KGEIR: Knowledge Graph-Enhanced Iterative Reasoning for Multi-Hop Question Answering

Tianda Sun, Dimitar Kazakov
University of York
{tianda.sun, dimitar.kazakov}@york.ac.uk

Abstract

Multi-hop question answering (MHQA) requires systems to retrieve and connect information across multiple documents, a task where large language models often struggle. We introduce Knowledge Graph-Enhanced Iterative Reasoning (KGEIR), a framework that dynamically constructs and refines knowledge graphs during question answering to enhance multi-hop reasoning. KGEIR identifies key entities from questions, builds an initial graph from retrieved paragraphs, reasons over this structure, identifies information gaps, and iteratively retrieves additional context to refine the graph until sufficient information is gathered. Evaluations on HotpotQA, 2WikiMultiHopQA, and MuSiQue benchmarks show competitive or superior performance to state-of-the-art methods. Ablation studies confirm that structured knowledge representations significantly outperform traditional prompting approaches like Chain-of-Thought and Tree-of-Thought. KGEIR's ability to explicitly model entity relationships while addressing information gaps through targeted retrieval offers a promising direction for integrating symbolic and neural approaches to complex reasoning tasks. Details of the project and the code are published at https://github.com/TiandaSun/KGEIR

1 Introduction

Multi-hop question answering (MHQA) presents a significant challenge in natural language processing, requiring systems to retrieve and connect information from multiple documents to answer complex questions (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). Unlike traditional question answering, which typically relies on information from a single passage, MHQA demands reasoning across disparate pieces of information, making it a more accurate reflection of human information-seeking behaviour (Chen et al., 2017). Despite recent advances in large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023), their ability to perform structured reasoning over multiple sources remains a challenging area, particularly when evidence must be gathered from diverse documents without explicit connections (Qi et al., 2019).

Existing approaches to MHQA typically follow a retrieve-then-read paradigm (Lewis et al., 2020; Karpukhin et al., 2020), where relevant documents are first retrieved based on the question, followed by a reading comprehension step to extract the answer. However, this sequential process often struggles with complex questions requiring multi-step reasoning, as the initial retrieval may fail to capture all necessary documents when relationships between different pieces of evidence are not explicitly considered [11]. Furthermore, most systems lack an effective mechanism to identify and address information gaps through iterative refinement (Trivedi et al., 2023). The increasing availability of powerful LLMs has opened new possibilities for MHQA, as these models demonstrate impressive reasoning capabilities (Wei et al., 2023; Wang et al., 2023). However, their application in multi-hop settings is often limited by several factors: (1) the inability to understand relationships between entities across different passages (Han et al., 2025), (2) the lack of structured representation of knowledge (Sun et al., 2024; Edge et al., 2025), and (3) the absence of systematic processes to identify and fill information gaps (He et al., 2024).

To address these limitations, we propose a novel Knowledge Graph-Enhanced Iterative Reasoning (KGEIR) framework for multi-hop question answering. Our approach combines the reasoning capabilities of LLMs with the structured representation of knowledge graphs, enabling more effective multi-hop reasoning through explicit modelling of entity relationships across documents. The key insight of our approach is that dynamically constructing and refining a knowledge graph during

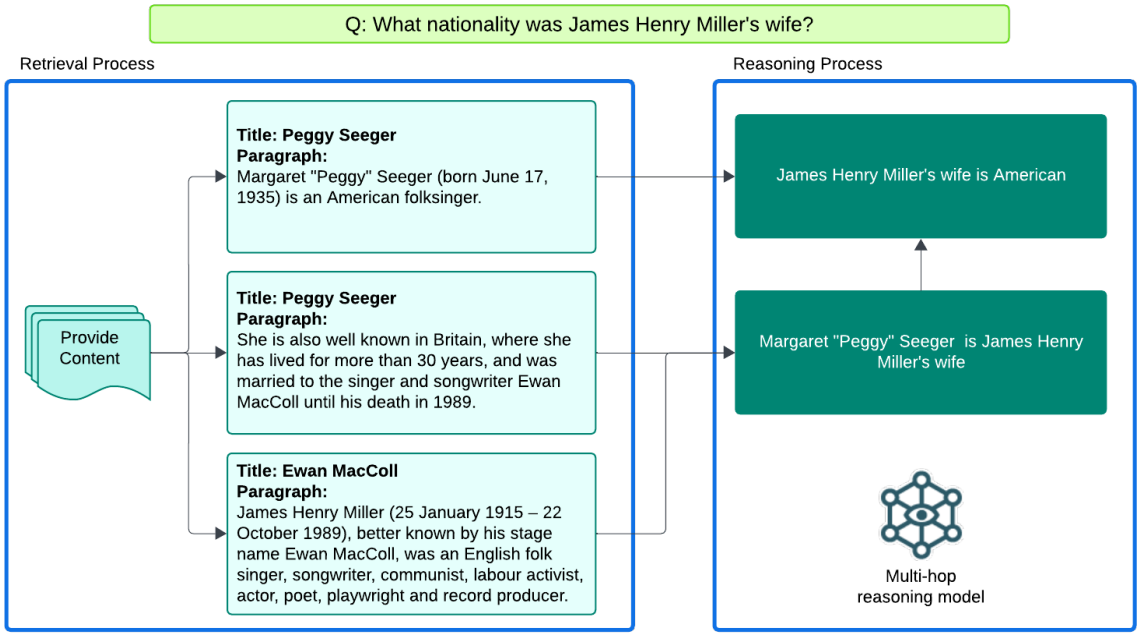

Figure 1: A common workflow on the MHQA task with an example from HotpotQA dataset. A regular MHQA question cannot get the answer from one single document but needs to retrieval multiply paragraphs from different documents. In here, the model firstly needs to retrieve the relevant paragraph across 2 different documents and identify that 'James Henry Miller' and 'Ewan MacColl' is one and the same person. Then it can make the connection between the fact that Peggy Seeger is his wife and the knowledge about her nationality (American).

the question-answering process provides an effective scaffold for reasoning, while also identifying information gaps that can guide targeted retrieval of additional context.

Our KGEIR framework operates through an iterative process: (1) initial retrieval of relevant paragraphs based on entities extracted from the question, (2) dynamic construction of a knowledge graph from retrieved paragraphs, (3) reasoning over the knowledge graph to attempt answering the question, (4) identification of information gaps in the knowledge graph, (5) targeted retrieval of additional paragraphs to fill these gaps, and (6) refinement of the knowledge graph and reasoning process. This iterative approach continues until sufficient information is gathered to answer the question confidently or a maximum number of iterations is reached.

We evaluate our approach on multiple multi-hop QA datasets, including HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022), demonstrating that KGEIR achieves significant improvements over strong baselines. Our analysis shows that the knowledge graph structure effectively guides the reasoning process of LLMs, while the iterative refinement process substantially improves answer accuracy by addressing information gaps identified during reasoning. The contributions of this paper are threefold:

1. A novel framework that leverages knowledge graphs to enhance multi-hop reasoning capabilities of large language models.

2. An iterative information-seeking approach that identifies and addresses knowledge gaps through targeted retrieval.

3. A comprehensive evaluation demonstrating the effectiveness of our approach on challenging multi-hop QA benchmarks.

Our results suggest that structuring information as explicit entity-relation graphs significantly enhances the multi-hop reasoning capabilities of LLMs, potentially opening new avenues for combining symbolic and neural approaches to complex question answering.

2 Related Works

This section examines recent advances in multi-hop question answering (MHQA), organised into complementary research directions that inform our KGEIR framework. By analysing the strengths and limitations of existing approaches, we demonstrate the need for our integrated framework.

2.1 Retrieval and Knowledge Structure Approaches

Recent retrieval methods for MHQA have progressed beyond simple matching to incorporate logical relevance and multi-hop connections. While dense retrievers like BGE (Xiao et al., 2024) perform well on single-hop tasks, they often struggle with capturing bridging information needed for complex reasoning. HopRAG (Liu et al., 2025) represents a significant advancement by introducing a logic-aware retrieval mechanism that connects passages through pseudo-queries and employs a retrieve-reason-prune paradigm. Their work demonstrated that indirectly relevant passages can serve as stepping stones to reach relevant ones, achieving notable results across multiple datasets. The approach, however, focuses primarily on enhancing retrieval rather than constructing explicit knowledge representations for reasoning.

Astute RAG (Wang et al., 2024) addresses imperfect retrieval by developing mechanisms to overcome knowledge conflicts and reasoning failures. They revealed that approximately 70% of retrieved passages do not directly contain true answers, highlighting the limitations of pure similarity-based retrieval. Similarly, BRIGHT (Su et al., 2025) demonstrates through their benchmark that even state-of-the-art retrievers struggle with multi-step reasoning tasks.

Knowledge structure approaches have emerged to provide explicit representations of entity relationships. G-Retriever (He et al., 2024) introduced a retrieval-augmented generation framework that enhances retrieval quality by leveraging graph structures to identify relevant information through entity-relation patterns. GraphRAG (Edge et al., 2025) builds hierarchical graph indices with knowledge graph construction and recursive summarisation, demonstrating the value of graph structures for organising complex information. Extract, Define, Canonicalise (Gutiérrez et al., 2025) presents an LLM-based framework for knowledge graph construction that systematically extracts entities and relations from text without extensive training or predefined schemata.

A key limitation across these approaches is their reliance on static construction processes and lack of explicit mechanisms to identify information gaps and iteratively refine knowledge representations, which our KGEIR framework specifically addresses.

2.2 Reasoning and LLM-Based Approaches

Recent reasoning approaches have increasingly leveraged structured representations to guide LLMs through complex multi-hop questions. Graph-based reasoning methods have shown particular promise in organising the reasoning process. Graph Elicitation (Park et al., 2024) decomposes multi-hop questions into sub-questions to form a graph and guides LLMs to answer based on the chronological order of the graph. Structure-Guided Prompting (Cheng et al., 2024) instructs LLMs in multi-step reasoning by exploring graph structures extracted from text. While effective, these approaches typically construct graphs as static scaffolds rather than dynamic structures that evolve through iterative refinement.

Graph Chain-of-Thought (Jin et al., 2024) augments LLMs by incorporating reasoning on graphs into the generation process, demonstrating that graph structures can significantly enhance LLMs' reasoning capabilities on tasks requiring structured knowledge. Reasoning with Graphs (Han et al., 2025) most closely aligns with our approach by structuring implicit knowledge into explicit graphs through multiple rounds of verification. Their results show significant improvements across logical reasoning and multi-hop question answering tasks, though their approach does not incorporate an iterative retrieval mechanism to address information gaps identified during reasoning.

The reasoning capabilities of LLMs have been extensively studied, revealing both strengths and limitations. Yang et al. (Yang et al., 2024) found that while models can connect information across sources, they benefit significantly from explicit guidance in complex scenarios, particularly as reasoning hops increase. Huang et al. (Huang et al., 2024) demonstrated that even advanced LLMs struggle to identify and correct errors in their reasoning without external guidance, underscoring the importance of providing explicit structures to guide the reasoning process.

Various approaches have been proposed to enhance LLMs' reasoning. Self-RAG (Asai et al., 2023) introduced a framework for retrieval, generation, and critique through self-reflection, while REFEED (Yu et al., 2023) employs a multi-round retrieval-generation framework using feedback to refine retrieval steps. SAFE-RAG (Liang et al., 2025) highlighted the importance of reliable reasoning over retrieved information, showing that without proper verification mechanisms, LLMs can produce inconsistent responses. However, these approaches typically lack explicit mechanisms to identify specific information gaps or leverage structured representations for reasoning.

2.3 Iterative Refinement Approaches

Iterative approaches to information retrieval and reasoning have gained significant traction, addressing multi-hop question answering challenges through progressive refinement. When compared with tree-structured RAG approaches like RAPTOR (Sarthi et al., 2024) and SiReRAG (Zhang et al., 2025), graph-structured approaches, such as HopRAG (Liu et al., 2025) demonstrate superior performance by enabling flexible logical modelling, cross-document organisation, and efficient construction.

The HippoRAG framework (Yang et al., 2024) introduces a neurobiologically inspired approach to long-term memory for LLMs, implementing a system that prioritises relevance signals and iteratively refines its understanding. However, their approach does not explicitly model the graph evolution process or use graph structures to identify information gaps. ActiveRetrieval (Jiang et al., 2023) actively queries a corpus during the generation process, using intermediate reasoning states to guide retrieval. This approach demonstrates the value of dynamically adjusting retrieval based on the current reasoning state, a principle that our KGEIR framework incorporates through gap-aware retrieval.

While these existing approaches have made significant strides in different aspects of the MHQA challenge, they typically address only part of the problem. KGEIR uniquely integrates dynamic knowledge graph construction, gap identification, and iterative refinement into a unified framework that addresses the full spectrum of challenges in multi-hop question answering, differentiating it from existing approaches that typically address only part of the problem.

3 Methods

We introduce KGEIR (Knowledge Graph-Enhanced Iterative Reasoning), a novel framework for multi-hop question answering that combines the reasoning capabilities of large language models with the structural advantages of knowledge graphs. This section describes our approach, which dynamically constructs and refines knowledge graphs to support iterative reasoning over multiple documents.

3.1 Problem Analysis

Multi-hop question answering requires integrating information across multiple sources to derive answers that cannot be found in any individual source. We formalize this task as follows: Given a question q and a corpus of documents $\mathcal{D} = \{d_1, d_2, ..., d_N\}$, MHQA aims to produce an answer a by reasoning over a subset of supporting documents $\mathcal{S} \subset \mathcal{D}$ where:

- No single document $d_i \in \mathcal{S}$ contains sufficient information to answer q.
- The answer a requires establishing relationships between information in different documents.
- The reasoning process can be represented as a sequence of hops between documents, forming a path:

$$d_{i_1} \rightarrow d_{i_2} \rightarrow ... \rightarrow d_{i_k} \rightarrow a$$

Traditional approaches follow a retrieve-then-read paradigm that can be formalised as:

$$a = \mathcal{R}(\mathcal{T}(q, \mathcal{D}), q)$$

Where $\mathcal{T}$ is a retrieval function that selects relevant documents, and$\mathcal{R}$ is a reading function that extracts the answer. This approach faces challenges with multi-hop questions as the initial retrieval often fails to capture all necessary information.

Our KGEIR framework reformulates this problem by introducing an iterative graph-based approach:

$$a = \mathcal{R}(G_k, q)$$

Where G_k is a knowledge graph constructed and refined through k iterations of retrieval and reasoning. Each iteration identifies information gaps and retrieves additional context to fill these gaps, progressively enriching the graph until sufficient information is gathered to answer the question.

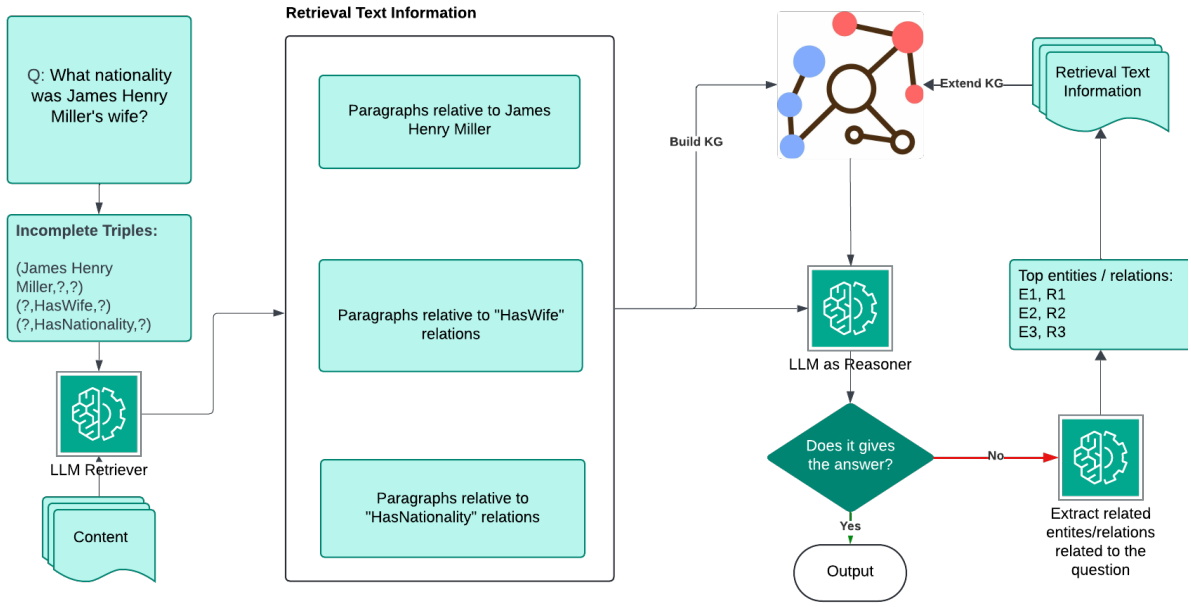

Figure 2: KGEIR framework workflow for multi-hop QA, illustrated with the example "What nationality was James Henry Miller's wife?" The process begins with extracting incomplete triples from the question (left), followed by multi-faceted retrieval extracting paragraphs relevant to the main entity and relations. The initial knowledge graph is constructed from retrieved paragraphs (centre), enabling structured reasoning by the LLM. If the current graph lacks sufficient information, the system identifies missing entities and relations to guide targeted retrieval (right), iteratively refining the knowledge graph until a confident answer can be produced. This dynamic enhancement process addresses the limitations of static retrieval approaches by adaptively exploring the information space based on reasoning requirements.

3.2 Initial Knowledge Graph Construction

The first component of KGEIR is constructing an initial knowledge graph from the question and corpus. As shown in Figure 1, this process involves question analysis, initial retrieval, and graph formation. Given a question, we first identify key entities and potential relationships required to answer it. For the question "What nationality was James Henry Miller's wife?" (Figure 1), we extract incomplete triples: (James Henry Miller, ?, ?), (?, HasWife, ?), and (?, HasNationality, ?). These incomplete triples capture both explicit entities mentioned in the question and implicit relations necessary to answer it. We only crop the corpus if it reaches the limitation of LLM's context length. Then, combining the text corpus with these extracted entities and relations, we design a prompt for the LLM to retrieve the relevant paragraphs from the corpus. This multi-faceted retrieval approach targets documents containing information about the entities ("James Henry Miller"), relations ("HasWife"), and properties ("HasNationality") in the query. This strategy ensures a broader coverage than traditional retrieval methods that focus on entities only. All our prompts used throughout the paper are available upon request.

From the retrieved paragraphs, we extract entities and relationships to construct the initial knowledge graph. This process creates a structured representation of the information contained in the retrieved documents, converting unstructured text into an explicit entity-relation graph. While valuable, this initial graph often lacks crucial connections needed to answer complex multi-hop questions, necessitating our dynamic enhancement approach.

3.3 Dynamic Knowledge Graph Enhancement

The core innovation of KGEIR is its dynamic approach to enhancing the knowledge graph through targeted retrieval and iterative refinement, as shown in the right portion of Figure 1. After constructing the initial graph, an LLM reasoner attempts to answer the question using the available knowledge structure. Upon determining that the current graph does not contain sufficient information for a confident answer to the question, the system initiates an enhancement cycle that operates through a

systematic information-seeking paradigm.

The enhancement process begins with a gap analysis mechanism that examines the knowledge graph structure to identify missing entities and relations. This mechanism employs specialised prompting techniques to direct the LLM's attention to specific structural deficiencies in the current graph. As illustrated in Figure 1, the system identifies entities that would be most informative for answering the query, modelling information necessity rather than merely information relevance. Following gap identification, the system employs relation-aware retrieval to efficiently locate documents containing the missing information. This targeted retrieval strategy differs significantly from traditional similarity-based approaches by formulating queries specifically designed to bridge identified knowledge gaps. The retrieval component employs both entity-centric and relation-centric query formulation to ensure comprehensive coverage of the missing information. The retrieved information undergoes structured extraction and integration into the existing knowledge graph through our graph extension mechanism (labelled "Extend KG" in Figure 1). This integration process preserves existing graph structure while incorporating new entities and relations, creating a progressively more comprehensive knowledge representation with each iteration. The enhancement cycle continues iteratively, with each cycle refining the knowledge graph until it contains sufficient information for confident reasoning or reaches a predetermined iteration limit. This dynamic refinement process enables KGEIR to overcome the limitations of static retrieval approaches, adaptively exploring the information space as directed by reasoning requirements rather than surface-level query similarity.

3.4 Knowledge-Guided Reasoning and Assessment

The final component of KGEIR leverages the dynamically enhanced knowledge graph to perform multi-hop reasoning and answer the question. Unlike approaches that reason directly over retrieved text, KGEIR reasons over the structured entity-relation graph, allowing for more precise navigation through complex information. The reasoning process leverages both the graph structure and the original retrieved passages, combining the advantages of structured knowledge representation with the contextual richness of the original text. As illustrated in Figure 1, the LLM reasoner identifies relevant paths through the knowledge graph that connect question entities to potential answers. For our example question, the reasoner would identify paths connecting "James Henry Miller" through the "HasWife" relation to his spouse, and then through the "HasNationality" relation to the target answer. By traversing these explicit relationship paths, the system effectively performs multi-hop reasoning while maintaining clarity about the evidence supporting each hop.

The reasoning process includes a simple verification step (the decision node in Figure 1) where the LLM determines if the current graph provides direct supporting information to answer the question. If more information is needed, the system triggers another enhancement cycle; otherwise, it proceeds to generate the final answer.

4 Experiment

4.1 Setup

Dataset and Retrieval Parameter For comprehensive evaluation of KGEIR, we conducted experiments on three established multi-hop QA benchmarks: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2023). These datasets represent varying degrees of reasoning complexity, including 2-hop, 3-hop, and 4-hop inference chains. Following established evaluation practices in this domain (Zhang et al., 2025), we selected a sample of 1,000 questions from each dataset's validation set. For the hyperparameter setting, we set the number of retrieved paragraphs to five for each iteration of the enhancement cycle, for a maximum of three iterations. If the model still cannot find the answer at this point, the question is marked as failed.

Baselines We evaluated KGEIR against representative methods spanning different approaches to multi-hop reasoning. We included both sparse retrieval with BM25 (Robertson and Zaragoza, 2009) and dense retrieval with BGE (Xiao et al., 2024) to establish performance baselines for non-structured approaches. We compared against the published results in Liu et al. (2025) on leading tree-structured systems, including RAPTOR (Sarthi et al., 2024) and SiReRAG (Zhang et al., 2025), as well as graph-based approaches namely GraphRAG (Edge et al., 2025) and HippoRAG (Gutiérrez et al., 2025). For all structured systems, GPT-4o has been used to maintain consistency between implementations.

Evaluation Metric We assessed performance using exact match (EM) and F1 scores as same as the setting in HopRAG model (Liu et al., 2025), which measures the precision of answer generation at different granularities. The EM metric requires exact correspondence with reference answers, while F1 combines precision and recall at the token level to provide a more nuanced measure of partial correctness. We focused exclusively on answer quality metrics rather than retrieval metrics, as several baseline systems generate synthetic content (such as summaries) that would make direct retrieval comparison inequitable.

4.2 Result Analysis

Table 1 presents a comprehensive evaluation of our proposed KGEIR framework against established baselines across three multi-hop QA datasets. Thus, KGEIR achieves competitive results with state-of-the-art methods, our novel mechanisms of dynamic knowledge graph construction and iterative reasoning.

On MuSiQue, KGEIR achieves 44.50% EM (exact matches) and 53.12% F1, showing modest improvements over HopRAG (42.20% EM, 54.90% F1). For HotpotQA, we observe performance of 63.15% EM and 76.77% F1, slightly higher than HopRAG's 62.00% EM and 76.06% F1. On 2WikiMultiHopQA, our approach achieves 59.13% EM and 69.55% F1, which is competitive though slightly lower than HopRAG's 61.10% EM and 68.26% F1. These results demonstrate that KGEIR achieves comparable performance to the current state-of-the-art while introducing a fundamentally different approach to multi-hop reasoning. The primary contribution of KGEIR is not a significant leap in raw performance metrics, but rather the introduction of a novel framework that enhances the reasoning process through explicit knowledge modelling and iterative refinement.

In terms of approach, KGEIR differs from HopRAG in several key aspects. While HopRAG prioritises logical connectivity between passages through pseudo-queries and multi-hop traversal, KGEIR focuses on dynamically constructing and refining explicit knowledge representations. Unlike HopRAG, which integrates similarity with logical relations when constructing edges, KGEIR explicitly models information gaps and uses these to guide targeted retrieval. The performance comparisons with traditional retrievers (BM25: 31.77% avg. EM, BGE: 36.17% avg. EM) highlight the significant advantages of structured approaches. Meanwhile, GraphRAG's lower performance (22.10% avg. EM) suggests that static knowledge graph construction alone is insufficient without iterative refinement mechanisms. Similar to how HopRAG positioned itself against SiReRAG by emphasising its streamlined graph structure without additional summary nodes, KGEIR introduces a novel dynamic knowledge graph construction process that evolves throughout reasoning. Our approach does not require pre-constructed knowledge graphs or complex graph preprocessing, instead, it builds and refines graph representations as reasoning progresses.

4.3 Ablation Experiment and Discussion

To evaluate the effectiveness of KG-based reasoning in our framework, we performed an ablation study comparing different reasoning methods following the retrieval phase. We examined four distinct approaches: (1) Vanilla (direct LLM reasoning without prompting), (2) Chain-of-Thought (CoT) (Wei et al., 2023), (3) Tree-of-Thought (ToT) (Yao et al., 2023), and (4) our complete KGEIR approach with knowledge graph reasoning. For all experiments, we used Gemma3-27B as the base model and maintained consistent dataset settings with our main evaluation. Performance was measured based on semantic correctness relative to ground truth answers.

As shown in Table 2, KGEIR consistently outperforms all baseline reasoning methods across all datasets, achieving an average improvement of 2.26% over ToT. The performance improvement is particularly pronounced on the HotpotQA dataset, where KGEIR achieves 62.20% accuracy compared to 57.70% for ToT—a 4.50% absolute improvement. All results suggest that our knowledge graph approach is very effective for complex bridging questions that require connecting information across multiple documents.

Table 2 shows a clear progressive improvement pattern (Vanilla $\rightarrow$ CoT $\rightarrow$ ToT $\rightarrow$ KGEIR), demonstrating the value of increasingly structured reasoning approaches. While CoT provides modest gains over vanilla reasoning (48.11% vs. 48.47% on average), ToT's tree-structured exploration offers more substantial improvements (53.85%). However, KGEIR's explicit modelling of entity relationships through dynamic knowledge graphs

Table 1: Comparison of RAG methods across datasets with baseline results from the cited literature.

Method	MuSiQue		2WikiQA		HotpotQA		Average	
	EM [%]	F1 [%]	EM [%]	F1 [%]	EM [%]	F1 [%]	EM [%]	F1 [%]
BM25	13.80	21.50	40.30	44.83	41.20	53.23	31.77	39.85
BGE	20.80	30.10	40.10	44.96	47.60	60.36	36.17	45.14
GraphRAG	12.10	20.22	22.50	27.49	31.70	42.74	22.10	30.15
RAPTOR	36.40	49.09	53.80	61.45	58.00	73.08	49.40	61.21
SiReRAG	40.50	53.08	59.60	67.94	61.70	**76.48**	53.93	65.83
HopRAG	42.20	**54.90**	**61.10**	68.26	62.00	76.06	55.10	**66.40**
KGEIR	**44.50**	53.12	59.13	**69.55**	**63.15**	73.77	**55.59**	65.48

Table 2: Comparison of ablation study between different reasoning methods across datasets.

Method	MuSiQue	2WikiQA	HotpotQA	Average
Vanilla (LLM 'as is')	42.60	62.21	40.62	48.47
CoT (LLM with CoT prompt)	44.50	57.25	42.57	48.11
ToT (LLM with ToT prompt)	45.65	58.21	57.70	53.85
KGEIR	46.45	59.69	62.20	56.11

provides the most effective reasoning framework (56.11%).

Interestingly, on 2WikiQA, the performance gap between reasoning methods is less pronounced, with vanilla LLM reasoning achieving a surprisingly high 62.21%. This suggests that for certain types of questions, the base reasoning capabilities of modern LLMs may be sufficient when retrieving appropriate context. Nevertheless, KGEIR still provides the most consistent performance across all datasets, demonstrating the robustness of our approach to different question types and reasoning complexities.

These results validate our hypothesis that structuring multi-hop reasoning through explicit knowledge graphs enhances LLMs' ability to connect information across documents, particularly for complex questions requiring multiple reasoning steps. The dynamic construction and refinement of knowledge representations provide a more interpretable and effective reasoning process compared to traditional prompting methods.

5 Conclusion

In this paper, we presented KGEIR, a novel framework that enhances multi-hop question answering through dynamic knowledge graph construction and iterative refinement. Unlike traditional retrieve-then-read paradigms, KGEIR explicitly models entity relationships across documents and systematically identifies information gaps to guide targeted retrieval. This iterative knowledge refinement process provides both a structured scaffold for LLM reasoning and an effective mechanism to address the inherent limitations of similarity-based retrieval for complex questions.

Our comprehensive evaluation across three multi-hop QA benchmarks demonstrates KGEIR's effectiveness, achieving competitive or superior performance compared to state-of-the-art methods. The most significant improvements appear on complex bridging questions, confirming our approach's strength in scenarios requiring cross-document reasoning. Ablation experiments reveal that structured knowledge graph reasoning consistently outperforms traditional prompting methods, with our full KGEIR model providing absolute improvements of up to 4.50% over Tree-of-Thought prompting.

The integration of dynamic knowledge graph construction with iterative reasoning represents a promising direction for addressing complex information needs in NLP systems. By bridging symbolic and neural approaches, KGEIR offers a principled solution to the challenges of information fragmentation and implicit relationships that characterise multi-hop reasoning tasks. We may extend this framework to incorporate uncertainty handling and conflicting information resolution, potentially expanding its applicability to a broader range of knowledge-intensive applications beyond question-answering.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. ArXiv:2310.11511 [cs].

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Kewei Cheng, Nesreen K. Ahmed, Theodore Willke, and Yizhou Sun. 2024. Structure Guided Prompt: Instructing Large Language Model in Multi-Step Reasoning by Exploring Graph Structure of the Text. ArXiv:2402.13415 [cs].

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. ArXiv:2404.16130 [cs].

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2025. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. ArXiv:2405.14831 [cs].

Haoyu Han, Yaochen Xie, Hui Liu, Xianfeng Tang, Sreyashi Nag, William Headden, Hui Liu, Yang Li, Chen Luo, Shuiwang Ji, Qi He, and Jiliang Tang. 2025. Reasoning with Graphs: Structuring Implicit Knowledge to Enhance LLMs Reasoning. ArXiv:2501.07845 [cs].

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. ArXiv:2402.07630 [cs].

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.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large Language Models Cannot Self-Correct Reasoning Yet. ArXiv:2310.01798 [cs].

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. ArXiv:2305.06983 [cs].

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, and Jiawei Han. 2024. Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 163–184, Bangkok, Thailand. 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.

Xun Liang, Simin Niu, Zhiyu Li, Sensen Zhang, Hanyu Wang, Feiyu Xiong, Jason Zhaoxin Fan, Bo Tang, Shichao Song, Mengwei Wang, and Jiawei Yang. 2025. SafeRAG: Benchmarking Security in Retrieval-Augmented Generation of Large Language Model. ArXiv:2501.18636 [cs].

Hao Liu, Zhengren Wang, Xi Chen, Zhiyu Li, Feiyu Xiong, Qinhan Yu, and Wentao Zhang. 2025. HopRAG: Multi-Hop Reasoning for Logic-Aware Retrieval-Augmented Generation. ArXiv:2502.12442 [cs].

Jinyoung Park, Ameen Patel, Omar Zia Khan, Hyunwoo J. Kim, and Joo-Kyung Kim. 2024. Graph Elicitation for Guiding Multi-Step Reasoning in Large Language Models. ArXiv:2311.09762 [cs].

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. 2019. Answering Complex Open-domain Questions Through Iterative Query

Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2590–2602, Hong Kong, China. Association for Computational Linguistics.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. ArXiv:2401.18059 [cs].

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. 2025. BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval. ArXiv:2407.12883 [cs].

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. ArXiv:2307.07697 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

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. Place: Cambridge, MA Publisher: MIT Press.

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.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. 2024. Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models. ArXiv:2410.07176 [cs].

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. ArXiv:2203.11171 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed Resources For General Chinese Embeddings. ArXiv:2309.07597 [cs].

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do Large Language Models Latently Perform Multi-Hop Reasoning? ArXiv:2402.16837 [cs].

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. ArXiv:2305.10601 [cs].

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving Language Models via Plug-and-Play Retrieval Feedback. ArXiv:2305.14002 [cs].

Nan Zhang, Prafulla Kumar Choubey, Alexander Fabbri, Gabriel Bernadett-Shapiro, Rui Zhang, Prasenjit Mitra, Caiming Xiong, and Chien-Sheng Wu. 2025. SiReRAG: Indexing Similar and Related Information for Multihop Reasoning. ArXiv:2412.06206 [cs].