# LLM4QA: Leveraging Large Language Model for Efficient Knowledge Graph Reasoning with SPARQL Query

Mingjing Lan, Yi Xia *, Gang Zhou, Ningbo Huang, Zhufeng Li, and Hao Wu
State Key Laboratory of Mathematical Engineering and Advanced Computing, Information Engineering University, Zhengzhou, China
Email: lanmingjing@126.com (M.L.); summer_one520@163.com (Y.X.); gzhougzhou@126.com (G.Z.); rylynn_ab@163.com (N.H.); 20086538@qq.com (Z.L.); wuhao186000@163.com (H.W.)
*Corresponding author

***Abstract*—As one of the core technologies of general artificial intelligence, knowledge graph reasoning aims to infer new knowledge from existing knowledge in the knowledge base, providing decision support for knowledge-driven intelligent information services such as information retrieval, question answering, and recommendation systems. However, there are still some issues, such as poor interpretability and low reasoning efficiency, always decrease the current knowledge reasoning performance. To tackle the challenges, this paper proposes a knowledge graph reasoning method *LLM4QA*, which leverages fine-tuned large language models with chain-of-thought to generate graph query languages SPARQL (i.e., SPARQL Protocol and RDF Query Language) for reasoning. Firstly, an efficient instruction fine-tuning method is applied to fine-tune open-source large language models with chain-of-thought. Then, the fine-tuned open-source large model is used to convert natural language questions into logical forms. Finally, we utilize unsupervised entity relationship retrieval to generate graph database query languages, realizing a natural language knowledge graph question-answering framework. Experimental results demonstrate that this method achieves well performance in terms of inference accuracy and significantly improves model retrieval efficiency.**

***Keywords*—Large Language Model(LLM),knowledge graph, question answering system, chain of thought**

## I. Introduction

Knowledge reasoning is an indispensable part of human intelligence and a core challenge in artificial intelligence systems. Its applications are wide-ranging, and further advancements have significant impacts on various domains such as question answering systems, recommender systems, and other information retrieval systems [1].

Traditional knowledge reasoning has primarily focused on symbolic reasoning, which involves using explicit symbols such as ontologies and logical rules for inferencing knowledge. In recent years, neural network systems, which are represented by Large Language Models (LLMs) have achieved great success in tasks across different domains. In the field of natural language processing, neural network systems, such as ChatGPT, trained on massive corpora, have demonstrated excellent performance across various tasks, showcasing their potential for general artificial intelligence.

While large-scale models have demonstrated excellent performance in certain reasoning tasks [2], they often suffer from the issue of being uninterpretable, making it inconvenient for people to analyze and understand the reasoning process and outcomes. The lack of interpretability in models significantly affects the reliability of the model's decisions, leading to widespread concern about their reliability and robustness. Particularly in crucial applications areas such as national defense, healthcare, and law, ensuring transparency and interpretability in the decisions made by systems is essential and necessary.

In recent years, Knowledge Graphs (KG) arise great concern among researchers [3]. Knowledge graph can achieve data integration from different sources into a unified structure, linking multi-source information through nodes and relationships to form a network of relationships, providing interpretable factors such as concepts, relationships, and properties for various reasoning tasks in the real world [4]. Therefore, knowledge reasoning techniques based on knowledge graphs have gradually garnered increasing attention among scholars, becoming one of the core technologies in the field of cognitive intelligence. Some recent work leverage sequence-to-sequence models to conduct semantic parsing over knowledge graphs [5]. The work convert natural language to the corresponding logical forms and executable graph query, and then apply them in the knowledge graph environment. However, generating the corresponding long logical forms for more complex questions is a challenge, symbolic knowledge reasoning itself lacks robustness and is sensitive to noise in the data. Therefore, in this work, we further apply chain-of-thought to decompose natural language questions into simpler logical forms for better performance.

This paper aims to apply large language models to achieve knowledge reasoning through semantic parsing. Firstly, an open-source large language model is fine-tuned using an efficient instruction-based fine-tuning method. Then, the fine-tuned open-source large model is used to convert natural language questions into logical forms. Finally, an unsupervised entity relationship retrieval model is utilized to generate graph query language SPARQL (i.e., SPARQL Protocol and RDF Query Language), which can be executed over knowledge graphs, realizing a natural language knowledge graph question-answering framework. Experimental results show that this method achieves the best performance in terms of reasoning accuracy and

Manuscript received June 7, 2024; revised June 25, 2024; accepted July 4, 2024; published October 21, 2024.

significantly improves the model retrieval efficiency.

We organize the paper as follows. The background and related work are illustrated in Section II. After that, we describe the proposed model in Section III, and further detail the experimental study and relevant analysis in Section IV. Finally, in Section V, we make a concludion and look forward the further work on our paper.

## II. Background and Related Work

Two basic and vital definitions are given as follows.

***Definition (Knowledge Graph).*** A knowledge graph $\mathcal{KG} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}\}$, is a set of facts stored as triples of the form $(e_s, r_q, e_o)$, where $\mathcal{E}$ denotes the entity set, $e_s$ and $e_o \in \mathcal{E}$. $\mathcal{R}$ denotes the set of relations, $r_q \in \mathcal{R}$. $\mathcal{T} = \{(e_s, r_q, e_o)\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ denotes the triple set.

***Definition (Logic Form).*** The Logic form refers to the abstract structure or representation of a natural language statement that captures its logical meaning independently. It is used in formal semantics and logic to analyze and understand the underlying logical relationships within sentences or propositions.

There have been many approaches proposed for knowledge graph reasoning, which aim to infer missing facts according to existing factual triples [6]; these could generally be classified into three categories: embedding-based methods, logic rule-based methods, and path-based methods. Recent advances and their features are briefly reviewed here.

Embedding-based methods, known as representation learning on KGs [4], [7], learn representations for entities and relations in KGs, and reasons from the representations. The methods model various relational patterns via designing different loss functions. Logic rule-based methods extract logical rules from facts in the KG [5], [8], and reason unknown facts based on the extracted rules. The methods could incorporate various domain knowledge, thus providing well explanations for the reasoning outcomes. Path-based reasoning methods intend to conduct efficient path exploration in KGs, and then learn the path co-occurrence patterns to reason the result. Some recent approaches model the problem as the multi-hop reasoning task applying Reinforcement Learning (RL) [9], [10]. This approach can not only efficiently obtain reasoning results, but also provide intermediate paths to indicate the reasoning process.

Many current knowledge reasoning methods are evaluated through downstream question-answering tasks, which can better test the model's understanding of natural language and facilitate the integration of external database knowledge for downstream tasks. Therefore, this paper primarily assesses the model's reasoning capabilities through knowledge graph question-answering tasks. Zhang *et al*. [11] proposed a subgraph retrieval method, which designs a subgraph retriever decoupled from the inference process to efficiently retrieve relevant subgraphs for questions. Combined with a subgraph-oriented reasoner, this approach allows the model to focus on more relevant and smaller-scale subgraphs, thereby enhancing the model's inference and question-answering performance. Shu *et al*. [12] improved the robustness of pretrained language models in various generalization scenarios and enhanced knowledge base question-answering performance by retrieving knowledge base content at multiple granularities, including entities, logical forms, and architectural items. Lan *et al*. [13] introduced a method called QGG, which first generates query graphs with good scalability for multi-hop questions, combined with constraints, to strengthen the ability to answer complex questions. Bhutani *et al*. [14] proposed a divide-and-conquer approach to dealing with questions, breaking down complex questions into multiple simple queries and using a semantic matching model combined with the results of subqueries to improve the reasoning performance for complex questions.

Some researchers have employed semantic parsing methods, utilizing strategies such as step-by-step query graph generation and search. Liu *et al.* [15] proposed a unified modeling approach for semantic parsing in question answering directed at both knowledge bases and databases through three modules: primitive enumeration, ranking, and combination. Jiang *et al*. [16] introduced the UniKGQA method, which implemented a unified architecture for retrieval and reasoning over knowledge graphs. This architecture, combining a semantic matching module and an information propagation module, was further enhanced by pre-training and fine-tuning strategies, leading to improved QA performance and enabling collaborative multi-hop knowledge QA. Others have adopted sequence-to-sequence models to generate S-expressions and provided various enhancements for the semantic parsing process. Yu *et al*. [17] proposed DECAF, a method that integrates retrieval results from knowledge bases with a sequence-to-sequence framework to generate answers to questions, thereby facilitating the generation of logical forms and QA inference.

Large language models have recently shown advanced in-context learning ability through pretraining on large-scale corpora. Recent works have utilized large language models as a tool to enhance various methods [18], [19]. Therefore, we proposes a method that enhances knowledge reasoning with LLM. We convert natural language question to logic SPARQL query, and then execute over knowledge graph to retain the answer. By leveraging LLM to generate logical queries, the reasoning outcome can be effectively retrieved from a knowledge graph in the form of SPARQL queries, which perform effective knowledge graph reasoning.

## III. Our Proposed Method

Our model is generally a generate-then-retrival knowledge reasoning framework. We leverage fine-tuned LLMs to convert the query to logical forms, and then execute it over knowledge graphs to obtain the answers.

This section begins with a formal definition of the proposed reasoning method, introducing relevant problem definition. Subsequently, a detailed presentation is provided on the overall framework of the proposed model and its various innovative aspects. Finally, an overview of the overall process of the model and the execution of graph query language is presented.

### A. Problem Definition

Given a query $q$ and knowledge graph $G$, the purpose of our reasoning for KGs is to convert the given query $q$ to the logic form $f$, and then we convert it to the equivalent

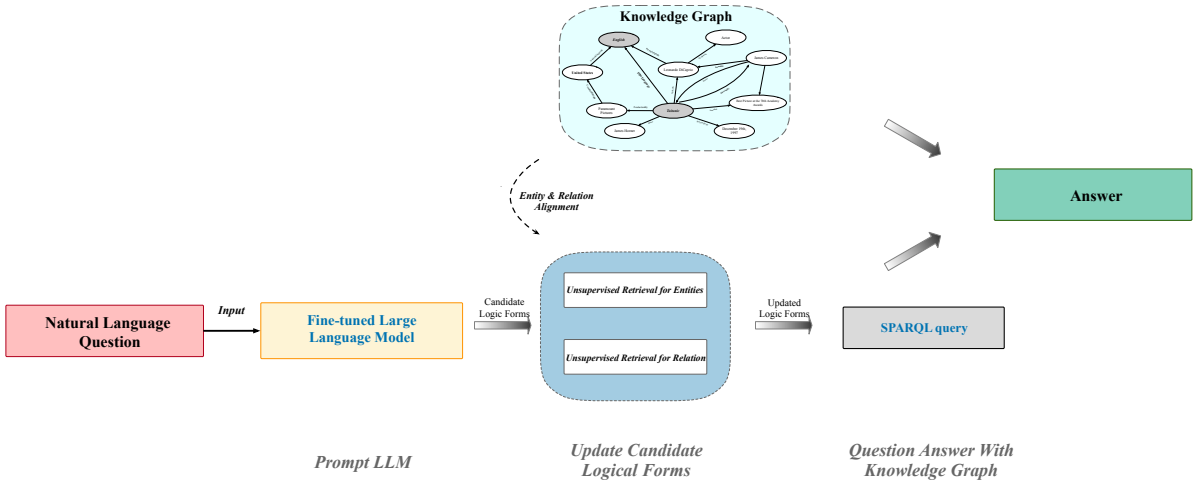


Fig. 1. Illustration of the Fine-tuned Large Language Model Reasoning Process over Knowledge Graphs.

SPARQL query. At last, the SPARQL query is executed over knowledge graph $G$ to seek a set of answer $E_o$, where $e_o \in E_o$.

### B. Efficent Fine-Tunning on Open-Source LLMs

In this work, we need to efficiently fine-tune the open-source large-scale models, so that these models can generate the corresponding logical forms as efficiently and effectively as possible.

Firstly, we need to create the training data for instruction fine-tuning. *LLM4QA* initially transforms the SPARQL queries associated with the natural language questions from the training set in the KBQA dataset into corresponding logical forms. However, in the knowledge question-answering dataset, entities described in natural language questions are represented by ID numbers, which present challenges for large language models adept at natural language understanding, thereby impacting performance. Therefore, we then substitute the entity IDs in the generated logical forms with corresponding entity labels, to facilitate a deeper understanding of the specific meanings of entities by the large-scale models.

Take an example of the above process. When we encounter the natural language question *'The national anthem Afghan National Anthem is from the country which practices what religions?'*, the fine-tuned model can onvert it to the logic form *'( JOIN ( R [ location , religion percentage , religion ] ) ( JOIN ( R [ location , statistical region , religions ] ) ( JOIN [ location , country , national anthem ] ( JOIN [ government , national anthem of a country , anthem ] [ Afghan National Anthem ] ) ) ) )'*, and then we align the entities in the knowledge graph and convet the ID number to corresponding entity. The aligned query can then execute over knowledge graphs to obtain the answer.

Further experimental results indicate that *LLM4QA* performs remarkably well in simple tasks such as one-hop and two-hop queries, mainly due to the similarity in the involved logical forms, which can be effectively generated by finely-tuned LLMs [19]. However, generating long logical forms corresponding to more complex queries presents a challenge. Therefore, we leverage Chain of Thought (CoT) technique to decompose and refine the resolution of natural language queries, aiming for improving complex question answering performance. We adopted the method from the KD-COT [20] to construct the COT dataset.

Finally, *LLM4QA* adopts the Parameter Efficient Fine-Tuning (PEFT) technique to fine-tune the LLM, which reduce the cost of fine-tuning LLMs with a small number of parameters. PEFT includes various methods such as LoRA, QLoRA, P-tuning v2, and Freeze. Additionally, *LLM4QA* can seamlessly switch between several state-of-the-art open-source LLMs, including Llama-2-7B, ChatGLM2-6B, and Baichuan2-7B.

### C. Prompt Fine-Tuned LLMs with Chain-of-Thought

By means of fine-tuning, the large-scale model has acquired a certain level of proficiency in semantic parsing, enabling it to somewhat transform natural language queries into the logical format of graph query language. Next, we will employ the fine-tuned large language model for semantic parsing on new queries in the test set.

As shown in the figure 1, we input the natural language to the fine-tuned large language model, and the fine-tuned LLMs then generate corresponding logic form. However, we noted that many of the generated entities and relations do not correspond to those in the knowledge graph, which could significantly impact the performance of reasoning in the KG. Therefore, prior to executing it over the knowledge graph, we align the entities and relations in the logic form with those in the knowledge graph. We apply the state-of-the-art unsupervised method SimCSE [21] to identify the most semantically similar candidates for the logic form. Furthermore, we update the logic form and execute it over knowledge graph to obatain the final answer. Compared with the state-of-the-art method ChatKBQA [19], we leverage the context about the anchor entity in knowledge graph, and construt more chain of thought example for in-context learning. The improvement can make our proposed method tackle with complex question more effectively.

## IV. Experiments

This section begins with an explanation of the dataset, baseline models, evaluation metrics settings used in the proposed reasoning method. Subsequently, a detailed presentation of the experimental reasoning results of the proposed

TABLE I
DATASETS AND THEIR STATISTICS.

| Datasets | #Question | #Entity | #Relation | #Train | #Valid | #Test |
|---|---|---|---|---|---|---|
| WebQSP | 4,737 | 2,461 | 628 | 3,098 | – | 1,639 |
| CWQ | 34,689 | 11,422 | 845 | 27,639 | 3,519 | 3,531 |

TABLE II
COMPARISON OF THE LINK PREDICTION RESULTS OF *LLM4QA* WITH OTHER BASELINES ON KBQA DATASETS.

| Model | WebQSP | | | CWQ | | |
|---|---|---|---|---|---|---|
| | ACC | Hit@1 | F1 | ACC | Hit@1 | F1 |
| Subgraph Retrival* [11] | – | 69.5 | 64.1 | – | 50.2 | 47.1 |
| Topic Units [7] | – | 68.2 | 67.9 | – | 39.3 | 36.5 |
| QGG [13] | – | 73.0 | 74.0 | – | 44.1 | 40.4 |
| UniKGQA* [16] | – | 77.2 | 72.2 | - | 51.2 | 49.4 |
| CBR-KBQA [22] | 69.6 | – | 72.8 | 67.1 | 70.4 | 70.0 |
| RnG-KBQA [23] | 71.1 | – | 75.6 | – | – | – |
| Program Transfer* [5] | – | 74.6 | 76.5 | – | 58.1 | 58.7 |
| GMT-KBQA [8] | 73.1 | – | 76.6 | 72.2 | – | 77.0 |
| DECAF [17] | – | 82.1 | 78.8 | – | 70.4 | – |
| HGNet [2] | 70.7 | 76.9 | 76.6 | 57.8 | 68.9 | 68.5 |
| SructGPT* [18] | – | – | 72.6 | – | – | – |
| ToG* [10] | – | 82.6 | – | – | 69.5 | – |
| ChatKBQA [19] | 73.8 | 83.2 | 79.8 | 73.3 | 82.7 | 77.8 |
| ChatKBQA* [19] | 77.8 | 86.4 | 83.5 | **76.8** | **86.0** | **81.3** |
| Ours (*LLM4QA*) | 74.1 | 83.2 | 79.7 | 73.1 | 85.3 | 77.6 |
| Ours (*LLM4QA**) | **78.0** | **86.9** | **84.1** | 76.5 | 85.8 | 81.0 |

method is provided, along with the comparison with other baseline models.

### A. Experimental Setup

We adopted five typical datasets for the training and evaluation of our *LLM4QA*, and the basic statistics are given in Table I.

To evaluate our model in terms of link prediction, we conducted experiments on the WebQSP and CWQ dataset. We applied the proportion of correct tail entity rankings in the top K (Hits@K) and Accuracy (Acc) as evaluation protocol. Meanwhile, we apply Llama-2-7b as our backbone LLM in Table II.

### B. Experimental Results

We conduct the KBQA experiment for three real-world datasets extracted from Freebase. The results of the models are mainly taken from their original paper. For our proposed method, we respectively display the results on WebQSP and CWQ. The best results are in bold. '*' in the Table II denotes applying the datasets entity annotation. It can be observed from the Table II that:

- **(i)** The *LLM4QA* framework exhibits a commendable track record, consistently delivering performances that are not only on par but often surpass those of its contemporaries across a multitude of evaluation criteria, with a particularly pronounced dominance observed on the WebQSP benchmark. This comparative advantage of *LLM4QA* over state-of-the-art reasoning models is notably gratifying, as it manifests a tangible enhancement in critical metrics. Specifically, when juxtaposed against the leading baseline model, *LLM4QA* achieves an average improvement of 2% in Accuracy (ACC), 0.8% in Hit@1, and a further 0.8% in F1 score. These increments, albeit seemingly modest at first glance, bear testament to the efficacy and refinement of our proposed methodology, underscoring its potential to significantly contribute to the advancement of question answering systems and knowledge base reasoning techniques. The systematic outperformance across these pivotal indicators substantiates the robustness and superiority of *LLM4QA*, positioning it as a formidable contender in the quest for enhanced reasoning capabilities within complex information landscapes.
- **(ii)** While *LLM4QA* has demonstrably achieved performances that are not merely competitive but frequently superior to those of preceding models, a pivotal strength that sets it apart lies in its unwavering commitment to maintaining interpretability throughout the reasoning process. This characteristic is emblematic of the model's effectiveness, underscoring its ability to deliver not just accurate outcomes but also insights that are accessible and comprehensible. A distinguishing feature of our approach is its capacity to generate an interpretable reasoning sequence via the utilization of SPARQL queries. This methodology engenders a level of transparency that is unparalleled, enabling users to trace the logical progression of the system's thought process with clarity and precision. The transparency afforded by this approach not only enhances trust in the model's decisions but also facilitates a deeper understanding of its operational mechanisms. Furthermore, the recourse to SPARQL queries enables a systematic approach to error analysis. In instances where the reasoning process yields unexpected results, the generated query serves as a roadmap, allowing for the meticulous backtracking of each step. This capability is invaluable, as it ensures that any discrepancies or inaccuracies can be identified and rectified with greater ease, making the entire process not only more plausible but also more acceptable to human evaluators. Such a meticulous attention to detail and the pursuit of interpretability are hallmarks of our dedication to advancing the field of machine learning while upholding the principles of clarity and accountability.

- **(iii)** The statistical assessment conducted further fortifies our findings, revealing that all link prediction outcomes have achieved a confidence interval of no less than 94% (±0.7) for the pertinent measurement metrics. This quantifiable affirmation underscores the remarkable consistency of our experimental results from a statistical standpoint, lending substantial credibility to the robustness and reliability of our methodology. It substantiates the hypothesis that the observed outcomes are not mere anomalies but rather, indicative of a consistent trend that holds firm under rigorous scrutiny, thus bolstering the validity of our conclusions within the realm of knowledge base completion.

### C. Ablation Analysis

In the work, *LLM4QA* uses two strategies for empowering reasoning over KGs, which denote fintune LLM with chain of thought (KG_COT) and unsupervised retrival for logical form (KG_retrival). We conducted several ablation studies that removed either the KG_COT (only open-source LLM was used) strategy or KG_retrival strategy, to evaluate their contributions to *LLM4QA*. These dual strategies are pivotal in augmenting the model's ability to reason effectively over the vast and complex data contained within KGs.

To rigorously assess the individual contributions of these strategies to *LLM4QA* overall performance, we conducted a series of ablation studies. In these experiments, we systematically excluded either the KG_COT strategy leaving the model reliant solely on its open source large language model capabilities or the KG_retrieval strategy, thereby isolating the impact of each component. Our findings reveal a clear decline in the model's question answering performance upon the removal of either strategy, which decreases the accuracy performance 5.2% and 4.1% respectively. This empirical evidence underscores the indispensable role played by both the KG_COT and KG_retrieval mechanisms in bolstering *LLM4QA* functionality.

Meanwhile, it can be observed that removing either KG_COT or KG_retrival decreases the question answering performance of the model, which demonstrates that both our proposed strategies contribute to *LLM4QA*. Removing KG_COT strategy have a more significant drop than removing KG_retrival strategy, suggesting that KG_COT strategy is more critical to *LLM4QA*. The chain of thought approach not only facilitates a more coherent and explicable reasoning process but also emerges as a critical factor in elevating the model's question answering prowess. It also proves that the chain of thought not only enables the reasoning process more reasonable and explainable but also plays an important role in question answering performance. We can also observe that in comparison with the baseline model ChatKBQA [24], both our proposed strategies make *LLM4QA* obtain a better performance, which indicates the effectiveness of our proposed strategies.

## V. Conclusion

In this study, we proposed a generate-then-retrival model *LLM4QA* for knowledge graph reasoning. Previous reasoning methods suffer from low retrieval efficiency and misleading generation in current knowledge reasoning processes, and the trustworthiness and reliability of the knowledge discover outcomes often thus interferes with human decision making. To overcome these obstacles, our proposed reasoning model *LLM4QA* leverage large language model to generate logical queries, and then the reasoning outcome can be effectively retrieved from a knowledge graph in the form of SPARQL queries. We convert natural language question to logic SPARQL query, and then execute over knowledge graph to retain the answer.

The experimental results indicate that our model performed better than existing reasoning models in terms of efficiency and effectiveness while retaining the advantages of interpretability and trackability. However, the methods proposed in this paper have not leveraged the intrinsic structural knowledge within the knowledge graph, a component that has the potential to significantly augment the model's performance. Overlooking these structural nuances may result in a loss of valuable insights and capabilities, indicating a direction for future enhancements where incorporating such structural information could lead to more refined and effective outcomes. For future work, we would like to integrate knowledge graph structural information into our model, which can boost our model with more structural pattern learning ability.

## Conflict of Interest

The authors declare no conflict of interest.

## Author Contributions

Mingjing Lan: Data processing, Original draft preparation. Yi Xia: Conceptualization, Methodology, Programming. Gang Zhou, Ningbo Huang: Visualization, Investigation. Zhufeng Li: Supervision. Hao Wu: Software, Validation. All authors had approved the final version.

## Funding

The authors would like to express appreciation sincerely to all the anonymous reviewers for taking their time to provide constructive comments, which are vital to improve our work. The work was supported by the National Natural Science Foundation of China (42371438) and the Science and technology project of Henan Province (222102210081, 222300420590).

## References

[1] Y. Xia, M. Lan, J. Luo, X. Chen, and G. Zhou, "Iterative rule-guided reasoning over sparse knowledge graphs with deep reinforcement learning," *Information Processing & Management*, vol. 59, no. 5, p. 103040, 2022.

[2] Y. Chen, H. Li, G. Qi, T. Wu, and T. Wang, "Outlining and filling: Hierarchical query graph generation for answering complex questions over knowledge graphs," *IEEE Transauction on Knowledge Data Engineering*, vol. 35, no. 8, pp. 8343–8357, 2023.

[3] Y. Xia, J. Luo, M. Lan, and G. Zhou, "Reason more like human: Incorporating meta information into hierarchical reinforcement learning for knowledge graph reasoning," *Applied Intelligence*, vol. 53, p. 13293–13308, 2023.

[4] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS 2013 - Volume 2*. Red Hook, NY, USA: Curran Associates Inc., 2013, p. 2787–2795.

[5] S. Cao, J. Shi, Z. Yao, X. Lv, J. Yu, L. Hou, J. Li, Z. Liu, and J. Xiao, "Program transfer for answering complex questions over knowledge bases," 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*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 8128–8140.

[6] Y. Li, X. Zhang, B. Zhang, and H. Ren, "Each snapshot to each space: Space adaptation for temporal knowledge graph completion," in *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings*, ser. Lecture Notes in Computer Science, U. Sattler, A. Hogan, C. M. Keet, V. Presutti, J. P. A. Almeida, H. Takeda, P. Monnin, G. Pirrò, and C. d'Amato, Eds., vol. 13489. Springer, 2022, pp. 248–266.

[7] Y. Lan, S. Wang, and J. Jiang, "Knowledge base question answering with topic units," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019, pp. 5046–5052.

[8] X. Hu, X. Wu, Y. Shu, and Y. Qu, "Logical form generation via multi-task learning for complex question answering over knowledge bases," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, N. Calzolari, C. Huang, H. Kim, J. Pustejovsky, L. Wanner, K. Choi, P. Ryu, H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S. Na, Eds. International Committee on Computational Linguistics, 2022, pp. 1687–1696.

[9] Y. Xia, J. Luo, G. Zhou, M. Lan, X. Chen, and J. Chen, "Dt4kgr: Decision transformer for fast and effective multi-hop reasoning over knowledge graphs," *Information Processing & Management*, vol. 61, no. 3, p. 103648, 2024.

[10] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, H. Shum, and J. Guo, "Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph," *CoRR*, vol. abs/2307.07697, 2023.

[11] J. Zhang, X. Zhang, J. Yu, J. Tang, J. Tang, C. Li, and H. Chen, "Subgraph retrieval enhanced model for multi-hop 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*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 5773–5784.

[12] Y. Shu, Z. Yu, Y. Li, B. F. Karlsson, T. Ma, Y. Qu, and C. Lin, "TIARA: multi-grained retrieval for robust question answering over large knowledge base," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, 2022, pp. 8108–8121.

[13] Y. Lan and J. Jiang, "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*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 969–974.

[14] N. Bhutani, X. Zheng, and H. V. Jagadish, "Learning to answer complex questions over knowledge bases with query composition," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, W. Zhu, D. Tao, X. Cheng, P. Cui, E. A. Rundensteiner, D. Carmel, Q. He, and J. X. Yu, Eds. ACM, 2019, pp. 739–748.

[15] Y. Liu, S. Yavuz, R. Meng, D. Radev, C. Xiong, and Y. Zhou, "Uniparser: Unified semantic parser for question answering on knowledge base and database," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, 2022, pp. 8858–8869.

[16] J. Jiang, K. Zhou, X. Zhao, and J. Wen, "Unikgqa: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[17] D. Yu, S. Zhang, P. Ng, H. Zhu, A. H. Li, J. Wang, Y. Hu, W. Y. Wang, Z. Wang, and B. Xiang, "Decaf: Joint decoding of answers and logical forms for question answering over knowledge bases," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[18] J. Jiang, K. Zhou, Z. Dong, K. Ye, X. Zhao, and J. Wen, "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*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 9237–9251.

[19] H. Luo, H. E, Z. Tang, S. Peng, Y. Guo, W. Zhang, C. Ma, G. Dong, M. Song, and W. Lin, "Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models," *CoRR*, vol. abs/2310.08975, 2023.

[20] K. Wang, F. Duan, S. Wang, P. Li, Y. Xian, C. Yin, W. Rong, and Z. Xiong, "Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering," *CoRR*, vol. abs/2308.13259, 2023.

[21] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 6894–6910.

[22] R. Das, M. Zaheer, D. Thai, A. Godbole, E. Perez, J. Y. Lee, L. Tan, L. Polymenakos, and A. McCallum, "Case-based reasoning for natural language queries over knowledge bases," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 9594–9611.

[23] X. Ye, S. Yavuz, K. Hashimoto, Y. Zhou, and C. Xiong, "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*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 6032–6043.

[24] R. Das, S. Dhuliawala, M. Zaheer, L. Vilnis, I. Durugkar, A. Krishnamurthy, A. Smola, and A. McCallum, "Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning," in *the 6th International Conference on Learning Representations, ICLR 2018, April 30 - May 3, 2018, Conference Track Proceedings*. Vancouver, BC, Canada: OpenReview.net, 2018.