Thinking Beyond the Local: Multi-View Instructed Adaptive Reasoning in KG-Enhanced LLMs

Minghan Zhang, Shu Zhao[*], Zhen Yang, Hongsheng Wu,
Yongxing Lin, Haodong Zou, Jie Chen, Zhen Duan
School of Computer Science and Technology, Anhui University, China
z17333121752@163.com, zhaoshuzs2002@hotmail.com
uscyz094@gmail.com, e23301289@stu.ahu.edu.cn
e23301224@stu.ahu.edu.cn, zou_hd@163.com
chenjie200398@163.com, dz@ahu.edu.cn

Abstract

Knowledge Graph-enhanced Large Language Models (KG-Enhanced LLMs) integrate the linguistic capabilities of LLMs with the structured semantics of Knowledge Graphs (KGs), showing strong potential in knowledge-intensive reasoning tasks. However, existing methods typically adopt query-driven iterative reasoning from a local perspective, which limits their ability to capture semantically distant but crucial information, leading to dual bottlenecks in efficiency and accuracy for complex multi-hop tasks. To address this issue, we propose MIAoG, a **M**ulti-view **I**nstructed **A**daptive reasoning of LLM **o**n K**G**, which is designed to overcome the limitations of local exploration by enabling LLMs to plan, evaluate, and adapt reasoning paths from a global perspective. Instead of query-anchored exploration, MIAoG first prompts the LLM to generate a multi-view instruction set that outlines diverse potential reasoning paths and explicitly specifies global reasoning intentions to guide the model toward coherent and targeted reasoning. During reasoning, MIAoG integrates a real-time introspection mechanism that evaluates the alignment between the current path and the instructions, adaptively pruning inconsistent trajectories to enhance global consistency while maintaining efficiency. Extensive experiments on multiple public datasets show that MIAoG achieves state-of-the-art performance in KG-enhanced LLM reasoning, particularly excelling in complex multi-hop scenarios. Our code is available at https://github.com/ahu-zmh/MIAoG.

1 Introduction

Large Language Models (LLMs) have witnessed rapid progress in recent years, showing remarkable superiority in a wide spectrum of natural language processing tasks (OpenAI, 2023; Touvron et al., 2023; Zeng et al., 2024). Despite their success, they continue to face fundamental challenges, among which hallucination remains a highly pervasive issue (Ji et al., 2023; Ye et al., 2024). Especially in complex scenarios requiring multi-hop reasoning and deep understanding, LLMs often exhibit reasoning biases, leading to outputs that deviate from factual correctness. To address these limitations, the incorporation of knowledge graphs (KGs) has emerged as a promising solution (Pan et al., 2024). By representing knowledge in a structured form through entities, attributes, and relationships, KGs construct semantically rich and logically coherent networks. This explicit structure provides LLMs with a reliable external knowledge source, offering a principled approach to support complex reasoning and effectively mitigate hallucinations in challenging tasks.

Building on this trend, Knowledge Graph-enhanced Large Language Models (KG-Enhanced LLMs) have become a rapidly growing research direction, demonstrating immense potential in complex knowledge-intensive reasoning and question answering (QA) tasks. One common approach is semantic parsing (Zhang et al., 2023; Xie et al., 2022; Ye et al., 2022; Li et al., 2023), where an LLM translates natural language questions into formal KG queries (e.g., SPARQL). This allows precise execution over the KG but heavily depends on the LLM's ability to correctly align natural language with the KG schema, which is often challenging for complex questions (Figure 1(a)).

Information retrieval is another approach that retrieves relevant triples or subgraphs from the KG through different retrieval strategies and integrates them with the LLM for downstream reasoning (Jiang et al., 2023; Guan et al., 2024; Liu et al., 2024; Wen et al., 2024). These methods anchor reasoning around a user query, prompting the LLM to explore paths within the Knowledge Graph to find evidence supporting the answer. While effective in simpler tasks, they typically operate from a local perspective—selecting the next hop based solely on

[*]Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2026, pages 6173–6188
March 24-29, 2026 ©2026 Association for Computational Linguistics

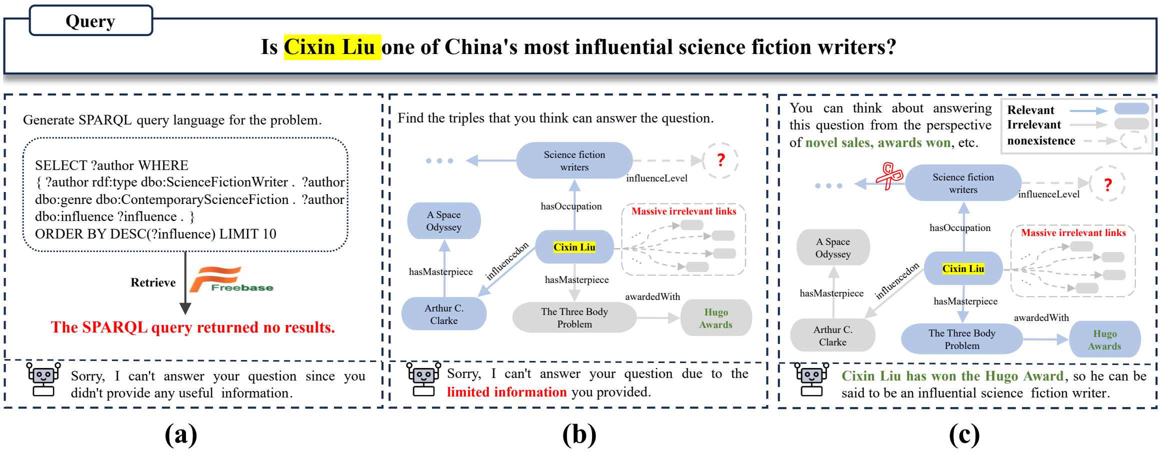
Query
Is Cixin Liu one of China's most influential science fiction writers?
Generate SPARQL query language for the problem.
SELECT ?author WHERE
{ ?author rdf:type dbo:ScienceFictionWriter . ?author dbo:genre dbo:ContemporaryScienceFiction . ?author dbo:influence ?influence . }
ORDER BY DESC(?influence) LIMIT 10
Retrieve
Freebase
The SPARQL query returned no results.
Sorry, I can't answer your question since you didn't provide any useful information.
Find the triples that you think can answer the question.
Science fiction writers
influenceLevel
?
hasOccupation
Massive irrelevant links
A Space Odyssey
Cixin Liu
influencedon
hasMasterpiece
Arthur C. Clarke
The Three Body Problem
awardedWith
Hugo Awards
Sorry, I can't answer your question due to the limited information you provided.
You can think about answering this question from the perspective of novel sales, awards won, etc.
Relevant
Irrelevant
nonexistence
Cixin Liu has won the Hugo Award, so he can be said to be an influential science fiction writer.
(a)
(b)
(c)

Figure 1: Different Methods of KG-enhanced LLM reasoning: (a) Semantic Parsing, which faces challenges in effectively aligning natural language queries with the underlying KG; (b) Reasoning Path Exploration, which tends to deviate from critical reasoning paths within the vast semantic search space; (c) Instruction-based Reasoning Path Exploration, which, by leveraging clear and structured instructions, achieves both precision and depth in the exploration process.

immediate neighbors. This lack of global foresight limits their efficacy in complex multi-hop scenarios involving semantically distant entities, where models are prone to deviating from the optimal path due to irrelevant graph distractions or query misinterpretation. As illustrated in Figure 1(b), when answering a question like *"Is Cixin Liu one of China's most influential science fiction writers?"*, the anchor entity *"Cixin Liu"* may connect to hundreds of surrounding entities. In the absence of task-oriented global guidance, the LLM is prone to veering off critical reasoning paths within the vast semantic search space, which in turn undermines both the accuracy and efficiency of the reasoning process. Although the latest methods (Sun et al., 2024; Chen et al., 2024; Wang et al., 2025) expand the search space through techniques such as beam search, self-correction, and relation abstraction, they still lack global planning, which limits their effectiveness in handling multi-hop reasoning tasks that span distant semantic contexts. Moreover, excessive reliance on search space expansion leads to inefficient resource allocation and further undermines the model's generalization ability.

In response to the aforementioned challenges, we propose MIAoG, a Multi-view Instructed Adaptive reasoning framework for KG-enhanced LLMs. As illustrated in Figure 1(c), MIAoG departs from conventional query-anchored reasoning by first prompting the LLM to conduct a fine-grained, multi-perspective semantic analysis of the query. This results in a diverse set of high-level reasoning instructions that explicitly define the goals and directions of the reasoning process, offering a global planning view across multiple candidate paths. Guided by these instructions, MIAoG enables the LLM to iteratively explore the knowledge graph along semantically meaningful reasoning trajectories. Crucially, the framework incorporates a real-time introspection mechanism that continuously evaluates the alignment between each reasoning path and the instruction set. Based on these alignment scores, paths lacking support from any instruction and instructions unsupported by all paths are adaptively pruned, ensuring tight coupling between planning and execution. This bidirectional adaptation balances reasoning accuracy with computational efficiency by preventing semantic drift and redundant exploration. Extensive experiments on three representative multi-hop KGQA datasets demonstrate the effectiveness and robustness of MIAoG in handling complex, multi-step reasoning tasks. The main contributions are summarized as follows:

- We propose a multi-view instructed KG-enhanced LLM reasoning framework that shifts reasoning from local query anchoring to global planning, enabling semantically comprehensive and coherent multi-hop reasoning.

- We design a dynamic focusing mechanism that continuously evaluates the alignment between reasoning paths and the instruction set, enabling real-time pruning of both irrelevant paths and unsupported instructions. This ensures tight coupling between planning and execution, effectively improving reasoning efficiency and accuracy.

- We conduct extensive experiments on three real-world KGQA datasets: CWQ, WebQSP, and GrailQA. The results demonstrate the effectiveness and efficiency of the MIAoG paradigm we proposed for KG-enhanced LLMs tasks.

2 Related Works

2.1 LLM Reasoning

In natural language processing, LLMs such as GPT and GLM can handle diverse tasks with simple prompts. However, standard prompting often yields direct answers, neglecting the reasoning steps behind complex problem-solving. Chain-of-Thought (CoT) (Wei et al., 2022) addresses this by deriving intermediate steps before producing answers, making reasoning more transparent and closer to human thinking. Building on this idea, researchers have explored automated CoT construction using LLMs' knowledge (Shao et al., 2023; Shum et al., 2023), and proposed advanced strategies such as self-consistency (Wang et al., 2023; Zelikman et al., 2022), thought trees (Yao et al., 2023), thought maps (Besta et al., 2024), and weighted CoT (Fang et al., 2025). These methods improve reasoning robustness and interpretability, helping models generalize better and reduce errors in multi-step inference. Nevertheless, most still rely on internal knowledge, lacking integration with external knowledge, which limits their reliability in complex reasoning.

2.2 Reasoning over KGs

Despite their outstanding performance in NLP tasks, LLMs still struggle with complex problems such as multi-hop and knowledge-intensive reasoning. KGs, storing knowledge triples in graph structures, are widely used to supplement LLMs. KG-based QA mainly retrieves evidence subgraphs and is generally divided into semantic parsing and information retrieval.

Semantic parsing translates natural language into KG-executable representations. Traditional methods construct query graphs via entity linking, attribute recognition, and constraint mounting, or employ Encoder-Decoder models to transform parsing into Seq2Seq problems with tree decoders (Lan and Jiang, 2020; Zhang et al., 2023; Xie et al., 2022; Ye et al., 2022). However, these approaches require large annotated data and lack transferability. Recent work (Li et al., 2023) explores using LLMs' in-context learning to generate graph queries directly, but this overly depends on LLMs and struggles with complex reasoning.

Information retrieval methods identify key entities via neural networks and extract candidate answers from KGs, reducing manual templates but suffering from low interpretability and mediocre performance. Recent works leverage LLMs to iteratively retrieve and generate interpretable evidence subgraphs. Sun et al. (2024) proposed an information interaction mechanism between KGs and LLMs for step-by-step reasoning. To further enhance structured reasoning, Jiang et al. (2023) introduced StructGPT with specialized interfaces and iterative reasoning procedures. In parallel, methods such as PoG (Chen et al., 2024) and ReKnoS (Wang et al., 2025) focus on adaptive and scalable graph reasoning through decomposition, self-correction, or relation abstraction. However, despite these advances, existing approaches still lack explicit global planning and often rely on inefficient search space expansion, which hinders their effectiveness in complex multi-hop reasoning. In contrast, our framework introduces multi-view semantic instructions and adaptive introspection to provide global guidance while ensuring efficient and accurate reasoning.

3 Methods

As shown in Figure 2, MIAoG first prompts the LLM to analyze the problem's intentions and generate multi-view instructions, which guide iterative exploration of reasoning paths on the KG. During exploration, both paths and instructions are adaptively pruned to progressively refine objectives until the LLM determines an answer. The process consists of three stages: Instruction Generation, Instructed Graph Exploration, and Answer Verification with Reasoning Introspection.

3.1 Instruction Generation

Given a question Q, we first utilize the natural language processing capabilities of the LLM to conduct multi-view analysis, generating a diverse set of instructive reasoning opinions $I = \{i_1, i_2, \ldots, i_n\}$. Simultaneously, the LLM extracts an initial set of subject entities $E^{(0)} = \{e_1^{(0)}, e_2^{(0)}, \ldots, e_m^{(0)}\}$ that

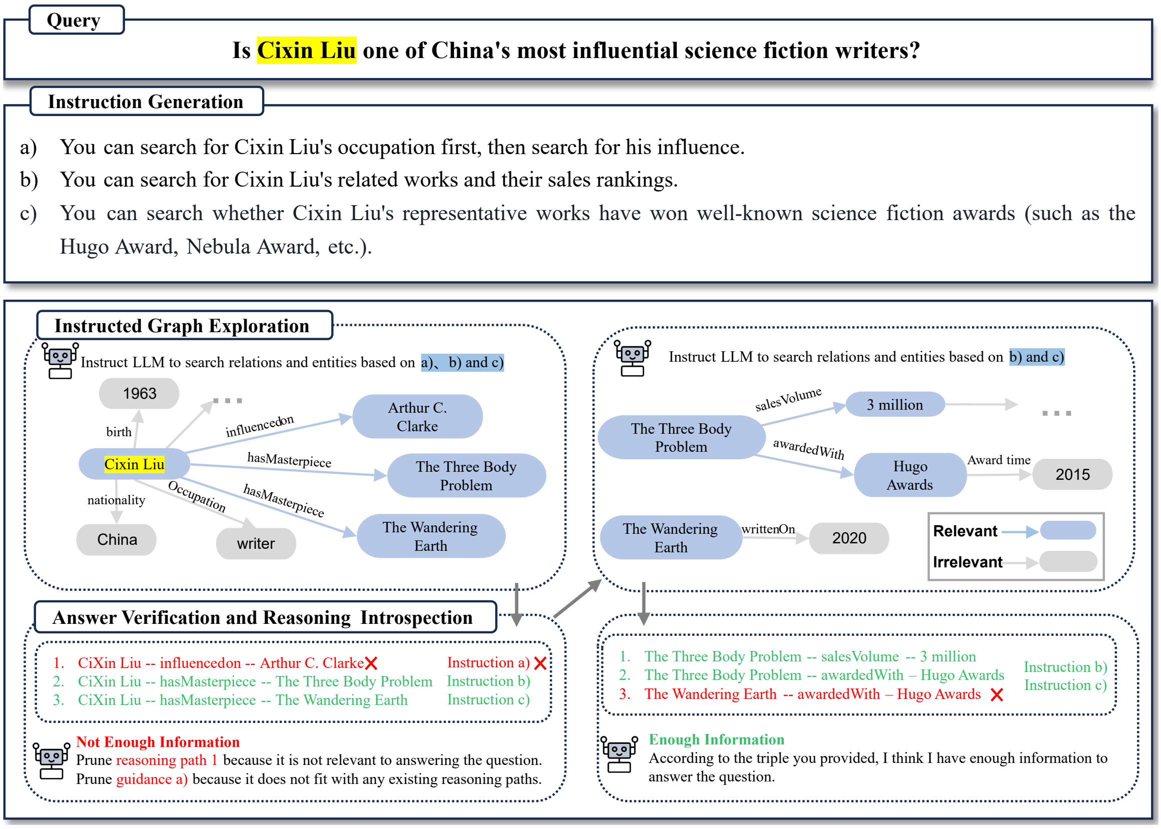
Query
Is Cixin Liu one of China's most influential science fiction writers?
Instruction Generation
a) You can search for Cixin Liu's occupation first, then search for his influence.
b) You can search for Cixin Liu's related works and their sales rankings.
c) You can search whether Cixin Liu's representative works have won well-known science fiction awards (such as the Hugo Award, Nebula Award, etc.).
Instructed Graph Exploration
Instruct LLM to search relations and entities based on a)、 b) and c)
1963
birth
Cixin Liu
influencedon
Arthur C. Clarke
hasMasterpiece
The Three Body Problem
nationality
Occupation
hasMasterpiece
China
writer
The Wandering Earth
Instruct LLM to search relations and entities based on b) and c)
salesVolume
3 million
The Three Body Problem
awardedWith
Hugo Awards
Award time
2015
The Wandering Earth
writtenOn
2020
Relevant
Irrelevant
Answer Verification and Reasoning Introspection
1. CiXin Liu -- influencedon -- Arthur C. Clarke ✗ Instruction a) ✗
2. CiXin Liu -- hasMasterpiece -- The Three Body Problem Instruction b)
3. CiXin Liu -- hasMasterpiece -- The Wandering Earth Instruction c)
Not Enough Information
Prune reasoning path 1 because it is not relevant to answering the question.
Prune guidance a) because it does not fit with any existing reasoning paths.
1. The Three Body Problem -- salesVolume -- 3 million
2. The Three Body Problem -- awardedWith – Hugo Awards
3. The Wandering Earth -- awardedWith – Hugo Awards ✗
Instruction b)
Instruction c)
Enough Information
According to the triple you provided, I think I have enough information to answer the question.

Figure 2: The framework overview of MIAoG.

are most relevant to Q, serving as the starting points for path reasoning. The reasoning process then iteratively explores KG under the guidance of I.

3.2 Instructed Graph Exploration

During the iterative process, we maintain a set of reasoning paths $P = \{p_1, p_2, \ldots, p_n\}$, each starting from an entity in $E^{(0)}$. At the t-th iteration ($t \geq 1$), the set of tail entities $E^{(t-1)} = \{e_1^{(t-1)}, e_2^{(t-1)}, \ldots, e_k^{(t-1)}\}$ from the current paths are taken as starting points. For each entity in $E^{(t-1)}$, we explore its associated relations and corresponding tail entities, appending the resulting triples to extend the paths in P. After t iterations, each path in P comprises t evidence triples. This exploration process consists of two key components: instruction-based relationship exploration and instruction-based entity exploration.

Instructed Relationship Exploration: At the t-th iteration ($t \geq 0$), we commence from the current entity set $E^{(t)} = \{e_1^{(t)}, e_2^{(t)}, \ldots, e_k^{(t)}\}$. For each entity $e_i^{(t)} \in E^{(t)}$, we consider the complete set of its incoming and outgoing relations from the KG, defined as:

$$\mathcal{R}(e_i^{(t)}) = \{r \mid (e_i^{(t)}, r, \cdot) \vee (\cdot, r, e_i^{(t)})\}. \quad (1)$$

Guided by the (potentially pruned) instruction set I, the LLM filters and ranks these relations. It assigns an importance score $s_r(r_j) \in [0, 1]$ to each relation $r_j \in \mathcal{R}(e_i^{(t)})$, selecting the top-$K$ relations to form a candidate relation set $R^{(t)} = \{r_1^{(t)}, r_2^{(t)}, \ldots, r_k^{(t)}\}$, where $K = \max(1, |I|)$. This process prioritizes relations that align with the reasoning directions specified by the instructions.

Instructed Entity Exploration: For each selected relation $r_j^{(t)} \in R^{(t)}$, we retrieve the set of all connected tail entities, denoted as $\mathcal{E}(r_j^{(t)})$. The LLM then evaluates each entity e_l in this set under the guidance of the instruction set I, assigning an entity relevance score $s_e(e_l) \in [0, 1]$. Simultaneously, the previously assigned relation score $s_r(r_j^{(t)})$ is recalled. A comprehensive triple score for the candidate $(e_i^{(t)}, r_j^{(t)}, e_l)$ is computed as the product of these scores:

$$s_{\text{triple}}(e_i^{(t)}, r_j^{(t)}, e_l) = s_r(r_j^{(t)}) \cdot s_e(e_l). \quad (2)$$

The top-K triples across all explored $(e_i^{(t)}, r_j^{(t)})$ pairs, ranked by s_{triple}, are selected. The tail entities from these top-K triples form the new entity set $E^{(t+1)} = \{e_1^{(t+1)}, e_2^{(t+1)}, \ldots, e_k^{(t+1)}\}$ for the next iteration, and the corresponding triples are appended to the reasoning paths in P.

For simple questions, such as *"Who influenced xx?"*, the LLM may generate only a single instruction, quickly identifying the key relation (*influenced_by*) and avoiding unnecessary computation, which enhances reasoning efficiency. For more complex multi-hop questions, the LLM initially explores a broader set of possible paths and, guided by I', progressively focuses on the most promising ones. This adaptive adjustment strategy, driven by real-time assessment of question complexity, optimizes the reasoning path while improving the generalization capability of the LLM.

3.3 Answer Verification and Reasoning Introspection

At the end of each iteration, the LLM evaluates whether the current reasoning paths, all grounded in knowledge graph triples, are sufficient to answer Q. To prevent hallucinations, MIAoG is explicitly prompted to either rely on such verifiable evidence or output "insufficient information." If the evidence-supported paths are deemed sufficient, the LLM generates a response. Otherwise, the instruction set I and the reasoning path set P are pruned based on the available reasoning clues, resulting in refined subsets I' (where $|I'| \leq |I|$) and P' (where $|P'| \leq |P|$). In the early stages of reasoning, the exploration breadth is intentionally large, allowing the model to explore diverse potential directions. As reasoning progresses, this breadth gradually narrows through pruning, which removes redundant or unpromising instructions and paths. This progressive narrowing not only concentrates on the most promising reasoning directions but also avoids unnecessary computational overhead.

To achieve this, the LLM evaluates the alignment between each instruction–path pair (i, p) by assigning an answerability score $s(i, p) \in [0, 1]$. This score reflects the likelihood that path p will fulfill the reasoning objective specified in instruction i. Based on these scores, a bidirectional pruning step is applied:

Path pruning: A reasoning path p is discarded if $\max_{i \in I} s(i, p) \leq \tau$, where τ is a predefined threshold. This retains only paths showing promise under at least one instruction and if all paths are below the threshold, the path with the highest total score is saved;

Instruction pruning: An instruction i is discarded if $\max_{p \in P} s(i, p) \leq \tau$. This removes instructions that no longer have supporting evidence in the current paths, thereby refining the focus of the subsequent exploration.

After pruning, we obtain the refined sets I' and P', which are aligned to ensure that subsequent reasoning focuses on evidence-supported directions until a path sufficient to answer the question is found or the maximum search depth is reached. Detailed prompts are provided in Appendix F.

4 Experiments

4.1 Experimental Setup

Datasets & Evaluation Metrics : To demonstrate the effectiveness of MIAoG on complex multi-hop reasoning tasks, we employ three representative multi-hop KGQA datasets, including WebQSP (Yih et al., 2016), CWQ (Talmor and Berant, 2018), and GrailQA (Gu et al., 2021), which contain up to four-hop questions. We used Freebase (Bollacker et al., 2008) as the data source. Freebase is a large-scale, multi-domain knowledge graph developed by Google, comprising over 250 million entities, thousands of relations, and attributes such as dates, locations, and numerical values. Built through collaborative and automated processes, it inevitably contains incompleteness and noise, serving as a challenging testbed for reasoning models under imperfect knowledge. For all datasets, we adopt the exact match accuracy (Hits@1) as the evaluation metric based on previous studies (Sun et al., 2024; Chen et al., 2024; Wang et al., 2025).

Detail: In our experiments, we utilize ChatGPT (gpt-3.5-turbo) and GPT-4o as backbone models for fair comparison with other baselines. During exploration, the temperature is set to a higher 0.6 to accommodate more diverse instructed opinions. During inference, the temperature parameter is set to 0 to ensure the accuracy of the inference. The maximum token length limit for generation is 256. In all experiments, we set the inference depth D to 3, and prompt the LLM to generate up to 3 instructions, as supported by our analysis in Appendix B. For pruning, the threshold is fixed at $\tau = 0.2$ unless explicitly stated otherwise.

Method	**WebQSP**	**CWQ**	**GrailQA**			
			overall	**I.I.D.**	**Compositional**	**Zero-shot**
			LLM only w/ ChatGPT			
IO prompt (Brown et al., 2020)	63.3	37.8	29.6	23.1	19.8	31.2
CoT (Wei et al., 2022)	61.8	38.2	28.1	22.3	21.2	32.1
SC (Wang et al., 2023)	61.2	40.1	29.8	24.5	23.0	30.1
			Semantic Parsing			
QGG (Lan and Jiang, 2020)	73.0	44.1	-	-	-	-
Rng-kbqa (Ye et al., 2022)	76.2	-	68.8	86.2	63.8	63.0
KB-BINDER (Li et al., 2023)	74.4	-	58.5	-	-	-
			Information Retrieval w/ ChatGPT			
StructGPT (Jiang et al., 2023)	72.6	54.3	-	-	-	-
ToG (Sun et al., 2024)	76.4	58.0	70.2	70.1	56.1	72.7
PoG (Chen et al., 2024)	82.0	63.2	76.5	76.3	62.1	81.7
ReKnoS (Wang et al., 2025)	81.9	63.1	76.8	76.5	63.0	81.2
MIAoG (Ours)	82.3	65.8	77.2	78.0	65.1	82.1
			Information Retrieval w/ GPT-4o			
StructGPT (Jiang et al., 2023)	79.5	64.7	-	-	-	-
ToG (Sun et al., 2024)	83.6	71.0	82.1	79.4	67.3	86.5
PoG (Chen et al., 2024)	86.4	74.1	84.2	86.6	69.1	88.2
ReKnoS (Wang et al., 2025)	86.1	73.2	83.6	86.1	68.4	87.5
MIAoG (Ours)	**86.8**	**74.9**	**84.9**	**87.2**	**70.2**	**89.0**

Table 1: Performance comparison of different methods on WebQSP, CWQ and GrailQA.

4.2 Baselines

We compare MIAoG with widely used baselines and state-of-the-art methods, which are mainly divided into three categories: 1) Question answering based on the LLM's own capabilities, including standard prompting (IO prompt) (Brown et al., 2020), Chain-of-Thought prompting (CoT) (Wei et al., 2022), and Self-Consistency (SC) (Wang et al., 2023). 2) Semantic parsing methods based on traditional or LLM, including QGG (Lan and Jiang, 2020), Rng-kbqa (Ye et al., 2022), and KB-BINDER (Li et al., 2023). 3) LLM-based information retrieval methods, including StructGPT (Jiang et al., 2023), ToG (Sun et al., 2024), PoG (Chen et al., 2024), and ReKnoS (Wang et al., 2025), which are most similar to our work. The descriptions of baselines are presented in Appendix A.

4.3 Main Result

As shown in Table 1, our method achieves the best performance on all three datasets. First, compare to using only the knowledge of the LLMs itself to answer questions, MIAoG achieves a significant improvement on all three datasets by retrieving external knowledge. This result highlights the importance of introducing external knowledge to alleviate the hallucination of LLM. The semantic parsing approach relies on transforming natural language questions into structured query language, a process that is relatively effective when the question structure is simple and the semantics are clear. However, when face with questions like those in CWQ and GrailQA, which involve multi-hop reasoning and complex semantic relationships, the accuracy of semantic parsing significantly declines, indicating the insufficiency of these methods in handling complex question parsing.

In the information retrieval paradigm, MIAoG demonstrates a significant advantage over the classic ToG baseline. Its overall performance on WebQSP, CWQ, and GrailQA is 4.55%, 5.85%, and 4.9% higher than ToG, respectively. MIAoG also maintains a sustained advantage when compared to the latest strong methods like PoG and ReKnoS.

Method	WebQSP	CWQ	GrailQA
	Llama3-8B		
CoT	54.1	34.1	28.1
MIAoG	65.8	46.2	43.3
Gain	**+11.7**	**+12.1**	**+20.5**
	Llama3-70B		
CoT	60.1	42.3	29.2
MIAoG	73.7	58.5	54.5
Gain	**+13.6**	**+16.2**	**+25.3**
	ChatGPT		
CoT	61.8	39.2	28.6
MIAoG	82.3	65.8	77.2
Gain	**+20.5**	**+26.6**	**+48.6**
	GPT-4o		
CoT	67.1	45.3	34.3
MIAoG	86.8	74.9	84.9
Gain	**+19.7**	**+29.6**	**+50.6**

Table 2: Performances of MIAoG using different backbone models on three datasets.

Furthermore, it is noteworthy that MIAoG achieves the most significant performance improvement on the more complex multi-hop reasoning tasks within the CWQ dataset. This is primarily attributed to MIAoG's design, where the macro-guided reasoning approach allows for deeper and more systematic reasoning on complex problems, effectively helping LLMs overcome potential knowledge and comprehension bottlenecks.

4.4 Backbone Models Comparison

Given that the core steps of MIAoG largely rely on the natural language processing capabilities of LLMs, we evaluate the impact of backbone models with different parameter sizes (including Llama3-8B and Llama3-70B) across three datasets. As shown in Table 2, integrating MIAoG's reasoning module significantly enhances all LLMs. Even with the smallest Llama3-8B, MIAoG surpasses GPT-4o's direct inference performance, demonstrating strong compatibility and adaptability across models of different scales.

Moreover, more powerful LLMs benefit more from MIAoG, demonstrating better synergy with structured KG knowledge. For the complex and diverse questions in CWQ and GrailQA, all models show larger gains compared to WebQSP, with GPT-4o achieving maximum improvements of 1.50 and 2.57 times, respectively. This indicates that MIAoG not only ensures consistent improvements across models of varying scales but also enhances synergy with stronger LLMs when handling complex multi-hop questions, underscoring its robustness and scalability.

Method	WebQSP	CWQ	GrailQA
MIAoG	**82.3**	**65.8**	**77.2**
w/o Instruction	77.9	57.2	69.2
w/ BM25	58.3	54.2	60.5
w/ SentenceBRET	62.9	57.7	62.3
w/o Instruction pruning	81.7	64.5	75.5
w/o Path pruning	82.3	65.6	77.1

Table 3: Performance after removing or replacing each mechanism.

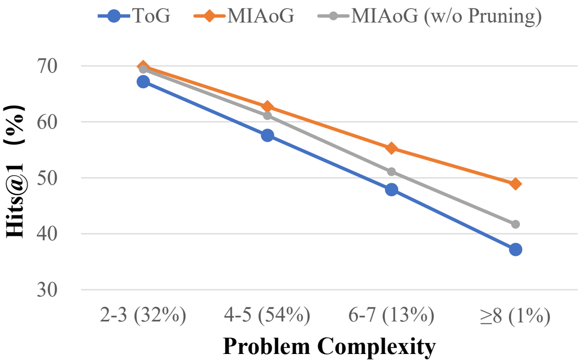

Figure 3: Performance comparison on the CWQ dataset stratified by reasoning depth. Reasoning depth is defined by the number of relations in the corresponding SPARQL query. The percentages in the x-axis labels indicate the proportion of questions falling within each depth range.

4.5 Ablation Study

To evaluate the effectiveness of multi-view instructions and adaptive focus, we conducted experiments on three datasets. For multi-view instructions, we explored evidence paths without multi-view instructions (w/o Instruction) and compared conditions using BM25 and SentenceBERT as pruning tools (w/ BM25 and w/ SentenceBERT) to validate the effectiveness of different exploration methods. For adaptive focus, we compared two modes: one without instruction pruning (w/o Instruction pruning) and one without path pruning (w/o path pruning), to evaluate the impact of pruning on performance.

Table 3 shows that both multi-view instructions

Question: What is there to see and do in the location where the Maltese breed originated?	
KB-BINDER	**Answer: [The generated SPARQL query language is not executable, and LLM refuses to answer.]** I'm sorry, could you please first briefly introduce the origin of the Maltese breed so that I can answer your question more accurately. Thank you.
ToG	**Paths Exploration:** Maltese → *biology.animal_breed.place_of_origin* → *Malta* → *location.country.capital*→ *Valletta* → *film.film_location.featured_in_films* → *Munich* ↓ ↘ *location.location.events* → *Great Siege of Malta* … **Answer: [No useful information was found in KG, LLM answered the question with its own knowledge]** *The Maltese dog originated in Malta. Its capital, Valletta, has many historical sites worth visiting, such as sites related to the Great Siege of Malta and the Megalithic Temple.*
MIAoG	**Instruction : [Explore *tourist attractions*, places of interest, and festivals in the region where the Maltese breed originated.]** **Paths Exploration:** Maltese → *biology.animal_breed.place_of_origin*→ *Malta* ***Fort Rinella*** ←*location.country.**tourist_attractions*** ↙ ↘ *location.country.capital* ↛ **Answer: [Under the instruction of clear macro opinions, LLM can easily find the key relationship of "*tourist_attractions*"]** The Maltese breed originated in Malta, and Malta has a tourist attraction, Fort Rinella, so Fort Rinella can be a place to visit.

Figure 4: A typical case to compare different methods to answer the complex multi-hop question.

Method	Tokens	calls	Time
ToG	7175	23.1	69.7
StructGPT	6325	12.8	29.1
PoG	5751	**9.3**	17.1
ReKnoS	4851	12.3	35.2
MIAoG	**4633**	10.5	**15.2**

Table 4: Efficiency comparison between our proposed MIAoG and several baseline approaches.

and adaptive focus contribute positively to performance, as their removal weakens results on complex QA tasks. Furthermore, while multi-view analysis provides richer semantic information, simple similarity-based pruning cannot replace LLMs' semantic understanding. In w/o Instruction pruning, we observe that without pruning instructions, misaligned guidance can adversely affect reasoning, leading to erroneous directions. In w/o path pruning, exploring multiple paths under clear instruction does not significantly boost performance but reduces efficiency due to irrelevant explorations. In contrast, MIAoG avoids incorrect path extensions through adaptive pruning of instruction suggestions, thereby substantially enhancing exploration efficiency. A more detailed analysis of the dynamic pruning behavior across datasets is provided in Appendix C, further illustrating MIAoG's ability to adjust its reasoning strategy based on question complexity.

4.6 Impact of Reasoning Depth

To evaluate the model's capability in complex scenarios, we stratified the CWQ dataset by reasoning depth. As shown in Figure 3, the performance of the baseline ToG drops significantly as depth increases, suffering from error accumulation in blind exploration. In contrast, MIAoG demonstrates superior robustness, outperforming ToG by 9.7% at depths of 8 hops or more.

Furthermore, the comparison with the w/o Pruning variant (defined as removing the adaptive pruning mechanism entirely, i.e., both instruction and path pruning) reveals the critical role of our adaptive mechanism. While the impact of pruning is marginal in shallow hops (+0.5%), the gap widens significantly to 5.2% in deep reasoning ($\geq$8 hops). This confirms that as the search space expands, adaptive pruning becomes indispensable for maintaining global consistency and correcting deviations from the planned instructions.

4.7 Efficiency Analysis

To demonstrate the efficiency of our adaptive focusing algorithm, we performed a detailed analysis of MIAoG against several other methods. The results presented in Table 4, which show the average performance across all three datasets.

MIAoG dramatically outperforms the classic ToG method across all efficiency metrics. By employing an adaptive focusing algorithm to avoid exploring irrelevant paths—unlike ToG's predefined breadth—MIAoG achieves superior precision with far fewer interactions. Specifically, it reduces average token consumption by 35.4%, cuts LLM invocations by 54.4%, and slashes the average time from 69.7 to a mere 15.2. Furthermore, when compared to recent methods like StructGPT, PoG, and ReKnoS, MIAoG maintains the highest overall efficiency. Although some competitors may slightly edge it out on a single metric, MIAoG achieves the lowest token consumption and the fastest time, making it the most efficient method overall.

4.8 Case Study

Figure 4 presents a typical complex multi-hop question from the CWQ dataset, aimed at comparing the performance and working mechanisms of MIAoG and ToG, the most representative methods from the Semantic Parsing (KB-BINDER) and Information Retrieval (ToG) paradigms. In KB-BINDER, due to the unexecutability of the SPARQL query and the lack of information retrieved from the KG, the LLM refuses to reply directly. Both ToG and MIAoG correctly identify the origin of the Maltese as Malta. However, when further exploring information related to Malta, the ToG model struggles to filter out the correct direction from nearly 200 relationships, leading to reasoning bias. While other approaches like PoG and ReKnoS also expand the search space to enhance reasoning, they still encounter similar reasoning bias in complex multi-hop tasks. In contrast, the MIAoG model, guided by a clear perspective, quickly locates the key relationship "tourist_attractions" and ultimately deduces Fort Rinella as the answer. Furthermore, while ToG engages in extensive futile exploration in incorrect directions when critical information is missing, the MIAoG model effectively avoids resource wastage through introspection and pruning mechanisms after each exploration, demonstrating higher efficiency and accuracy.

5 Conclusion

We propose MIAoG, a multi-view instructed adaptive reasoning framework that preprocesses user queries with semantic analysis to generate multi-view reasoning instructions, allowing LLMs to navigate multi-hop reasoning paths precisely from a macro perspective. During reasoning, MIAoG adaptively focuses the exploration scope and refines the direction, effectively avoiding unnecessary resource consumption and improving reasoning efficiency and accuracy. Results show that MIAoG significantly reduces reasoning costs, outperforms existing methods, and excels in complex multi-hop tasks.

Ethical Statement

This work utilizes publicly available datasets and does not involve personally identifiable information. While our proposed MIAoG framework improves reasoning efficiency and reduces computational costs, it relies on external Knowledge Graphs that may contain factual errors or inherent biases. Consequently, users should exercise caution and verify outputs when deploying the model in sensitive or high-stakes domains.

Limitations

The MIAoG framework, despite its strong performance in multi-hop reasoning over Knowledge Graphs (KGs), faces two main constraints. It's currently tailored for high-precision, single-answer reasoning and primarily assessed using Hits@1. This narrow focus limits its applicability to more complex scenarios that involve multiple valid answers or demand a richer evaluation captured by metrics like the F1-score. Furthermore, while it does improve efficiency with adaptive pruning, MIAoG's iterative reasoning still generates nontrivial computational overhead. Applying it to large-scale KGs with millions of entities and relations could lead to practical latency and scalability issues. Addressing both its limited scope of reasoning and its computational demands is essential for broadening MIAoG's utility and robustness.

Acknowledgments

Our work is supported by the National Natural Science Foundation of China (62476003), Anhui Province Excellent Scientific Research and Innovation Team (2024AH010004), Anhui Provincial Natural Science Foundation- Water Science Joint Fund (2408055US006), the University Synergy Innovation Program of Anhui Province (GXXT2023-050), and SMP-Zhipu AI Large Model CrossDisciplinary Fund (SMP-Zhipu20240210). We also acknowledge the support from Zhipu AIAnhui University Joint Research Center, and the High-Performance Computing Platform of Anhui University.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17682–17690. AAAI Press.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Liyi Chen, Panrong Tong, Zhongming Jin, Ying Sun, Jieping Ye, and Hui Xiong. 2024. Plan-on-graph: Self-correcting adaptive planning of large language model on knowledge graphs. 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*.

Yuanheng Fang, Guoqing Chao, Wenqiang Lei, Shaobo Li, and Dianhui Chu. 2025. Cdw-cot: Clustered distance-weighted chain-of-thoughts reasoning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 23878–23886. AAAI Press.

Yu Gu, Sue Kase, Michelle Vanni, Brian M. Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond I.I.D.: three levels of generalization for question answering on knowledge bases. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3477–3488. ACM / IW3C2.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18126–18134. AAAI Press.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9237–9251. Association for Computational Linguistics.

Yunshi Lan and Jing Jiang. 2020. Query graph generation for answering multi-hop complex questions from knowledge bases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 969–974. Association for Computational Linguistics.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023. Few-shot in-context learning on knowledge base question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6966–6980. Association for Computational Linguistics.

Jiaxiang Liu, Tong Zhou, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Enhancing large language models with pseudo- and multisource- knowledge graphs for open-ended question answering. *CoRR*, abs/2402.09911.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7):3580–3599.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 30706–30775. PMLR.

Kashun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12113–12139. Association for Computational Linguistics.

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. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 641–651. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Song Wang, Junhong Lin, Xiaojie Guo, Julian Shun, Jundong Li, and Yada Zhu. 2025. Reasoning of large language models over knowledge graphs with super-relations. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. 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*, pages 10370–10388. Association for Computational Linguistics.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, and 4 others. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 602–631. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2024. Cognitive mirage: A review of hallucinations in large language models. In *Proceedings of the First International OpenKG Workshop: Large Knowledge-Enhanced Models co-locacted with The International Joint Conference on Artificial Intelligence (IJCAI 2024), Jeju Island, South Korea, August 3, 2024*, volume 3818 of *CEUR Workshop Proceedings*, pages 14–36. CEUR-WS.org.

Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. 2022. RNG-KBQA: generation augmented iterative ranking for knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6032–6043. Association for Computational Linguistics.

Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, and 36 others. 2024. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793.

Lingxi Zhang, Jing Zhang, Yanling Wang, Shulin Cao, Xinmei Huang, Cuiping Li, Hong Chen, and Juanzi Li. 2023. FC-KBQA: A fine-to-coarse composition framework for knowledge base question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1002–1017. Association for Computational Linguistics.

A Baseline Descriptions

We compared MIAoG with widely used baselines and state-of-the-art methods, which are mainly divided into three categories:

1) Question answering based on the LLM's own capabilities:

- **IO** (Brown et al., 2020)**:** Standard input-output prompts are used for direct input-output testing of LLM.
- **Chain of Thought (CoT)** (Wei et al., 2022)**:** The LLM is encouraged to enhance its reasoning ability by generating a series of intermediate reasoning steps.
- **Self-Consistency (SC)** (Wang et al., 2023)**:** The answer is obtained by multiple sampling iterations and voting.

2) Semantic parsing methods based on traditional or LLM:

- **QGG:** (Lan and Jiang, 2020)**:** QGG proposes a phased query graph generation method for complex question answering. It integrates early constraints to prune the search space, boosting accuracy and efficiency in handling constrained and multi-relationship queries.
- **Rng-kbqa:** (Ye et al., 2022)**:** Rng-kbqa is a knowledge base question answering method that combines ranking and generation techniques, improving performance through iteratively trained rankers and T5-based generators, especially adept at handling unseen KB pattern problems.
- **KB-BINDER:** (Li et al., 2023)**:** KB-Binder method generates drafts of logical forms using LLM, and combines with the knowledge base for entity and relationship binding, achieving few-shot context learning without training, effectively solving the entity and relationship matching problem in knowledge base question answering.

3) Information retrieval methods based on LLM:

- **StructGPT:** (Jiang et al., 2023)**:** StructGPT introduces a general framework that enables large language models (LLMs) to perform reasoning over structured data such as knowledge graphs, tables, and databases. The framework employs specialized interfaces and an iterative reading-then-reasoning (IRR) procedure to facilitate structured data access and logical inference.
- **ToG:** (Sun et al., 2024)**:**Think-on-Graph technique allows for tight coupling interaction between LLMs and KGs, driving the LLM agent to step-by-step search and infer the optimal answer on the associated entities of the KG. This achieves traceability, error correction, and modification of knowledge.
- **PoG:** (Chen et al., 2024)**:** Plan-on-Graph (PoG) achieves self-correcting and adaptive reasoning by decomposing questions into subobjectives and leveraging large language models for adaptive path exploration, memory updating, and self-reflection.
- **ReKnoS:** (Wang et al., 2025)**:** ReKnoS introduces the concept of Super-Relations, which groups and abstracts related relations in a knowledge graph, enabling the simultaneous representation and exploration of multiple relation paths. This approach significantly expands the search space and improves the retrieval success rate.

B Effect and Robustness of Multi-view Instructions

B.1 Impact of Number of Instructions

To investigate the impact of the number of multi-view instruction on the performance of MIAoG, we conducted experiments with settings where the LLM was prompted to generate a maximum of 1-4 instruction. As shown in Figure 5, the performance of MIAoG improves with increasing number of instruction, accompanied by an expansion of the maximum exploration breadth. However, when the number of instruction exceeds 3, the marginal benefit decreases; therefore, in all experiments, we prompt the LLM to generate a maximum of 3 instruction.

B.2 Robustness to Imperfect Initial Instructions

Given that our framework relies on the quality of the initial multi-view instructions, one natural concern is whether biases or errors in these early instructions might negatively impact the reasoning process. To address this, MIAoG is designed

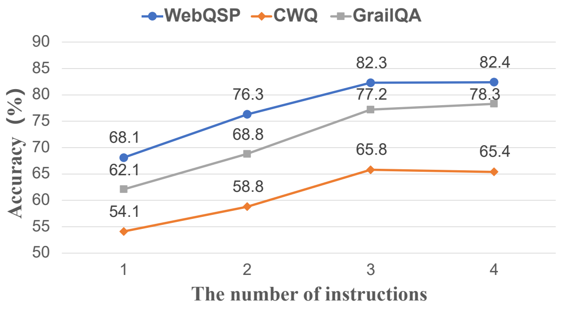

Figure 5: The impact of the number of instructions on performance.

Condition	WebQSP	CWQ
Baseline (3 LLM-generated)	82.3	65.8
Expanded Guidance (6 unfiltered)	81.2	64.9
Noisy Instructions Injected	80.5	62.8

Table 5: Performance of MIAoG under imperfect initial instructions. The expanded and noisy settings introduce redundant or irrelevant guidance to test robustness.

with two key mechanisms to ensure robustness: a breadth-first exploration strategy in the early stages, which prevents premature commitment to suboptimal paths, and an adaptive pruning mechanism that continuously filters out irrelevant or misleading instructions as reasoning progresses.

To further assess the system's tolerance to noisy or incomplete guidance, we conducted two stress-test experiments:

- In the first setting, we expanded the number of initial instructions from 3 to 6, all generated by the LLM without filtering. Many of these were redundant or only weakly relevant.

- In the second, we manually injected 3 unrelated instructions drawn from different questions.

As shown in Table 5, our adaptive pruning mechanism helps mitigate the impact of weak or misleading instructions by gradually filtering them out, allowing the model to focus on relevant reasoning paths. Although performance declines with increased noise, the drop remains modest, demonstrating MIAoG's robustness under suboptimal guidance.

C Dynamic Focusing and Pruning

C.1 Pruning Behavior under the Dynamic Focusing Mechanism

We conducted a systematic and comprehensive analysis of the dynamic focusing mechanism employed during the reasoning process of MIAoG. Figures 6 and 7 illustrate the frequency of instruction and reasoning path pruning performed by MIAoG across three benchmark datasets—WebQSP, CWQ, and GrailQA—when generating the final answer. The results indicate that MIAoG actively performs both instruction and path pruning in the majority of examples, clearly demonstrating its ability to dynamically adjust the reasoning process. However, the pruning behaviors exhibit significant variation across different types of questions.

Specifically, instruction pruning is most prominent in the WebQSP dataset, which contains structurally simpler and more direct questions. This can be attributed to the clarity of objectives in such questions, making redundant or non-essential instructions easier to identify and eliminate. In contrast, instruction pruning is less frequent in the more complex CWQ dataset, which features multi-hop reasoning tasks. These problems often require guidance from multiple perspectives, and overly aggressive pruning may hinder comprehensive exploration. In comparison, reasoning path pruning is most frequent in the CWQ dataset. Multi-hop questions inherently involve a larger search space and higher uncertainty, increasing the likelihood of encountering spurious or unproductive reasoning paths. By leveraging the dynamic focusing mechanism, MIAoG effectively filters out these paths, reducing reasoning bias and concentrating computational resources on more promising trajectories.

These findings suggest that MIAoG is capable of adapting its pruning strategies based on question complexity. While maintaining—or even improving—answer accuracy (as shown in Table 3), it significantly reduces unnecessary computation by selectively pruning instructions and reasoning paths. The dynamic focusing mechanism thereby contributes to more efficient and robust reasoning, offering a generalizable and effective solution for complex multi-hop question answering tasks.

C.2 Reliability of Pruning

To further evaluate the effectiveness and safety of our pruning strategy, we conducted a path-level

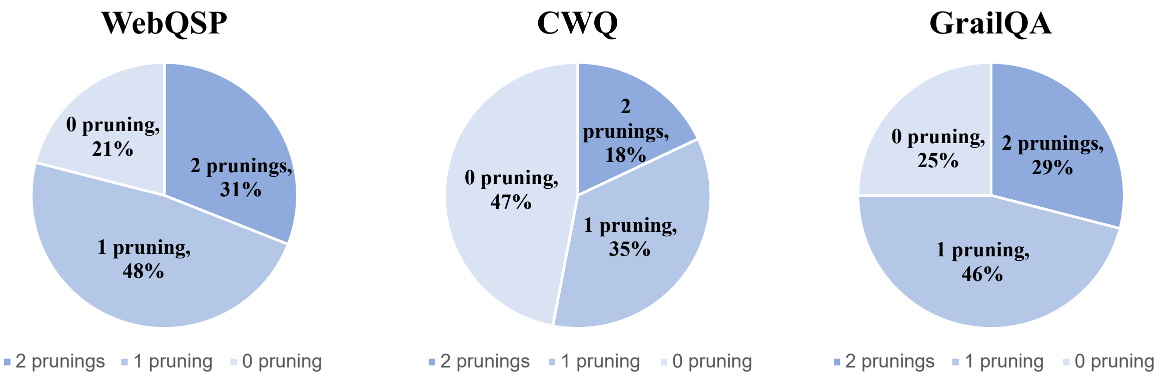

Figure 6: Distribution of instruction pruning counts on WebQSP, CWQ, and GrailQA.

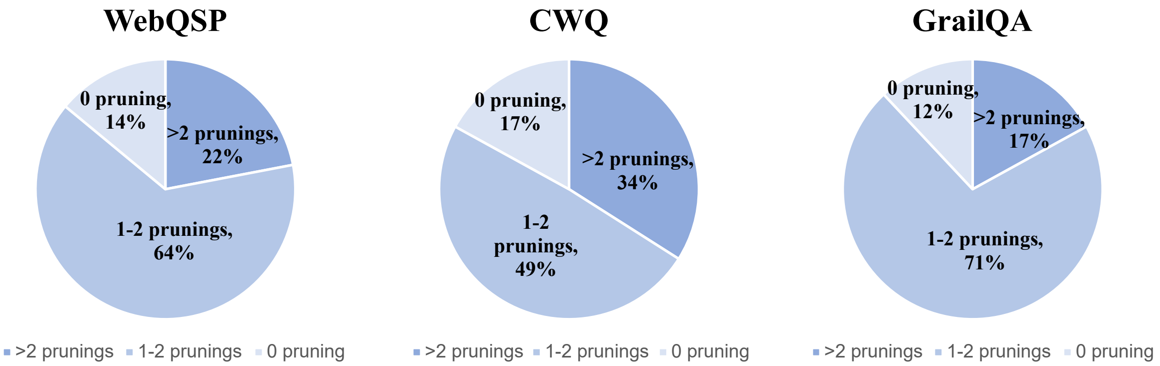

Figure 7: Distribution of path pruning counts on WebQSP, CWQ, and GrailQA.

Dataset	Correct Path Pruned (%)
WebQSP	1.7
CWQ	2.5
GrailQA	2.1

Table 6: Proportion of correct reasoning paths pruned during adaptive pruning across datasets.

analysis to assess the risk of mistakenly removing correct reasoning paths. As shown below, the proportion of gold answer paths that were inadvertently pruned remains low across datasets:

As shown in Table 6, our pruning strategy is generally conservative and reliable, effectively filtering out irrelevant or misleading paths while preserving those that are essential for accurate reasoning in the vast majority of cases.

D Impact of Entity Linking

To further address concerns regarding the reliance on gold entity linking and to ensure a fair comparison with methods that do not use gold entities (e.g., QGG, Rng-kbqa, and KB-BINDER), we also evaluated MIAoG in a setting without gold entities. In this scenario, we prompt the LLM to identify candidate entities and perform iterative reasoning to determine the correct starting points. As shown in Table 7, MIAoG maintains comparable performance levels even without gold entities. This demonstrates that MIAoG's effectiveness stems from its robust reasoning mechanism rather than a dependency on perfect entity linking, and it continues to significantly outperform semantic parsing baselines that do not use gold entities.

Method	WebQSP	GrailQA
QGG (Lan and Jiang, 2020)	73.0	-
Rng-kbqa (Ye et al., 2022)	76.2	68.8
KB-BINDER (Li et al., 2023)	74.4	58.5
MIAoG (w/o Gold Entity)	82.1	77.0
MIAoG (w/ Gold Entity)	**82.3**	**77.2**

Table 7: Performance comparison of MIAoG with and without gold entities against semantic parsing baselines.

E Stability and Sensitivity Analysis

We assess the stability of the LLM's scoring mechanism and the impact of the pruning threshold τ.

Score Type	Variance
Relation Score	0.0082
Entity Score	0.0065
Path Pruning Score	0.0042
Instruction Pruning Score	0.0049

Table 8: Variance of scoring components across five independent runs.

Threshold (τ)	Hits@1 (%)
0.1	64.8
0.2	**65.8**
0.4	65.1
0.6	65.3
0.8	64.6

Table 9: Sensitivity analysis of the pruning threshold τ on CWQ.

First, to address concerns regarding LLM stochasticity, we measured the variance of different scores across five independent runs on 200 random CWQ samples. As shown in Table 8, the variances are consistently low (< 0.01), confirming that the scoring mechanism is stable and reliable.

Second, we evaluated the model's sensitivity to the pruning threshold τ. Table 9 demonstrates that performance is robust within a reasonable range. The optimal performance is achieved at $\tau = 0.2$; lower thresholds introduce noise, while higher thresholds aggressively prune valid paths.

F Prompts

Here, we provide all the prompts used in MIAoG. To facilitate LLM output parsing, we require LLM to provide answers strictly in the way the examples are output.

Query Instruction Generation

Please analyze and think about the problem from different angles, give different instructions perspectives in concise language, and tell me what information should be unearthed to more effectively answer the question step by step. Provide the one to three most important instructions perspectives.

Example:
......
Question: {}
Instruction:

Table 10: Prompt for Query Instruction Generation.

Entities Exploration

Please refer to the instruction provided about the problem(If you think it helps you score the entity, otherwise ignore it), score the entities' contribution to the question on a scale from 0 to 1 (the sum of the scores of all entities is 1).

Example:
......
Question: {}
Instruction: {}
Relation: {}
Entites: {}
Score:

Table 11: Prompt for Entities Exploration.

Relations Exploration

Please pick the 3 relations (separated by semicolons) that contribute most to the problem from the relations I provide, and rate their contribution on a scale of 0 to 1 (the sum of the scores of 3 relations is 1).You can refer to the instructions if you think they are helpful.

Example:
......
Question: {}
Instruction: {}
Relation: {}
Entites:

Table 12: Prompt for Relations Exploration.

Reasoning Introspection

Given a question, the instruction to help answer the question, and the retrieved knowledge graph triple paths (entity, relation, entity), help me complete the following tasks.

Task 1: For each instruction and each path, assign an "answerability score" between 0 and 1.

- 0 means the path is completely irrelevant to this instruction and cannot help answer the question.

- 1 means the path is highly relevant and provides direct evidence to support or refute the instruction in answering the question.

- Values in between indicate partial relevance (the path may provide indirect or weak support).

Example:
......
paths: {}
Instruction: {}
Score:

Table 13: Prompt for Reasoning Introspection.

Answer Verification
Given a question and the associated retrieved knowledge graph triplets (entity, relation, entity), you are asked to answer whether it's sufficient for you to answer the question with these triplets and your knowledge (Yes or No). Example: ...... Question: {} Knowledge Triplets: {} Answer:

Table 14: Prompt for Answer Verification.

Answer
Given a question and the associated retrieved knowledge graph triplets (entity, relation, entity), you are asked to answer the question with these triplets and your knowledge. Example: ...... Question: {} Knowledge Triplets: {} Answer:

Table 15: Prompt for Answer.