Large Language Model and Knowledge Graph Entangled Logical Reasoning

Anonymous ACL submission

Abstract

Large language models (LLMs) and knowledge graphs (KGs) have complementary strengths and weaknesses for logical reasoning. LLMs exhibit strong semantic reasoning capacities, but they lack world knowledge and structured reasoning abilities. KGs contain extensive factual knowledge but have limited language understanding and reasoning flexibility. In this paper, we propose a framework LKLR that entangles LLMs and KGs for synergistic reasoning. A key technique is transforming the LLM's implicit reasoning chain into a grounded logical query over the KG, enabling seamless integration. Traversing this query grounds each inference step in KG facts while maintaining reasoning flow, combining the robust knowledge of KGs with the semantic reasoning of LLMs. Our approach synergistically integrates neural and symbolic reasoning to achieve hybrid reasoning capabilities. Experimental results on several QA benchmarks show that our proposed framework achieves state-of-the-art performance and provides transparent and reliable reasoning.

1 Introduction

LLMs have demonstrated powerful capabilities in language understanding and reasoning (Wei et al., 2022), owing to their pretraining on massive text corpora. A major strength is their ability to perform logical reasoning purely based on semantic patterns in language. However, LLMs have significant limitations in structured deductive reasoning. Firstly, their lack of world knowledge results in unconfident inferences despite strong semantic reasoning capacities (Ji et al., 2023). Secondly, LLM lacks the ability to accurately verify the correctness of their inferences, often generating logical but incorrect conclusions. Finally, LLMs struggle to reason about novel compositions of existing knowledge, limited to their pretraining data distribution.

In contrast, KGs (Bollacker et al., 2008; Vrandecic and Krötzsch, 2014) contain vast structured knowledge about the world in the form of entities and relations. This structured knowledge could make deductions through logical reasoning grounded in facts. However, pure KGs have notable limitations for logical reasoning. Firstly, there is a gap between semantic reasoning and structured reasoning, while mapping text to structured queries is challenging. Secondly, KGs contain facts but no predefined reasoning patterns tailored for specific questions. Manually engineering reasoning rules is difficult and leads to brittle performance. Finally, rule-based reasoning with KGs can fail for complex multi-hop reasoning, as pre-defined rules cannot cover all possible reasoning paths.

Despite their individual limitations, LLMs and KGs each possess complementary strengths that could enable more robust logical reasoning when combined synergistically. Firstly, the inability of LLMs to validate inferences could be augmented by leveraging the factual accuracy of structured KGs to correct unsupported leaps in reasoning. Secondly, the lack of semantic reasoning capabilities in KGs could be overcome by utilizing the language understanding capacities of LLMs to map text to formal queries. Thirdly, the flexible multi-step inferencing of LLMs could provide guidance to direct valid multi-hop reasoning paths on KGs. By compensating for their respective weaknesses in this complementary manner, a hybrid system could achieve greater factual accuracy, natural language understanding, and rigorous multi-step deductive reasoning than either LLMs or KGs alone. This provides strong motivation for developing a framework that entangles the capabilities of LLMs with KGs for logical reasoning.

While combining LLMs and KGs is promising, there are fundamental challenges that must be addressed. Simply incorporating external knowledge sources like search engines or knowledge graphs

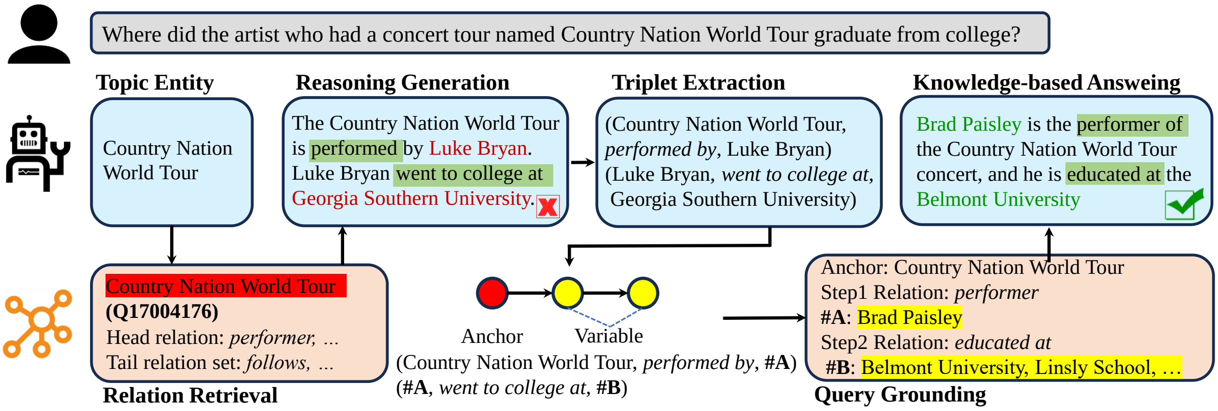

Where did the artist who had a concert tour named Country Nation World Tour graduate from college?
Topic Entity
Country Nation World Tour
Reasoning Generation
The Country Nation World Tour is performed by Luke Bryan. Luke Bryan went to college at Georgia Southern University.
Triplet Extraction
(Country Nation World Tour, performed by, Luke Bryan) (Luke Bryan, went to college at, Georgia Southern University)
Knowledge-based Answeing
Brad Paisley is the performer of the Country Nation World Tour concert, and he is educated at the Belmont University
Country Nation World Tour
(Q17004176)
Head relation: performer, ...
Tail relation set: follows, ...
Relation Retrieval
Anchor
Variable
(Country Nation World Tour, performed by, #A)
(#A, went to college at, #B)
Anchor: Country Nation World Tour
Step1 Relation: performer
#A: Brad Paisley
Step2 Relation: educated at
#B: Belmont University, Linsly School, ...
Query Grounding

Figure 1: An overview of LKLR, the whole framework consists of four components: (1) topic-based reasoning chain generation, (2) reasoning chain to logical query, (3) logical query grounded, (4) knowledge-based answering. For each step, LLM and KG are both used to enhance each other.

into existing LLMs has difficulties. Firstly, search engines (Zhao et al., 2023) return excessive irrelevant information without structure. Moreover, reformulating complex reasoning questions into executable program (Li et al., 2023d), such as SPARQL, exceeds current natural language generation capabilities. If consulting KGs step-wise, it will disrupt overall reasoning flow and coherence. More fundamentally, effectively integrating the symbolic knowledge representations in KGs with the distributed representations in LLMs needs more research. The reasoning mechanisms behind KGs and LLMs differ substantially—one is based on logical formalisms while the other relies on neural networks. Enabling synergistic reasoning between them requires thorough interaction and knowledge transfer across the symbolic and neural systems. In summary, the core challenge lies in developing integration between the neural reasoning of LLMs and the symbolic reasoning of KGs. This calls for innovative techniques to extract structured knowledge from LLMs, validate it against KGs, ground integrated reasoning chains in structured knowledge, and enable joint optimization between the two systems. Overcoming these challenges is the key to realizing the potential of hybrid systems for advanced logical reasoning.

To address these integration challenges, we propose a framework that entangles LLMs and KGs for synergistic reasoning. Our approach leverages the complementary strengths of both systems through an iterative workflow as shown in Figure 1: Firstly, we utilize the chain-of-thought reasoning capacity of the LLM to decompose the question into a logical query, since we hypothesize the LLM's reasoning approach implicitly contains a valid logical structure, though with spurious entities generated by the model. This query decomposition enables explicit integration of symbolic knowledge with the LLM's distributed representations. We then traverse the query step-by-step on the KG, with each inference grounded in factual knowledge while maintaining an overall reasoning flow guided by the LLM. Finally, the LLM contextualizes the extracted subgraphs to answer the original question. This framework enables tight interaction between the LLM and KG to validate inferences, ground them in structured knowledge, and leverage the reasoning capacities of both systems in a synergistic manner.

We conduct extensive experiments demonstrating the effectiveness of our proposed approach on logical reasoning tasks. We consider three challenging settings: multi-hop knowledge base question answering, open-domain question answering, and slot filling for entity-centric queries. Across four standard benchmarks for these tasks, our integrated framework achieves new state-of-the-art results.

Our contributions could be summarized as:

- We propose a novel framework, LKLR, that combines the complementary reasoning strengths of LLMs and KGs. This enables robust and rigorous reasoning by utilizing the advantages of each system.

- We develop an innovative technique to transform free-form reasoning questions into grounded logical queries over KGs.

- We conduct extensive experiments on four logical reasoning benchmarks, demonstrating the state-of-the-art performance of our integrated approach.

2 Preliminaries

Previously, we mentioned that we would transform a natural language question into a logical query. In this section, we will introduce the definition of logical query. Given a set of entities $\mathcal{V}$ and relations $\mathcal{R}$, a knowledge graph $\mathcal{G}$ is defined as $(\mathcal{V}, \mathcal{R}, \mathcal{T})$, where $\mathcal{T}$ represents triplets. A triplet, denoted as $r(e_i, e_j)$, signifies the existence of relation r between entities e_i and e_j, both belonging to $\mathcal{V}$.

In a logical query q, anchor entities are denoted by the set $\mathcal{V}_a \subseteq \mathcal{V}$, and existential quantified variable nodes are represented as $V_1, V_2, \ldots, V_k$. The target answer is expressed as the variable $V?$. Following the approach proposed by BetaE (Ren and Leskovec, 2020), the logical query is structured in disjunctive normal form, where it can be written as a disjunction of conjunctions:

$$q[V_?] := V_? : V_1, V_2, \ldots, V_k : c_1 \lor c_2 \lor \cdots \lor c_n.$$

Here, each c_i is a conjunction of literals a_{ij}, expressed as $c_i = a_{ij} \land \cdots \land a_{im}$. An atom or negation of an atom, such as $r(e_a, V)$, $\neg r(e_a, V)$, $r(V', V)$, or $\neg r(V', V)$, represents a_{ij}, where $e_a \in \mathcal{V}a$, $V, V' \in V_1, V_2, \ldots, V_k, V?$, and $V \neq V'$.

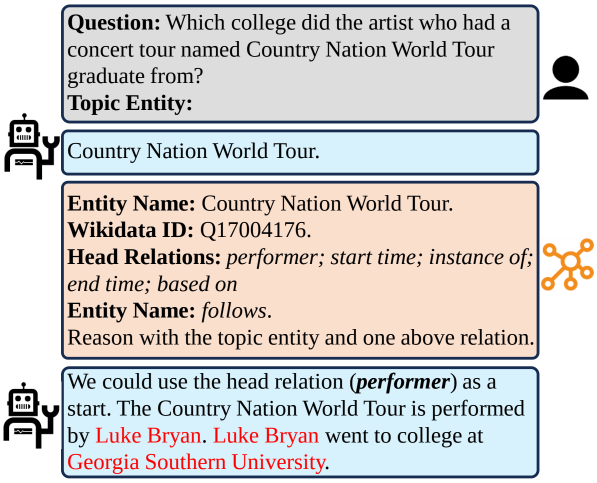
Logical queries incorporate variables, constants, relations, and logical operators, where variables signify entities for inference, constants anchor the query, relations indicate connections, and logical operators like intersection impose constraints on entity sets. A crucial aspect of our framework involves translating the reasoning chain of the LLM into a logical query so it can be executable over the KGs.

3 Method

In this section, we describe the methodology for our proposed framework, LKLR, to entangle LLMs and KGs handling logical reasoning. As outlined previously, our approach aims to leverage the complementary strengths of LLMs and KGs through an iterative process that grounds an LLM's implicit reasoning chain into the structured query. The framework comprises four stages: 1) reasoning chain generation by LLM based on the topic entity in question and relation in KG, 2) transformation from reasoning chain to a complex logical query, 3) execution of the logical query grounded in the KG, and 4) answering the original question with knowledge triplets. This section provides the technical details for each stage . We demonstrate how our techniques enable tight integration between the neural reasoning of the LLM and symbolic reasoning of the KG to achieve robust deductive reasoning that is both semantically driven and knowledge-grounded.

3.1 Topic-based Reasoning Chain Generation

The first stage of LKLR involves generating a reasoning chain for a given question as the example in Figure 2.



Question: Which college did the artist who had a concert tour named Country Nation World Tour graduate from?
Topic Entity:

Country Nation World Tour.

Entity Name: Country Nation World Tour.
Wikidata ID: Q17004176.
Head Relations: *performer; start time; instance of; end time; based on*
Entity Name: *follows.*
Reason with the topic entity and one above relation.

We could use the head relation (***performer***) as a start. The Country Nation World Tour is performed by Luke Bryan. Luke Bryan went to college at Georgia Southern University.

Figure 2: Example for the topic-based reasoning chain generation, the steps are started from the topic entity and the related relation.

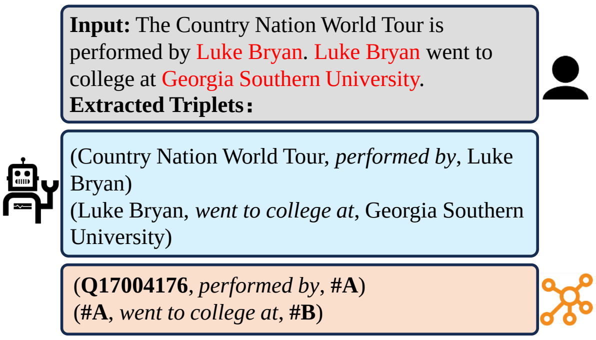
We identify the key topic entities in the question by prompting the LLM to extract the keywords, since we need an anchor entity for logical reasoning over the KG in the following stage. After that, all relations of a topic entity are retrieved in the KG to construct a head relation set and a tail relation set. Then the LLM needs to construct the full reasoning chain like the chain-of-thought (CoT), but the difference is that the reasoning must initiate from a topic entity and select a first-step relation from the two relation sets connected to the topic entity. It then continues reasoning based on the initiated relation and entity. The constraint on the first step of inference will make it easier for the transformed query to be grounded on KG in the following stage. If the model is allowed to freely construct the inference chain, the incompleteness of KG may result in the query not being grounded on KG.

We choose to leverage the LLM's own reasoning chain as a starting point because large language models can produce coherent reasoning flows, though individual factual statements may be unreliable. Our approach maintains the overall reasoning direction while replacing specific entities to ground the chain in knowledge.

This seeds the CoT with KG-based relations while leveraging the LLM's strengths in chaining logical inferences. The resulting CoT contains an implicit reasoning structure that will be made explicit in the next stage through query transformation.

3.2 Reasoning Chain to Logical Query

The second stage transforms the reasoning chain into a grounded logical query over the knowledge graph as the example in Figure 3.



Input: The Country Nation World Tour is performed by Luke Bryan. Luke Bryan went to college at Georgia Southern University.
Extracted Triplets:

(Country Nation World Tour, *performed by*, Luke Bryan)
(Luke Bryan, *went to college at*, Georgia Southern University)

(**Q17004176**, *performed by*, **#A**)
(**#A**, *went to college at*, **#B**)

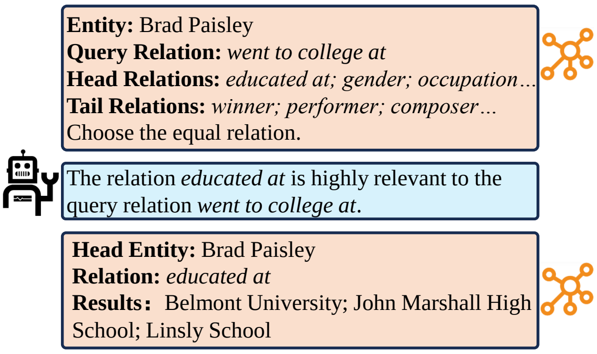
Figure 3: Example for the transformation from reasoning to logical query.

Firstly, we extract the chain into triplet facts using the LLM, as large models excel at this information extraction task (Li et al., 2023a; Chern et al., 2023). Secondly, we process the triplets to form variables for querying: entities presented in the original topic entities are kept, while other non-topic entities are replaced with variables. This grounds the query in the topic while allowing inference chaining. Crucially, the same entity is replaced with the same variable across all triplets, and the results of this variable correspond to the intersection of the different result sets of this variable in different atoms. Additionally, we filter triplets where both entities become variables, unless those variables could be linked with one of the topic entities through other variables. Otherwise, such triplets with disconnected variables will fail to be grounded during query execution due to a lack of grounding.

The resulting transformed triplets form a logical query with topic entities and intersecting variables for multi-hop inference. This makes the implicit reasoning structure explicit for execution over the KG while maintaining relevance to the original question through topic grounding.

3.3 Logical Query Grounding

The third stage executes the logical query through multi-hop reasoning over the KG, and the example in Figure 4 shows one step. We begin from the topic entities identified earlier and traverse the query triplets sequentially.



Entity: Brad Paisley
Query Relation: *went to college at*
Head Relations: *educated at; gender; occupation*...
Tail Relations: *winner; performer; composer*...
Choose the equal relation.

The relation *educated at* is highly relevant to the query relation *went to college at*.

Head Entity: Brad Paisley
Relation: *educated at*
Results: Belmont University; John Marshall High School; Linsly School

Figure 4: Example for logical query grounding.

A key problem is linking the query relations to the actual relations in the KG, as there may be name inconsistencies between the language-generated relation and the relations obtained from the KG. To address this, we leverage the language model to select the most appropriate KG relations that semantically match the current query relation.

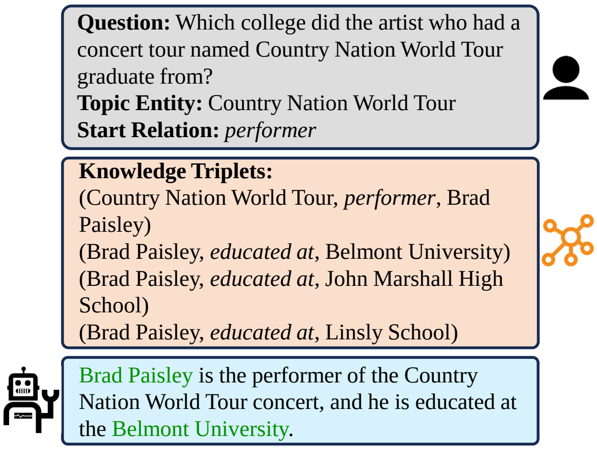
Executing a triplet can retrieve multiple entity candidates, but part of them may satisfy the current step query. However, they could not satisfy the overall query structure finally if the variable is also shown in other triplets. Our use of intersecting variables across triplets constrains the results to entities fulfilling the logical constraints. Performing intersection at each reasoning step prunes the search space and reduces final redundant results.

If the full reasoning chain fails to be grounded on the KG, we provide feedback to the LLM indicating where the failure occurred. This allows the model to update its reasoning approach and generate an alternative chain.

Overall, this stage grounds each reasoning step in the KG by eliminating the disambiguation of relations while leveraging the query structure to maintain relevance and validity. The output is extracted subgraphs containing inferred chains connected to the topic entities, with related knowledge triplets.

3.4 Knowledge-based Answering

The fourth stage involves the LLM utilizing the extracted knowledge triplets to answer the original question like the example in Figure 5.



Question: Which college did the artist who had a concert tour named Country Nation World Tour graduate from?
Topic Entity: Country Nation World Tour
Start Relation: *performer*

Knowledge Triplets:
(Country Nation World Tour, *performer*, Brad Paisley)
(Brad Paisley, *educated at*, Belmont University)
(Brad Paisley, *educated at*, John Marshall High School)
(Brad Paisley, *educated at*, Linsly School)

Brad Paisley is the performer of the Country Nation World Tour concert, and he is educated at the Belmont University.

Figure 5: Example for the final answering with the knowledge.

First, we provide the LLM with the question text, the topic entities and the first step relation. This primes the model to continue its initial line of reasoning. Next, we supply the LLM with the full set of subgraphs extracted and grounded over the KG with the logical query. These provide external knowledge to augment the LLM's fact inferences. Finally, we instruct the LLM to contextualize this knowledge by continuing its initial reasoning chain to generate the final answer to the question. Providing the grounded topic entities and relations focuses the LLM on reasoning paths most relevant to the question and extracted knowledge.

Importantly, even if the full query fails to be grounded completely due to KG incompleteness, the partial chains and entities retrieved could still provide useful knowledge. The LLM can utilize these grounded facts to improve its final answer.

This stage allows the LLM to interpret and synthesize the retrieved knowledge using its strong language capacities, answering the question by combining its reasoning with structured external facts.

4 Experiment

In this section, we detail the experimental setup to evaluate the proposed LKLR framework for entangling LLMs and KGs for synergistic reasoning. We aim to assess the effectiveness of our approach in logical reasoning tasks.

4.1 Dataset and Evaluation Metrics

We evaluate LKLR on four standard logical reasoning benchmarks: **WebQSP** (Yih et al., 2016) is tailored for multi-hop question answering over KGs. The result could demonstrate the model's proficiency in multi-hop reasoning. **QALD-10(en)** (Perevalov et al., 2022) is a benchmark for semantic web question answering, featuring questions from diverse domains. The dataset serves as a testbed for LKLR's capabilities in handling complex queries. **WebQuestion** (Berant et al., 2013) is designed for open-domain question answering. LKLR is tested on its capacity to answer diverse questions by reasoning over the knowledge graph, demonstrating its effectiveness in open-domain scenarios. **T-REx** (ElSahar et al., 2018) is a dataset containing large-scale high-quality alignments between DBpedia abstracts and Wikidata triples. We use this dataset to assess the performance of LKLR framework in handling the slot-filling task.

To evaluate the accuracy of the different, exact match accuracy (Hits@1) is used following previous works (Sun et al., 2023c).

4.2 Baselines

Standard prompting (IO prompt) (Brown et al., 2020): Models are provided with example questions, but the answers lack any explicit reasoning process, focusing solely on outputs.

Chain-of-thought prompting (CoT prompt) (Wei et al., 2022): It presents models with example questions, each accompanied by an explicit reasoning chain or process. This prompts the model to answer questions incorporating the understanding of intermediate steps in the reasoning process.

Self-Consistency prompting (SC prompt) (Wang et al., 2023): Guiding the language model with a CoT prompt, prompting the generated multiple reasoning paths through multiple samples, and selecting the most consistent answer via voting.

Think on Graph (ToG) (Sun et al., 2023c): The model conducts multi-step reasoning from the topic entity in the question. The exploration involves selecting relations using LLM and employs a beam-search approach to obtain multiple paths on KG with a max depth, and LLM scores each path to get the answer. The main difference between ToG and our model is that ToG only considers the current reasoning step, without holistically consider-

	Multi-Hop		Open-Domain	Slot-Filling
	WebQSP	**QALD10-en**	**WebQuestions**	**T-Rex**
		without knowledge graph		
IO prompt	63.3(+13.6%)	42(+27.4%)	48.7(+25.7%)	33.6(+141.7%)
CoT prompt	62.2(+15.6%)	42.9(+24.7%)	48.5(+26.2%)	32(+153.8%)
SC prompt	61.1(+17.7%)	45.3(+18.1%)	50.3(+21.7%)	41.8(+94.3%)
		with knowledge graph		
ToG	68.8(+4.5%)	50.2(+6.6%)	54.5*(+12.3%)	76.8(+5.7%)
LKLR	**71.9**	**53.5**	**61.2**	**81.2**

Table 1: The main result for baselines and LKLR on question answering. We use the OpenAI API to call GPT-3.5-turbo as the LLM, and the knowledge base used for models with external knowledge is Wikidata (Vrandecic and Krötzsch, 2014) except the result *, which is based on Freebase (Bollacker et al., 2008). The baseline results are from (Sun et al., 2023c). The best results are marked with bold, and the numbers in parentheses represent the proportion of improvement in the best result compared to that result.

ing previous choices. In contrast, LKLR follows a complete reasoning framework for solving the problem, making the connections between steps more coherent.

4.3 Main Results

The Table 1 presents the experimental results across multiple question-answering tasks and datasets, comparing baseline methods with our proposed LKLR. Notably, LKLR consistently outperforms baselines across tasks, showcasing its versatility and robust performance. The collaborative integration of LLMs and knowledge graphs within LKLR positions it as a powerful framework, delivering notable improvements across diverse reasoning tasks.

Our model exhibits a significant performance boost compared to models without the integration of external knowledge. The key enhancement stems from our method's adeptness at seamlessly combining the inferential capabilities of LLM with the. This fusion enables a more comprehensive understanding of natural language and a nuanced interpretation of complex queries. Furthermore, the integration of external knowledge plays a crucial role in addressing the limitations inherent in LLM. While LLMs excel in semantic reasoning, they often fall short in multi-step deductive reasoning, world knowledge validation, and reasoning about novel compositions of existing knowledge. By grounding the LLM's implicit reasoning chain in the knowledge graph through a logical query, LKLR overcomes these limitations. Each step of the reasoning process is aligned with factual knowledge from the KG, enhancing the model's ability to validate inferences and produce more reliable results.

Compared to the ToG method, our approach excels by strategically pre-planning reasoning paths from a holistic perspective. Unlike ToG's dynamic exploration, LKLR leverages the large language model's chain-of-thought reasoning to precompose a structured reasoning chain. This proactive approach ensures purposeful and directed reasoning, enhancing coherence and alignment with the overall question. The use of beam-search further enables adaptability. Overall, LKLR stands out for its strategic planning, providing a more focused and effective approach to multi-step reasoning compared to ToG's exploratory nature.

4.4 Analysis

We conducted an in-depth analysis of the knowledge effectiveness in our experimental results, providing quantitative evidence to demonstrate the tangible performance improvement achieved by our method. We specifically focused on the extent of knowledge acquisition and the impact of knowledge completeness on the experimental outcomes. Here we choose the QALD10-en, T-Rex, and WebQSP datasets. We do not use WebQuestions since WebQSP is based on it and has higher quality, which is more representative when conducting analysis.

In the process of grounding query, potential failures could occur, such as unsuccessful triplet extraction or the incompleteness of the knowledge

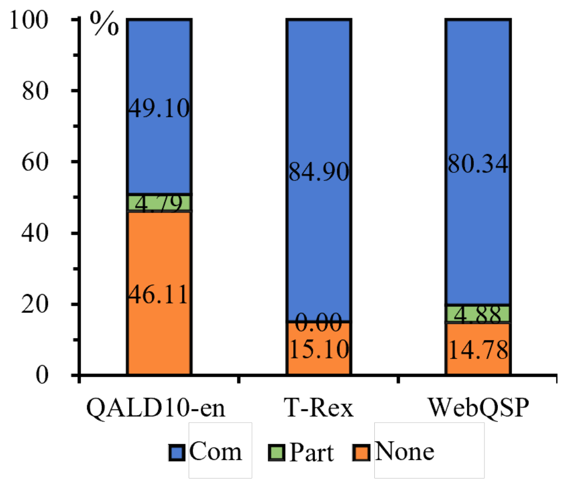


Figure 6: The proportion of different types of query grounding. Complete is denoted as Com.

graph, leading to the inability to obtain ideal knowledge relevant to the questions. We categorized the results based on the degree of query instantiation into three types: "Complete" signifies obtaining a logical query and fully instantiating it, acquiring the inference triplets; "Part" indicates obtaining a logical query but instantiating only a portion of it, resulting in partial knowledge relevant to the questions; the final category "None" is when it's impossible to extract appropriate queries or the logical queries cannot be instantiated, in which case we directly employ the CoT method for prediction. Figure 6 illustrates our statistical findings. We conducted statistical analyses on three datasets, revealing that the majority of the data aligns with our expectations, allowing us to obtain complete knowledge as anticipated. Instances where incomplete knowledge was obtained or no knowledge was acquired constitute a minority in our statistical analyses.

	Complete	**Part**	**None**	**Overall**
QALD10-en	**64.02**	62.50	40.91	53.50
T-Rex	**93.05**	-	14.57	81.20
WebQSP	**73.93**	70.59	61.17	71.90

Table 2: Accuracy of different types of query grounding.

Furthermore, we conducted a detailed analysis of the accuracy of answers in various scenarios, as shown in Table 2. It is evident that when complete knowledge is obtained, the accuracy of answers is consistently the highest across all datasets. This robustly demonstrates that our model's ability to acquire effective knowledge contributes significantly to enhancing overall performance. Moreover, even in scenarios where only partial knowledge is acquired, the accuracy of answers is notably higher than in situations where no knowledge is obtained at all. This aligns with our previous inference that, for generative models, accuracy in capturing partial facts can effectively elevate the overall accuracy of generated results. Future studies could explore further synergies between large models and structured knowledge, emphasizing the potential for more comprehensive investigations into their consistent integration.

4.5 Case Study

We selected two questions that appeared in the experiments for a case study as Table 3. The purpose of choosing these two examples is not only to demonstrate the effectiveness of the answers but also to showcase how our chosen knowledge can provide more informative responses. In the first case, our answer is more detailed, offering a finer granularity of information. In the second case, while the answer to the question itself has deficiencies, it can be considered correct according to evaluation metrics. However, in comparison to our answer, which provides intuitively accurate knowledge triplets, the combination of a large model's response and our answer is evidently more comprehensive and accurate. This enhances the persuasiveness of the response to the questioner.

5 Related Work

Recent research has demonstrated that prompting methods such as chain-of-thought (Wei et al., 2022; Wang et al., 2023; Kojima et al., 2022; Sun et al., 2023a; Besta et al., 2023; Sun et al., 2023b) can significantly enhance the reasoning capabilities of large language models. By constructing prompt sequences that encode reasoning logic and causal relationships, CoT provides step-by-step guidance for LLM prediction-making under few-shot learning settings. This allows LLMs to achieve improved performance on complex reasoning tasks.

However, while prompting strategies bolster the reasoning ability of LLMs, models prompted in this way still suffer from issues like hallucination and factual inconsistency (Ji et al., 2023) that stem from their reliance on implicit knowledge derived solely from training data (Lee et al., 2022). In order to further improve the reliability and factual

Q&A	Where does Michelle Pfeiffer live now? Orange County
CoT	Michelle Pfeiffer currently resides in the United States.
LKLR	*Query:* ("Q159778", "resides in", "#A")
	Knowledge: Michelle Pfeiffer, residence, Orange County Michelle Pfeiffer, country of citizenship, United States of America
	Answer: Based on the knowledge triplets, Michelle Pfeiffer resides in Orange County.
Q&A	When did carolina panthers go to superbowl? Super Bowl XXXVIII
CoT	The Carolina Panthers participated in the Super Bowl in two different years: 2004 and 2016.
LKLR	*Query:* ("Q330120", "participated in", "#A")
	Knowledge: Carolina Panthers, participating team, Super Bowl XXXVIII Carolina Panthers, participating team, Super Bowl 50
	Answer: The Carolina Panthers participated in Super Bowl XXXVIII and Super Bowl 50. Super Bowl XXXVIII took place on February 1, 2004, and Super Bowl 50 took place on February 7, 2016.

Table 3: Gnerated examples of CoT and LKLR. Note that we only represent the reasoning part of the answer, and the exact answer part to calculate the metric is not shown in the table.

accuracy of LLM responses, complementary approaches based on structured knowledge retrieval and evidence integration have been explored.

To alleviate the hallucination problem in LLMs, some approaches incorporate external knowledge (Yao et al., 2023; Sun et al., 2023c) to help generate more accurate responses. By searching for query-relevant information from external knowledge sources such as the Web and providing it as part of the prompt to the LLM, these methods are able to generate answers with higher correctness to some extent (Lu et al., 2023; Liu et al., 2023; Li et al., 2023b). However, while supplying additional query-relevant context can reduce the risks of hallucination, it lacks comprehensive correctness guarantees for multi-step reasoning tasks. Furthermore, simplistic search brings redundant information, hampering model inference. To further enhance reliability and factual consistency, recent work explores retrieve-after-generate paradigms that automatically filter (Gao et al., 2023; He et al., 2023; Peng et al., 2023; Zhao et al., 2023) or edit LLM outputs based on evidence from structured knowledge graphs (Li et al., 2023d; Guan et al., 2023; Li et al., 2023c; Baek et al., 2023). Integrating such structured external knowledge provides a way to refine model generations while mitigating hallucination and inaccuracies.

Our proposed approach builds on these insights to combine the benefits of reasoning-focused prompting and structured knowledge retrieval. We utilize knowledge graphs as dependable external knowledge sources to refine LLM responses after initial prompting-based generation. This allows us to reduce hallucinations and enhance factual consistency without extensive re-training. The integration of explicit reasoning guidance and structured external knowledge seeks to complement the strengths of pre-trained LLMs, addressing the limitations of previous work.

6 Conclusion

In this work, we proposed a novel framework that entangles large language models and knowledge graphs for advanced logical reasoning. The key innovation is transforming implicit reasoning chains into executable logical queries, enabling multi-hop inference grounded in structured knowledge. This addresses the limitations of both neural and symbolic approaches. We demonstrate significant performance improvements on multiple logical reasoning datasets, including multi-hop QA, open-domain QA, and slot-filling tasks. The framework represents an important advance toward more robust reasoning in AI by combining neural creativity with logical validation. Further directions include extending the approach to broader knowledge sources and more complex reasoning. We believe this synergistic reasoning paradigm will open new frontiers in artificial intelligence.

Limitation

While our proposed framework makes significant progress in integrating neural and symbolic reasoning, it has some limitations that could be addressed in future work. A key limitation is that the entities involved in the reasoning must exist in the KG. If the reasoning contains entities not present in the KG, then relevant knowledge cannot be provided for such questions. This places requirements on the completeness of the KG for broad reasoning coverage. Future work could focus on techniques to handle reasoning about unknown entities, such as searching external sources or generating plausible knowledge. Additionally, our current framework relies on a single knowledge graph, while combining multiple heterogeneous knowledge sources could provide more diverse reasoning capabilities.

References

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *CoRR*, abs/2306.04136.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of thoughts: Solving elaborate problems with large language models. *CoRR*, abs/2308.09687.

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, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528.

Hady ElSahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: researching and revising what language models say, using language models. 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 16477–16508. Association for Computational Linguistics.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2023. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. *CoRR*, abs/2311.13314.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Rethinking with retrieval: Faithful large language model inference. *CoRR*, abs/2301.00303.

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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. 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*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. 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 8424–8445. Association for Computational Linguistics.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *CoRR*, abs/2304.11633.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jingyuan Wang, Jian-Yun Nie, and Ji-Rong Wen. 2023b. The web can be your oyster for improving large language models. *CoRR*, abs/2305.10998.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023c. 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.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq R. Joty, and Soujanya Poria. 2023d. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *CoRR*, abs/2305.13269.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 4549–4560. ACM.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *CoRR*, abs/2304.09842.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, abs/2302.12813.

Aleksandr Perevalov, Dennis Diefenbach, Ricardo Usbeck, and Andreas Both. 2022. Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers. In *16th IEEE International Conference on Semantic Computing, ICSC 2022, Laguna Hills, CA, USA, January 26-28, 2022*, pages 229–234. IEEE.

Hongyu Ren and Jure Leskovec. 2020. Beta embeddings for multi-hop logical reasoning in knowledge graphs. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiashuo Sun, Yi Luo, Yeyun Gong, Chen Lin, Yelong Shen, Jian Guo, and Nan Duan. 2023a. Enhancing chain-of-thoughts prompting with iterative bootstrapping in large language models. *CoRR*, abs/2304.11657.

Jiashuo Sun, Yi Luo, Yeyun Gong, Chen Lin, Yelong Shen, Jian Guo, and Nan Duan. 2023b. Enhancing chain-of-thoughts prompting with iterative bootstrapping in large language models. *CoRR*, abs/2304.11657.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023c. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. *CoRR*, abs/2307.07697.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

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*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. 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 5823–5840. Association for Computational Linguistics.