Few-Shot Knowledge Graph Question Generation via LLM Abstraction-to-Instantiation Conversion and Small Model Collaboration

Junze Tan[1], Runhao Zhao[1], Jiuyang Tang[1], and Lei He[2]

[1] Laboratory for Big Data and Decision, National University of Defense Technology
{tanjunze,runhaozhao,jiuyang_tang}@nudt.edu.cn
[2] College of System Engineering, National University of Defense Technology
helei@nudt.edu.cn

Abstract. Knowledge graph question generation (KGQG) refers to the task of generating natural language questions from knowledge graphs(KGs). Although this problem has been deeply studied in the past few years, at present, the training of small models heavily relies on a large amount of labeled data, and large language models (LLMs) also require a large number of parameters and high training costs. In reality, a significant amount of time and financial resources are required for manual annotation, which results in a scarcity of annotated data in practice. Therefore, the methods of these models are not very realistic. To address the above problems, we propose the Few-Shot Knowledge Graph Question Generation via LLMs Abstraction-to-Instantiation Conversion and Small Models Collaboration model (AIC-KGQG), which automatically generates labeled data through few-shot examples and collaborates between large and small model. We compare this framework with four mainstream knowledge graph question generation methods. The results show that AIC-KGQG achieves state-of-the-art performance under few-shot conditions, balancing data dependency and deployment efficiency, and providing a practical solution for KGQG in resource-constrained fields.

Keywords: KGQG · LLMs · Abstraction-to-Instantiation Conversion

1 Introduction

The knowledge graph question generation (KGQG) task is to generate logically coherent natural language questions based on the given knowledge graph. Question generation effectively addresses the scarcity of high-quality question-answering datasets in specific domains [6], avoids expensive and time-consuming manual annotation, and aids in enhancing the efficiency of question-answering systems, making it a vital component of the knowledge mining field.

Significant progress has already been made in the research on KGQG in the current field [3–5, 8, 11, 18, 22]. Specifically, the template based approach relies on

pre-designed rules and has relatively poor generalization ability. With the development of deep learning technologies, methods based on neural networks have been extensively studied. SATHISH proposed a model named KG2Question-RNN that can generate simple questions. The G2S model has achieved significant improvement in generation quality through training with a large amount of labeled data.

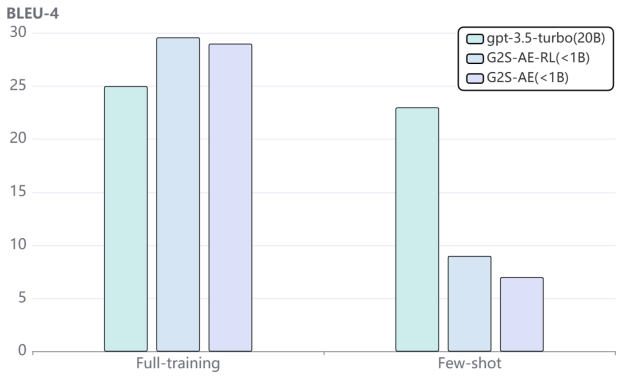

Fig. 1. The performance of three advanced KGQG methods on the WebQuestions dataset under two annotation data scale conditions: full-training and few-shot. "Full-training" refers to training with all annotated data, while "few-shot" means less than 5% of the annotated data.

Most existing KGQG models place high demands on the quality and scale of datasets [13]. When the sample size is insufficient, KGQG performance tends to collapse significantly (see Fig. 1). The stark difference in KGQG performance between full-sample and few-shot scenarios highlights its heavy dependence on high-quality datasets. However, in some domains, high-quality datasets are scarce, and manual annotation incurs substantial time and financial costs. Therefore, developing methods for automatically generating annotated data under few-shot conditions is of utmost importance.

Recent studies have shown that LLMs like ChatGPT [16], due to their strong generalization capabilities, have demonstrated outstanding performance in generation, driving progress in the KGQG field [15]. As shown in Figure 1, LLMs can also exhibit good performance under few-shot conditions (see Fig. 1) [24]. However, LLMs require a large amount of training data, have a huge number of parameters, have high deployment costs, and overly rely on high computing resources. Therefore, exploring lightweight deployment solutions is of crucial importance [19].

To address the above challenges, we propose a framework for training small models using question-answering pairs generated by LLMs under few-shot constraints. Drawing inspiration from the chain-of-thought approach, we introduce an abstraction-to-instantiation conversion mechanism to guide LLMs through

staged reasoning for extracting key entity names and knowledge graph relation paths. Based on the information extracted during the prompting process, LLMs instantiate natural language questions to form question-answering datasets. By training small models with these generated datasets, we effectively combine the generalization and generation capabilities of LLMs with the lightweight deployment advantages of small models, aiming to improve the quality of generated questions.

To validate the effectiveness of AIC-KGQG, we compared it with multiple state-of-the-art models. Experimental results demonstrate that AIC-KGQG performs remarkably well under few-shot conditions, essentially matching the performance of state-of-the-art models.

The primary contributions of this article are as follows: (1) We propose AIC-KGQG, a KGQG framework based on the abstraction-to-instantiation prompting mechanism, aiming to guide LLMs through step-by-step thinking to reason and generate high-quality questions for training small models. Through innovations in the chain of thought, we significantly reduce reasoning costs and improve generation quality. (2) We attempt to collaborate LLMs with small models, which not only reduces KGQG's dependency on large-scale training data but also lowers training and deployment costs. (3) Under few-shot conditions, AIC-KGQG significantly outperforms all baseline models and approaches the performance of state-of-the-art models under full-sample conditions, demonstrating the broad prospects of this research.

2 Related works

Knowledge graph question generation Early KGQG research was implemented through manually predefined rules [3, 5, 8, 18]. The SPARQL-based rule parsing framework in [11] generates query statements by matching entity and relation templates in questions. Such methods exhibit weak generalization capabilities and struggle to achieve cross-domain expansion. In recent years, Transformer architectures based on attention mechanisms have become mainstream. Vaswani et al. [22] utilized self-attention mechanisms to capture global semantic information, dynamically focusing on key details to generate high-quality queries. With the proposal of meta-learning frameworks, the DSM model [7]addresses the semantic diversity of subgraphs by using graph contrastive learning to retrieve semantically similar KG subgraphs, constructing specific semantic tasks to enhance the performance of downstream knowledge graph question answering tasks. Despite continuous technological breakthroughs and integrations improving KGQG performance, these models perform poorly under few-shot conditions. Pre-trained models represented by GPT and BERT have demonstrated significant potential in few-shot scenarios. The TEGTOK model [20] integrates task-specific knowledge with knowledge bases, injecting dense retrieval into encoding and decoding stages to improve entity accuracy in generation tasks. However, pre-trained models still exhibit semantic relation understanding biases, inevitable hallucination issues, low BLEU-4 scores, and insufficient

phrase matching precision, necessitating further optimization of the relation understanding process [12, 18].

Large language model Although LLMs have strong generation capabilities, their training costs and model parameter quantities are extremely high [1]. Consequently, numerous studies have emerged to combine LLMs with small models to solve the mentioned problem. The LPKG framework [23] generates planning training data using KG subgraph patterns, fine-tuning small models to learn complex question decomposition. By integrating external knowledge supplemented via LLMs' retrieval enhancement with small models' reasoning path planning, this approach improves KGQG task efficiency and accuracy. Additionally, while knowledge distillation is a common paradigm for synergizing LLMs and small models, it is mostly applicable to specific scenarios and difficult to directly implement in KGQG [9, 21]. Related research has found that LLMs perform inadequately in few-shot-based KGQG tasks, producing samples with poor entity matching accuracy.

3 Methodology

3.1 Task Definition

Knowledge graph is a structured knowledge base graph composed of multiple triples $G = \{(s_i, r_i, o_i)\}_{i=1}^{n}$,wheres_i represents the subject, $r_i$represents the relation, and o_i represents the object. KGQG task aims to extract information from such structured knowledge graphs, enabling models to generate logically clear questions. These questions need to accurately guide models to match target answers. For a knowledge graph subgraph G composed of the aforementioned triples, based on a given set of correct target answers, a natural language question with high semantic consistency and logical correctness is to be generated. According to the task definition, the input content of few-shot examples is required to include the answer a_i ,the question q_i ,and the subgraph G_i.

3.2 Framework Overview

We outline the overall framework of the AIC-KGQG model (see Fig. 2). AIC-KGQG is a few-shot KGQG framework that integrates the advantages of LLMs and small models, effectively addressing real-world challenges such as high deployment costs and hallucination issues of LLMs, as well as the excessive dependence of traditional KGQG models on manually annotated data.

3.3 Abstraction-to-Instantiation Conversion Based Few-Shot

Traditional sequence models represented by RNN and Transformer struggle to directly encode the graph structure information of knowledge graphs, while LLMs,

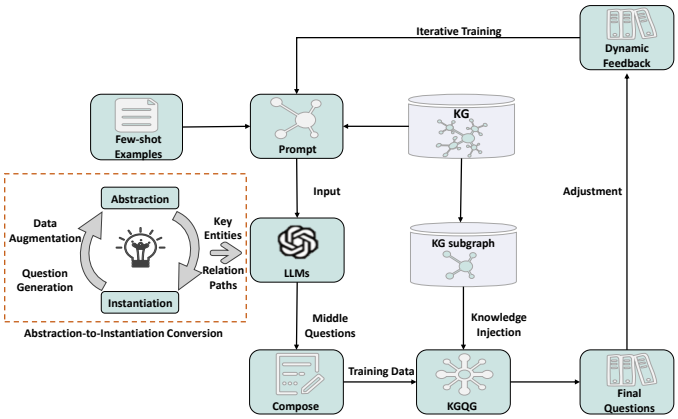

Fig. 2. The core framework of AIC-KGQG.

despite their powerful generation capabilities, tend to produce "hallucinated" entities or relations inconsistent with subgraphs. To address this issue, we introduce the abstraction-to-instantiation conversion mechanism by referencing the core mechanisms of chain-of-thought and structured prompt engineering, guiding LLMs to generate logically and semantically consistent candidate questions in stages. We use ChatGPT-3.5-turbo as LLM(temperature=0.4). The prompting process consists of three parts: abstract analysis stage, structure parsing stage, and instantiation generation stage. Zhang's research shows that providing LLMs with a small number of input examples as prompts can significantly improve the accuracy and normativity of generated questions. Few-shot examples demonstrate that the abstraction-to-instantiation conversion mechanism effectively guides LLMs to mimic examples through step-by-step reasoning, enhancing the quality and efficiency of question generation (see Fig. 3).

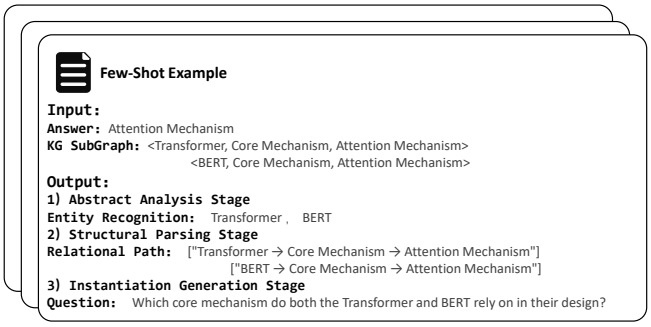

Fig. 3. A Few-shot Example.

3.4 Small Model Collaborative Training

Compared to LLMs, small models have far fewer parameters and lower training costs. We generate training data by combining the questions Q_G produced by LLMs through the abstraction-to-instantiation conversion based few-shot module with input answers a_i and subgraphs G_i.These automatically generated training data effectively address the scarcity of datasets while avoiding the expensive and time-consuming process of manual data annotation. We adopt G2S-AE, one of the state-of-the-art models in the question generation field, as the small model [2, 3]. This model performs Levi graph transformation on the knowledge graph, converting the structure of triples into a bipartite graph form. For example, the triple (Rome, Captain, Italy) is transformed into "Rome→Captain→Italy". In knowledge graph subgraphs with sparse entity-relation density, the Levi-transformed graph exhibits a predominantly linear network structure. During encoding, the model employs a BiGraphSeq architecture [17, 14]:

$$\phi_{\mathcal{U}_{in}(\mathcal{V})}^{(t)} = MEAN\left(\left\{\phi_{\mathcal{V}}^{(t-1)}\right\} \cup \left\{\phi_{\mathcal{U}}^{(t-1)} \mid \mathcal{U} \in \mathcal{U}_{in}(\mathcal{V})\right\}\right) \tag{1}$$

MEAN represents the average pooling process, which captures directional semantics by separately aggregating the information of outgoing and incoming edges.

3.5 Dynamic Feedback Optimization

We design a dynamic feedback module to form a closed-loop iteration for gradually improving model performance through real-time evaluation of generation quality. The evaluation primarily uses BLEU-4 and ROUGE-L scores [3]:

$$\text{BLEU} - 4 = \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n\right) \cdot \min\left(1, \frac{L_{Generation}}{L_{Reference}}\right) \tag{2}$$

$$ROUGE - L = \frac{\left(1+\beta^2\right) \cdot P_{LCS} \cdot R_{LCS}}{\beta^2 \cdot P_{LCS} + R_{LCS}}, \quad \beta = 1.2 \tag{3}$$

The similarity between the generated text and the reference text is measured by the matching length of the Longest Common Subsequence (LCS). For example, if the reference text sequence is Researchers developed a new algorithm to improve image classification accuracy and the generated text sequence is A novel algorithm was proposed to enhance the accuracy of classifying images, the longest common subsequence of these two texts is "algorithm to accuracy image", with a length of 4. L_{LCS} represents the length of the longest common subsequence.

$$R_{LCS} = \frac{L_{LCS}}{L_{Reference}}, P_{LCS} = \frac{L_{LCS}}{L_{Generation}} \tag{4}$$

The model is evaluated by the average of BLEU-4 and ROUGE-L to achieve iterative optimization. When the model shows no improvement for three consecutive evaluations, the training process is promptly truncated to enhance training and evaluation efficiency.

4 Experiments

In this section, we elaborate on the details of experimental settings, including the datasets used and baseline models for comparison. The framework is then evaluated from multiple dimensions, such as ablation experiments, iterative analysis, and case analysis.

4.1 Datasets

WebQuestions (WQ) is a question-answering dataset constructed from real Google user queries, primarily composed of WebQuestionsSP and ComplexWebQuestions. It is mainly used for training and evaluating single-hop knowledge base question-answering systems, providing questions, answers, and annotated subgraphs [10, 25].

PathQuestions (PQ) is another important dataset in the KGQG field, mainly consisting of multi-hop questions with relations based on complex path reasoning. It is primarily used to evaluate the ability of question-answering models to understand semantic relations and perform path reasoning [10, 25].

4.2 Baselines

We compared AIC-KGQG with the following benchmark models under different sample size conditions. It includes Standard prompt, Chain of Thought (CoT), L2A [4], KQG-CoT [13] and G2S-AE [2, 3].

4.3 Evaluation Metrics

Based on the research status of KGQG, we designed a multi-dimensional evaluation system, mainly including BLEU-4 and ROUGE-L metrics, which are currently used in the KGQG field to measure the accuracy and recall of text generation tasks.

4.4 Main Results

Table 1 shows the performance of AIC-KGQG and other LLM-based baseline models on the WQ and PQ datasets. Experiments indicate that AIC-KGQG outperforms all baseline models. The CoT method demonstrates significant improvements compared to the standard prompt, showing that guiding LLMs through step-by-step prompting can effectively enhance task completion and generate high-quality questions.

Table 2 presents the experimental results of AIC-KGQG and other small-model baseline models under full-sample and few-shot conditions. Under few-shot conditions, the collaborative method of large and small model significantly improves performance across all datasets, and AIC-KGQG outperforms all small models under few-shot conditions in all metrics on all datasets. Additionally, AIC-KGQG performs impressively under full-sample conditions, basically approaching the current state-of-the-art G2S-AE model.

Table 1. Results of Baseline Methods and AIC-KGQG Method based on LLMs

Model	WQ		PQ	
	BLEU-4	**ROUGE-L**	**BLEU-4**	**ROUGE-L**
Standard prompt	20.23	47.28	50.97	71.26
CoT	23.67	49.93	51.28	73.94
Few-shot CoT	27.18	53.12	53.95	75.58
KQG-CoT	**28.71**	53.85	57.22	76.22
AIC-KGQG(ours)	28.57	**54.77**	**59.51**	**76.41**

Table 2. Results of Baseline Methods and AIC-KGQG Method Based on Small-Parameter Models

Model (Sample Size)	WQ		PQ	
	BLEU-4	**ROUGE-L**	**BLEU-4**	**ROUGE-L**
L2A (Full Sample)	6.01	25.24	17.00	50.38
Transformer (Full Sample)	8.94	32.61	56.43	73.64
G2S-AE (Full Sample)	**29.40**	**55.23**	**59.59**	75.20
G2S-AE (Few Shot)	0.46	7.97	1.07	10.76
AIC-KGQG (Few Shot)	28.57	54.77	59.51	**76.41**

4.5 Ablation Study

We conducted ablation experiments to evaluate the impact of each module in AIC-KGQG on the overall performance of the model.

Table 3. Results of Ablation Studies Under Few-Shot Conditions

Model	WQ		PQ	
	BLEU-4	**ROUGE-L**	**BLEU-4**	**ROUGE-L**
w/o Abstract Understanding	14.93	39.63	40.11	63.21
w/o Structural Analysis	15.97	42.80	45.48	65.28
w/o LLM	0.46	7.97	1.07	10.76
w/o Dynamic Adjustment	17.46	44.47	56.88	76.01

The experiment shows that if the abstract understanding step is removed, the model will be unable to effectively extract all entity information, resulting in the occurrence of entity ambiguity in the generated questions, and subsequently leading to alignment failure. If the structural analysis step is removed, the model will be unable to capture the key relationship paths, causing too many prompts when generating questions based on the answers, and resulting in poor quality of the generated questions. It is worth noting that when the entire LLMs is removed, the overall performance of the model significantly declines under the condition of small samples, indicating that the data generated by the abstract-

to-instance transformation mechanism of LLMs plays a crucial role in the entire model.

4.6 Iteration Study

We analyzed the process of iterative training. The experimental results in Figure 4 show that collaborative training of large and small models has a significant impact on the matching accuracy of text generation, particularly for the matching accuracy of long-distance word order. In addition to enhancing generation accuracy, iterating on the collaborative training module for large and small models has also improved the coverage of effective information. Under the WQ dataset, the ROUGE-L metric achieved a substantial improvement, but the improvement was less significant under the PQ dataset, indicating that the optimization of AIC-KGQG primarily focuses on the coverage and matching of the longest common subsequence (see Fig. 4).

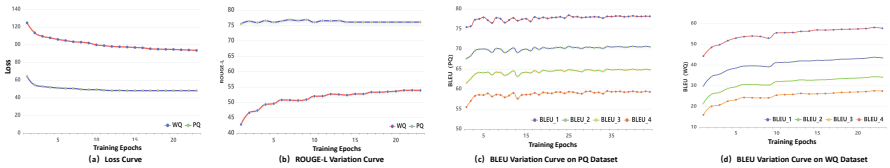

Fig. 4. Figure a shows the loss curve during the training process of the AIC-KGQG model. Figure b depicts the increasing curve of ROUGE-L scores as the training iterations progress. Figures c and d illustrate the changes in BLEU scores as the number of training epochs increases.

4.7 Manual Evaluation

To further assess generated question quality, 20 questions were randomly sampled from the WQ test set for expert evaluation, focusing on relevance and fluency (scored 15, with higher scores indicating better quality). To minimize evaluation costs, comparisons were made among representative LLM-based methods, small-model KGQG approaches, and AIC-KGQG. Results in Table 4 show that while LLM-based methods yield highly fluent questions, their relevance is low, likely due to hallucination. Our AIC-KGQG achieves the best performance in both relevance and fluency.

5 Conclusion

At present, the training of small models in state-of-the-art KGQG overly relies on high-quality question-answering datasets, while LLMs suffer from issues

Table 4. Results of manual evaluation on the WQ dataset

Model	Overall Average	Relevance	Fluency
COT (Few-shot)	3.25	2.59	3.91
G2S-AE (Full-sample)	3.38	3.42	3.33
AIC-KGQG (Few-shot)	**3.87**	**3.62**	**4.12**

such as high training costs and excessively large parameter sizes. Based on this, we propose a collaborative framework for large and small models, AIC-KGQG, which aims to combine the efficient deployment capabilities of small models with the generative and generalization abilities of LLMs. The framework integrates an abstraction-to-instantiation conversion mechanism and a collaborative training mechanism for large and small models. Through targeted prompting, it guides LLMs to engage in multi-stage thinking to improve the quality of question generation. Experimental results show that the proposed framework achieves state-of-the-art performance on BLEU-4 and ROUGE-L metrics across multiple public datasets. In the future, more efficient approaches warrant further exploration.

Acknowledgments. This work was partially supported by NSFC under grant No. 62302513.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. An, R., Yang, S., et al.: Unictokens: Boosting personalized understanding and generation via unified concept tokens. CoRR **abs/2505.14671** (2025). https://doi.org/10.48550/ARXIV.2505.14671, https://doi.org/10.48550/arXiv.2505.14671
2. Chen, Y., Wu, L., Zaki, M.J.: Reinforcement learning based graph-to-sequence model for natural question generation. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020)
3. Chen, Y., Wu, L., Zaki, M.J.: Toward subgraph-guided knowledge graph question generation with graph neural networks. IEEE Trans. Neural Networks Learn. Syst. **35**(9), 12706–12717 (2024). https://doi.org/10.1109/TNNLS.2023.3264519
4. Du, X., Shao, J., Cardie, C.: Learning to ask: Neural question generation for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. pp. 1342–1352. Association for Computational Linguistics (2017). https://doi.org/10.18653/V1/P17-1123
5. Fei, Z., Zhou, X., Gui, T., Zhang, Q., Huang, X.: LFKQG: A controlled generation framework with local fine-tuning for question generation over knowledge bases. In: Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022. pp. 6575–6585. International Committee on Computational Linguistics (2022)

6. Guo, S., Liao, L., Zhang, J., Wang, Y., Li, C., Chen, H.: SGSH: stimulate large language models with skeleton heuristics for knowledge base question generation. In: Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024. pp. 4613–4625. Association for Computational Linguistics (2024). https://doi.org/10.18653/V1/2024.FINDINGS-NAACL.287
7. Guo, S., Zhang, J., Wang, Y., Zhang, Q., Li, C., Chen, H.: DSM: question generation over knowledge base via modeling diverse subgraphs with meta-learner. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. pp. 4194–4207. Association for Computational Linguistics (2022). https://doi.org/10.18653/V1/2022.EMNLP-MAIN.281
8. Jia, R., Liang, P.: Data recombination for neural semantic parsing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics (2016). https://doi.org/10.18653/V1/P16-1002
9. Kelvinius, F.E., Georgiev, D., Toshev, A.P., Gasteiger, J.: Accelerating molecular graph neural networks via knowledge distillation. 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 (2023)
10. Kumar, V., Hua, Y., Ramakrishnan, G., Qi, G., Gao, L., Li, Y.: Difficulty-controllable multi-hop question generation from knowledge graphs. In: The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11778, pp. 382–398. Springer (2019). https://doi.org/10.1007/978-3-030-30793-6_22, https://doi.org/10.1007/978-3-030-30793-6_22
11. Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W.X., Wen, J.: A survey on complex knowledge base question answering: Methods, challenges and solutions. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021. pp. 4483–4491. ijcai.org (2021). https://doi.org/10.24963/IJCAI.2021/611
12. Li, J., Tang, T., Zhao, W.X., Wei, Z., Yuan, N.J., Wen, J.: Few-shot knowledge graph-to-text generation with pretrained language models. In: Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021. Findings of ACL, vol. ACL/IJCNLP 2021, pp. 1558–1568. Association for Computational Linguistics (2021). https://doi.org/10.18653/V1/2021.FINDINGS-ACL.136
13. Liang, Y., Wang, J., Zhu, H., Wang, L., Qian, W., Lan, Y.: Prompting large language models with chain-of-thought for few-shot knowledge base question generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023. pp. 4329–4343. Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.EMNLP-MAIN.263
14. Lin, C.: Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? In: Kando, N., Ishikawa, H. (eds.) Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, NTCIR-4, National Center of Sciences, Tokyo, Japan, June 2-4, 2004. National Institute of Informatics (NII) (2004)

15. OpenAI: GPT-4 technical report. CoRR **abs/2303.08774** (2023). https://doi.org/10.48550/ARXIV.2303.08774
16. Ouyang, L., Wu, J., et al.: Training language models to follow instructions with human feedback. 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 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. pp. 311–318. ACL (2002). https://doi.org/10.3115/1073083.1073135
18. Seyler, D., Yahya, M., Berberich, K.: Knowledge questions from knowledge graphs. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017. pp. 11–18. ACM (2017). https://doi.org/10.1145/3121050.3121073
19. Sun, Z., Lyu, C., Li, B., Wan, Y., Zhang, H., Li, G., Jin, Z.: Enhancing code generation performance of smaller models by distilling the reasoning ability of llms. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy. pp. 5878–5895. ELRA and ICCL (2024)
20. Tan, C., Gu, J., Tao, C., Ling, Z., Xu, C., Hu, H., Geng, X., Jiang, D.: Tegtok: Augmenting text generation via task-specific and open-world knowledge. In: Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022. pp. 1597–1609. Association for Computational Linguistics (2022). https://doi.org/10.18653/V1/2022.FINDINGS-ACL.125
21. Tian, Y., Pei, S., Zhang, X., Zhang, C., Chawla, N.V.: Knowledge distillation on graphs: A survey. ACM Comput. Surv. **57**(8), 189:1–189:16 (2025). https://doi.org/10.1145/3711121
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017)
23. Wang, J., Chen, M., Hu, B., Yang, D., Liu, Z., Shen, Y., Wei, P., Zhang, Z., Gu, J., Zhou, J., Pan, J.Z., Zhang, W., Chen, H.: Learning to plan for retrieval-augmented large language models from knowledge graphs. In: Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024. pp. 7813–7835. Association for Computational Linguistics (2024)
24. Zhao, R., Tang, J., Zeng, W., Chen, Z., Zhao, X.: Zero-shot knowledge graph question generation via multi-agent llms and small models synthesis. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024. pp. 3341–3351. ACM (2024). https://doi.org/10.1145/3627673.3679805
25. Zhou, M., Huang, M., Zhu, X.: An interpretable reasoning network for multi-relation question answering. In: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. pp. 2010–2022. Association for Computational Linguistics (2018), https://aclanthology.org/C18-1171/