG not as a rigid map,

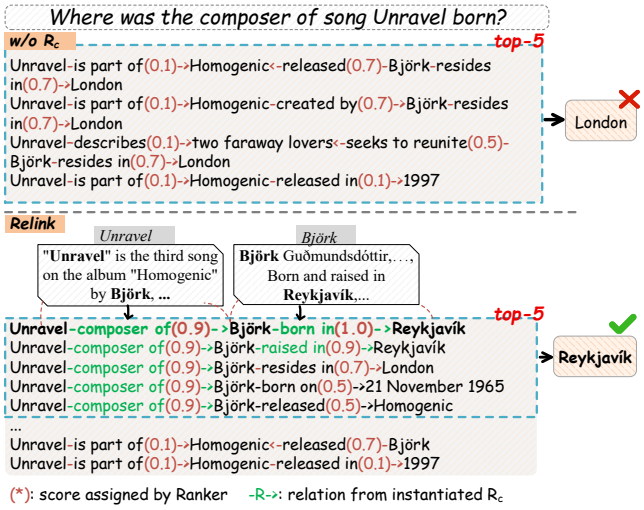

Figure 4: A case study contrasting static reasoning with Relink's dynamic approach. The static baseline (*w/o* $\mathcal{R}_c$) is misled by the highly relevant `resides in` distractor. In contrast, Relink succeeds by dynamically constructing the correct reasoning chain (`composer of` $\rightarrow$ `born in`) and using its query-driven ranker to prioritize it.

but as a high-precision scaffold to be augmented. Upon identifying the knowledge gap, it proactively uses textual evidence to instantiate the necessary `composer of` relation. This newly constructed fact is then evaluated in a unified ranking step against existing facts from the KG, including the `resides in` distractor. The query-driven ranker, seeking to fulfill the specific semantic constraint of "...was **born**?", decisively prioritizes the path containing the correct `born in` relation. This is not mere pathfinding; it is the on-the-fly construction of a compact and query-specific evidence graph.

Ultimately, this comparison highlights a necessary paradigm shift. The case study serves as a microcosm of our central thesis: static reasoning methods are fundamentally limited by their passive, retrieval-based nature, making them vulnerable to both incompleteness and noise. Relink's dynamic approach, which integrates reasoning with on-demand knowledge construction, directly addresses these core challenges. It demonstrates that for robust and precise multi-hop reasoning, the system's focus must shift from navigating what a graph **has** to intelligently constructing what a query **needs**.

Conclusion

In this work, we proposed **Relink**, a framework designed to challenge the reliance of GraphRAG on static KGs. We proposed a shift to a ***"reason-and-construct"*** paradigm, where Relink dynamically constructs query-specific reasoning paths by re-linking relations from both explicit KGs and latent text. Our experiments validate this approach: Relink significantly outperforms leading methods and maintains robust performance under the knowledge sparsity that cripples static models. This underscores the value of flexible, on-demand knowledge completion.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62120106008), the Hefei Key Generic Technology Research and Development Program (No. 2024SGJ010), the Youth Talent Support Program of the Anhui Association for Science and Technology (Grant No. RCTJ202420), the Anhui Provincial Science and Technology Fortification Plan (Grant No. 202423k09020015), and the Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China) under Grant No. BigKEOpen2025-03. The computation was completed on the HPC Platform of Hefei University of Technology. Y. He was not supported by any of these funds.

References

Arora, S.; Lewis, P. S. H.; Fan, A.; Kahn, J.; and Ré, C. 2023. Reasoning over Public and Private Data in Retrieval-Based Systems. *Trans. Assoc. Comput. Linguistics*, 11: 902–921.

Biswas, R.; Sack, H.; and Alam, M. 2024. MADLINK: Attentive multihop and entity descriptions for link prediction in knowledge graphs. *Semantic Web*, 15(1): 83–106.

Bu, C.; Chang, G.; Chen, Z.; Dang, C.; Wu, Z.; He, Y.; and Wu, X. 2025. Query-Driven Multimodal GraphRAG: Dynamic Local Knowledge Graph Construction for Online Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, 21360–21380.

Chen, B.; and Bertozzi, A. L. 2023. AutoKG: Efficient Automated Knowledge Graph Generation for Language Models. In He, J.; Palpanas, T.; Hu, X.; Cuzzocrea, A.; Dou, D.; Slezak, D.; Wang, W.; Gruca, A.; Lin, J. C.; and Agrawal, R., eds., *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, 3117–3126. IEEE.

Chen, H.; Shen, X.; Lv, Q.; Wang, J.; Ni, X.; and Ye, J. 2024a. SAC-KG: Exploiting Large Language Models as Skilled Automatic Constructors for Domain Knowledge Graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 4345–4360. Association for Computational Linguistics.

Chen, W.; Bai, T.; Su, J.; Luan, J.; Liu, W.; and Shi, C. 2024b. KG-Retriever: Efficient Knowledge Indexing for Retrieval-Augmented Large Language Models. *CoRR*, abs/2412.05547.

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; and Larson, J. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *CoRR*, abs/2404.16130.

Genest, P.; Portier, P.; Egyed-Zsigmond, E.; and Goix, L. 2022. PromptORE - A Novel Approach Towards Fully Unsupervised Relation Extraction. In *Proceedings of the 31st ACM CIKM, 2022*, 561–571. ACM.

Guo, Q.; Wang, X.; Zhu, Z.; Liu, P.; and Xu, L. 2023. A knowledge inference model for question answering on an incomplete knowledge graph. *Appl. Intell.*, 53(7): 7634–7646.

Guo, Z.; Xia, L.; Yu, Y.; Ao, T.; and Huang, C. 2024. LightRAG: Simple and Fast Retrieval-Augmented Generation. *CoRR*, abs/2410.05779.

Gupta, S.; Ranjan, R.; and Singh, S. N. 2024. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. *CoRR*, abs/2410.12837.

Gutierrez, 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 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

He, J.; Ma, M. D.; Fan, J.; Roth, D.; Wang, W.; and Ribeiro, A. 2024a. Give: Structured reasoning of large language models with knowledge graph inspired veracity extrapolation. *arXiv preprint arXiv:2410.08475*.

He, X.; Tian, Y.; Sun, Y.; Chawla, N. V.; Laurent, T.; LeCun, Y.; Bresson, X.; and Hooi, B. 2024b. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Proceedings of the Conference on NeurIPS 2024*.

Ho, X.; Duong Nguyen, A.-K.; Sugawara, S.; and Aizawa, A. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th ICCL*.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025a. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2): 42:1–42:55.

Huang, M.; Bu, C.; He, Y.; and Wu, X. 2025b. How to Mitigate Information Loss in Knowledge Graphs for GraphRAG: Leveraging Triple Context Restoration and Query-Driven Feedback. *CoRR*, abs/2501.15378.

Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, X.; and Wen, J. 2023. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on EMNLP 2023*, 9237–9251. Association for Computational Linguistics.

Kamalloo, E.; Dziri, N.; Clarke, C. L. A.; and Rafiei, D. 2023. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In *Proceedings of the 61st ACL 2023*, 5591–5606. Association for Computational Linguistics.

Ma, S.; Xu, C.; Jiang, X.; Li, M.; Qu, H.; Yang, C.; Mao, J.; and Guo, J. 2025. Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-guided Retrieval Augmented Generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Min, B.; Grishman, R.; Wan, L.; Wang, C.; and Gondek, D. 2013. Distant Supervision for Relation Extraction with an Incomplete Knowledge Base. In Vanderwende, L.; III,

H. D.; and Kirchhoff, K., eds., *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, 777–782. The Association for Computational Linguistics.

Mo, B.; Yu, K.; Kazdan, J.; Mpala, P.; Yu, L.; Cundy, C.; Kanatsoulis, C. I.; and Koyejo, S. 2025. KGGen: Extracting Knowledge Graphs from Plain Text with Language Models. *CoRR*, abs/2502.09956.

Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7): 3580–3599.

Panda, P.; Agarwal, A.; Devaguptapu, C.; Kaul, M.; and P, P. A. 2024. HOLMES: Hyper-Relational Knowledge Graphs for Multi-hop Question Answering using LLMs. In *Proceedings of 62nd Conference on ACL*, 13263–13282. ACL.

Patel, A.; Li, B.; Rasooli, M. S.; Constant, N.; Raffel, C.; and Callison-Burch, C. 2023. Bidirectional Language Models Are Also Few-shot Learners. In *Proceedings of 11th Conference on ICLR 2023*. OpenReview.net.

Peng, B.; Zhu, Y.; Liu, Y.; Bo, X.; Shi, H.; Hong, C.; Zhang, Y.; and Tang, S. 2024. Graph Retrieval-Augmented Generation: A Survey. *CoRR*, abs/2408.08921.

Sarthi, P.; Abdullah, S.; Tuli, A.; Khanna, S.; Goldie, A.; and Manning, C. D. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Sun, J.; Qian, S.; Han, Z.; Li, W.; Qian, Z.; Yang, D.; Cao, J.; and Xue, G. 2025. LKD-KGC: Domain-Specific KG Construction via LLM-driven Knowledge Dependency Parsing. *CoRR*, abs/2505.24163.

Sun, J.; Xu, C.; Tang, L.; Wang, S.; Lin, C.; Gong, Y.; Ni, L. M.; Shum, H.; and Guo, J. 2024. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. In *Proceedings of Twelfth Conference-ICLR 2024*. OpenReview.net.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Trans. Assoc. Comput. Linguistics*, 10: 539–554.

Wu, X.; Huang, M.; and Bu, C. 2025. BEKO: Bidirectional Enhancement with a Knowledge Graph for Large Language Models. *Chinese Journal of Computers*, 48(7): 1572–1588.

Xu, Y.; He, S.; Chen, J.; Wang, Z.; Song, Y.; Tong, H.; Liu, G.; Zhao, J.; and Liu, K. 2024. Generate-on-Graph: Treat LLM as both Agent and KG for Incomplete Knowledge Graph Question Answering. In *Proceedings of the 2024 Conference on EMNLP*, 18410–18430. Association for Computational Linguistics.

Yang, L.; Chen, H.; Li, Z.; Ding, X.; and Wu, X. 2024. Give us the Facts: Enhancing Large Language Models With Knowledge Graphs for Fact-Aware Language Modeling. *IEEE Trans. Knowl. Data Eng.*, 36(7): 3091–3110.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the Conference on EMNLP 2018*, 2369–2380. ACL.

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. *CoRR*, abs/2501.13958.

Zhao, F.; Li, Y.; Hou, J.; and Bai, L. 2022. Improving question answering over incomplete knowledge graphs with relation prediction. *Neural Comput. Appl.*, 34(8): 6331–6348.

Zhong, L.; Wu, J.; Li, Q.; Peng, H.; and Wu, X. 2024. A Comprehensive Survey on Automatic Knowledge Graph Construction. *ACM Comput. Surv.*, 56(4): 94:1–94:62.

Zhou, Y.; Su, Y.; Sun, Y.; Wang, S.; Wang, T.; He, R.; Zhang, Y.; Liang, S.; Liu, X.; Ma, Y.; and Fang, Y. 2025. In-depth Analysis of Graph-based RAG in a Unified Framework. *CoRR*, abs/2503.04338.

Zhuo, X.; Pan, S.; Wang, J.; Wu, G.; Zhang, Z.; Li, R.; Wei, Z.; and Wu, X. 2025a. Progressive Prefix-Memory Tuning for Complex Logical Query Answering on Knowledge Graphs. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, 3716–3724. ijcai.org.

Zhuo, X.; Wang, J.; Wu, G.; Pan, S.; and Wu, X. 2025b. Effective Instruction Parsing Plugin for Complex Logical Query Answering on Knowledge Graphs. In Long, G.; Blumestein, M.; Chang, Y.; Lewin-Eytan, L.; Huang, Z. H.; and Yom-Tov, E., eds., *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, 4780–4792. ACM.

Zhuo, X.; Wu, G.; Zhang, Z.; and Wu, X. 2024. Geometric-Contextual Mutual Infomax Path Aggregation for Relation Reasoning on Knowledge Graph. *IEEE Trans. Knowl. Data Eng.*, 36(7): 3076–3090.