A Comprehensive Evaluation Method for KG-Augmented Large Language Models

Xingyu Chen[1], Ligang Dong[1(✉)], and Meng Han[2]

[1] College of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China
donglg@zjsu.edu.cn

[2] College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China
mhan@zju.edu.cn

Abstract. Integrating factual information from knowledge graphs (KGs) into large language models (LLMs) has emerged as a promising approach to mitigate hallucination issues inherent in LLMs. This augmentation not only addresses the problem of generating inaccurate or fictional content but also offers avenues for tailoring LLMs to specific domains. Despite the potential benefits, the current body of research lacks a thorough examination of how KG augmentation influences various large language models.

This paper introduces a comprehensive evaluation method specifically designed for assessing the performance of KG-augmented LLMs. By evaluating these frameworks across multiple dimensions, the proposed method aims to provide a nuanced understanding of the strengths and limitations associated with integrating KGs into LLM-based question-answering systems. The systematic evaluation is expected to offer valuable insights, guiding future research endeavors and facilitating enhancements in this emerging field. This approach contributes to advancing the integration of KGs with LLMs and fostering the development of more robust and context-aware language models tailored to specific knowledge domains.

Keywords: Large Language Model · Knowledge Graph · Augmented

1 Introduction

Recently, Large Language Models (LLMs) like GPT-4 [1], Baichuan [2], and ChatGLM [3] have achieved remarkable success in Natural Language Processing (NLP) tasks. LLMs have demonstrated human-level performance across a wide spectrum of NLP tasks, including question answering, translation, and information extraction, among others. Despite their outstanding general capabilities,

X. Li et al. (Eds.): SmartGift 2024, LNICST 600, pp. 112–121, 2025.
https://doi.org/10.1007/978-3-031-78806-2_7

LLMs still suffer significant challenges, including hallucination, a lack of interpretability, and the absence of domain-specific knowledge.

Firstly, hallucinations in LLMs are attributed to the fact that the knowledge embedded in these models relies on pre-trained data, which may contain misinformation and biases. This can result in issues such as missing domain-specific knowledge and reliance on outdated information. Furthermore, the random content generation based on probability in LLMs makes them prone to hallucinations. Secondly, as black box models, the knowledge within LLMs is internalized in the parameters, which cannot explain and verify the knowledge obtained by LLM to produce answers. Finally, the currently available LLMs are designed for general purposes and lack specialization for specific fields. This limitation is particularly critical in applications like medical diagnosis, where an erroneous result poses a significant risk to the health and safety of patients.

To address these issues, researchers have proposed to integrate knowledge graphs (KGs) into LLMs [4]. KGs are structured representation of knowledge, where the fundamental unit is a triplet, i.e., (entity, relationship, entity). KGs play a crucial role in providing explicit knowledge to LLMs, ensuring the generation of interpretable results. Moreover, KGs support the integration of new knowledge, effectively mitigating the problem of knowledge obsolescence. Additionally, there are numerous domain-specific KGs, including those in finance, law, education, etc., which furnish LLMs with precise and reliable domain knowledge. This approach effectively addresses the issue of missing domain knowledge in LLM and eliminates the need for expensive fine-tuning. The framework that combines KGs with LLMs is collectively referred to as KG-LLM in this paper.

As the demand for large models continues to grow in the specific field, researchers have introduced an increasing number of KG-augmented methods. However, a notable gap exists as there is currently no established framework for evaluating the effectiveness of these approaches. To bridge this gap, this paper proposes a comprehensive evaluation method for KG-augmented methods. This method utilizes two KGs from different domains to evaluate the capabilities of the state-of-the-art KG-LLM framework from six perspectives. It reveals limitations in their performance across diverse domains.

2 Related Work

While LLM has demonstrated strong performance in numerous tasks, its performance in knowledge-intensive tasks has been less than satisfactory. For example, in Knowledge Graph Question Answering (KGQA) tasks, LLMs may suffer from hallucinations, inaccurate answers, and outdated knowledge [5]. In recent years, a considerable amount of research has been conducted to enhance LLM's question answering capabilities using KGs. Specifically, the approach involves extracting relevant triplets from the KGs that are related to the given question. These extracted knowledge pieces are then injected into the reasoning process of LLM, serving as prompts to generate more reliable answers.

To tackle the challenge of LLMs' performance in knowledge-intensive tasks, Wu, et al. [6] proposed a KG-To-Text method, which aims to represent the knowledge stored in a knowledge graph (KG) as natural language. By transforming KG knowledge into a textual format, LLMs can leverage this information more effectively to generate accurate and contextually appropriate answers. Beak, et al. [8] proposed a Zero-Shot Knowledge-Augmented language model Prompt-ING (KAPING) framework in which facts retrieved from KGs are added to questions passed to LLM to generate more reliable answers. Wang, et al. [10] proposed a knowledge-based PLM framework KP-PLM that can be combined with any mainstream LLM. Firstly construct a knowledge subgraph from KBs for each context, and then design multiple sequential prompt rules to turn the knowledge subgraph into a natural language prompt to help LM generate more accurate answers. To improve the zero-shot reasoning ability of LLMs on structured data in a uniform way, Jiang, et al. [9] proposed StructGPT, an Iterative Reading-then Reasoning method targeting at structured Data (such as KGs, data table, and database) to support LLMs to reason structured data with the help of external interfaces. Guan, et al. [7] introduced a novel framework called Knowledge Graph-based Retrofitting (KGR) to address the issue of factual hallucination during the reasoning process of LLMs. This framework leverages the factual knowledge stored in knowledge graphs (KGs) to mitigate the problem.

An increasing number of KG-augmented LLM methods have been proposed by researchers, but there lacks a unified framework to evaluate them. In this paper, the framework proposed by us aims to evaluate the performance of KG-enhanced LLM methods from multiple dimensions.

3 Evaluation Criteria

This paper proposes a framework for comparing KG-enhanced LLM Question Answering (QA) systems and defines fundamental criteria across multiple dimensions: answer accuracy, randomness, universality, robustness, reasoning ability on the complex questions, and knowledge updating capability. By evaluating performance along these dimensions, a more comprehensive understanding of the strengths and limitations of KG-enhanced LLM question-answering systems can be obtained, providing guidance for further research and improvement. Here are detailed explanations of the metrics mentioned:

Accuracy: In the KGQA task, the accuracy of the provided answers is an important metric in the framework proposed in this paper. When performing the KGQA task, accuracy is measured by comparing the answers to those contained in the ground truth. However, LLMs, as generative models, produce answers in an uncontrolled format that differs from the generated answers in KGQA, LLMs primarily generate text similar to human language and require semantic understanding to assess the accuracy of the generated answers. This paper compares the answers generated by the LLMs with the vertices from the KGs. To overcome these challenges, the paper performed manual evaluation to ensure fairness. The

calculation formula for the accuracy metric is as follows:

$$\text{Accuracy} = \frac{\text{Number of Correctly Answered Questions}}{\text{Total Number of Answered Questions}}$$

Randomness: As generative models, LLMs often exhibit a certain degree of randomness in the answers they generate when faced with the same context and given question. Therefore, assessing the randomness of the KG-LLM framework is also a crucial metric. To account for this factor, this paper employs a method of running each question three times and calculates accuracy based on the best answer obtained. This approach allows for a more comprehensive evaluation of the performance of the KG-LLM framework.

Universality: The universality of the KG-LLM framework refers to its ability to be applied to various knowledge graphs in different domains without retraining the model, while maintaining its ability to answer questions effectively. To evaluate the universality of the framework, this paper conducts experiments on four real-world knowledge graphs from two domains and benchmarks. The datasets includes questions of varying complexity and styles, carefully curated to ensure a wide coverage in evaluating the framework.

Robustness: In LLMs, robustness encompasses their ability to tolerate erroneous inputs, such as questions containing spelling or grammatical errors. These errors are quite common due to the free-form nature of human input. Despite the presence of errors, a well-trained LLM is still capable of generating correct answers. Furthermore, when presented with a question that is related to entities in the KGs but for which the answer does not exist in KGs, LLMs should either refuse to answer the question or provide a response indicating insufficient information.

Complex question reasoning: When tackling KGQA tasks, we often encounter both simple and complex questions. Simple questions refer to those that can be answered using a single relation path without additional reasoning, while complex questions require reasoning through multiple relation paths to arrive at answers. When users pose questions of varying complexity, such as temporal questions or multi-hop questions, evaluating the model's ability for complex question reasoning becomes crucial. LLMs typically encapsulate question understanding within a broader output generation process. For complex questions, we need to observe whether the model can generate accurate answers by combining reasoning and relation paths. By analyzing the model's generated answers, we can assess its ability for complex question reasoning and evaluate its performance.

Knowledge update: KGs are frequently updated by adding, deleting, or modifying facts. For instance, Wikipedia undergoes edits at an astonishing rate every second, which means hundreds of thousands of edits are made in the corresponding Wikidata KG every day. Users typically expect to receive the most up-to-date answers. In our evaluation, we take this criterion into consideration to ensure that our framework can adapt to evolving knowledge and provide the latest answers.

4 Experiment

In this section, we evaluated the performance of various KG-LLM frameworks and analyzed the results to summarize the main challenges currently faced. We conducted extensive testing on these frameworks and considered various metrics. Through an objective analysis of these issues, this paper provides valuable insights for further improvement and development of KG-LLM frameworks.

4.1 Datasets

To evaluate the performance of the KG-LLM framework, this paper has chosen two commonly used Knowledge Graph Question Answering (KGQA) datasets: WebQuestionsSP (WebQSP) [11] and MetaQA [12].

4.2 Main Results

We employed the proposed evaluation method in this paper to assess the performance of four KG-LLM frameworks. Table 1 summarizes the accuracy scores of each participant under various benchmarks. To ensure fairness in the evaluation, each KG-LLM framework utilized LLaMA2-Chat-7B as the LLM backbone. In the experiments presented in this paper, answers were categorized into three groups:

(1) Correct, where the LLM provided an answer semantically consistent with the correct answer;
(2) Incorrect, where the answer provided by the LLM did not align with the semantic meaning of the correct answer;
(3) No Answer, indicating that the LLM concluded the question was unanswerable. Only answers categorized as correct are considered as accurate in this study.

Table 1. Baselines results comparison on WebQSP and CWQ datasets.

Model	WebQSP		CWQ	
	Accuracy	F1	Accuracy	F1
KAPING	59.5	62.1	55.2	54.3
RoG	67.4	70.5	52.4	55.9
RRA	72.4	76.0	68.3	70.4
ChatKBQA	73.6	78.2	72.6	77.3

1. KAPING [8]: A zero-shot framework that forwards knowledge from KG to LLM to generate answers without additional training
2. RoG [13]:RoG initially generates relation paths grounded in KGs as faithful paths. Subsequently, these paths are employed to retrieve valid reasoning paths from the KGs, enabling LLMs to perform faithful reasoning.

3. Retrieve-Rewrite-Answer (RRA) [6]: A KG-to-Text approach is introduced, which is sensitive to answers, aiming to convert knowledge from the KGs into textual descriptions, thereby enhancing the accuracy of LLM answers.
4. ChatKBQA [14]: A generation-retrieval framework based on fine-tuned open source LLMs.

By analyzing the results in Table 1, it is evident that the performance of the four evaluated KG-LLM frameworks on WebQSP surpasses their performance on CWQ. RoG, by generating faithful paths grounded in KGs as reasoning paths, achieves better results than KAPING. RRA employs a KG-to-Text method, converting triples into free-format text, enabling LLMs to better comprehend the provided factual knowledge and thus enhancing their capabilities in KGQA. To ensure experimental fairness, the ChatKBQA framework continues to use an LLM without fine-tuning. The generation-retrieval approach utilized by ChatKBQA proves superior to the three aforementioned retrieval-generation methods, as the latter may introduce erroneous interference information in the retrieved data, impacting downstream LLM tasks.

In terms of universality, the experimental results indicate that all four KG-LLM frameworks are applicable to different datasets without the need for retraining the models, while maintaining accuracy in answering questions. It's worth noting that both the WebQSP and CWQ datasets encompass complex questions. In the CWQ dataset, specifically in the testset, over half of the questions involve multiple entities, and some questions require intricate reasoning. Importantly, the aforementioned KG-LLM frameworks demonstrate excellent performance in addressing complex questions, maintaining high accuracy even when confronted with intricate reasoning scenarios.

4.3 Randomness Analysis of KG-LLM Framework

In order to assess the randomness of the KG-LLM framework, this section randomly sampled 100 questions from the WebQSP and MetaQA datasets, running each question three times. The results of the runs can be categorized into three scenarios:

1. Consistent Semantics (CS) across all three responses;
2. Consistent Semantics between Two Responses (CSTR), with the third being inconsistent;
3. Inconsistent Semantics (IS) across all three responses.

Analyzing the experimental results in Figs. 1 and 2, it is observed that the four compared KG-LLM frameworks exhibit a high level of determinism with a certain degree of randomness in LLMs. During the experiments, explanations for some answers varied slightly, but the answers remained consistent. In certain cases, questions initially went unanswered but produced correct responses in subsequent iterations. Furthermore, for some questions, the results changed entirely across three attempts. The reason for this phenomenon is that LLMs, as

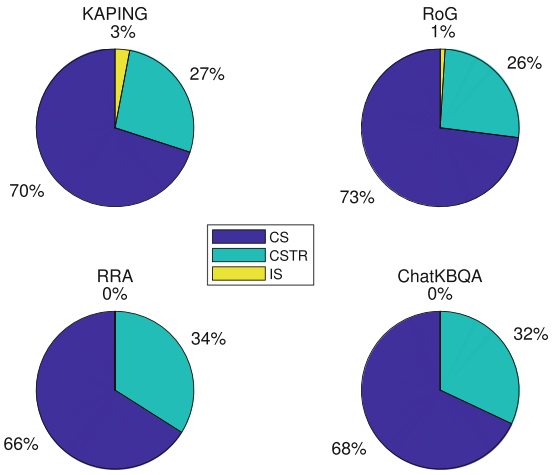

Fig. 1. The performance of KG-LLM frameworks running three times on WebQSP.

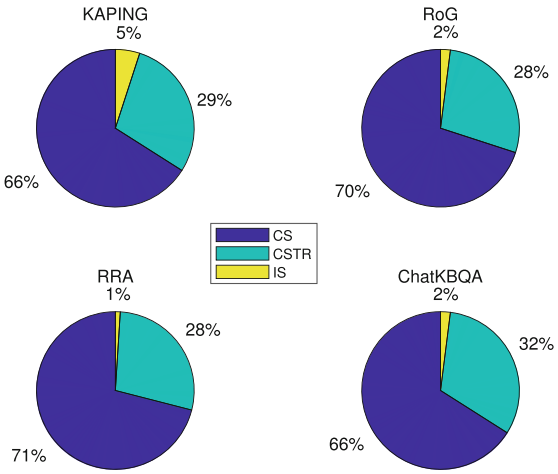

Fig. 2. The performance of KG-LLM frameworks running three times on CWQ.

generative models, rely on probabilistic calculations for each response, introducing inherent randomness into the answering process. Additionally, the process of retrieving knowledge from KGs also involves an element of randomness, making it challenging to ensure consistent results with each retrieval.

4.4 Robustness Analysis of KG-LLM Framework

In order to assess the robustness of the KG-LLM framework, this section initially randomly selected 10 questions from the WebQSP and MetaQA datasets. Four different versions of each question were created by introducing various spelling

errors and grammatical errors. There were two types of spelling errors, namely entity misspelling and non-entity misspelling, along with two types of grammar errors. Subsequently, based on the knowledge graph corresponding to the WebQSP and MetaQA datasets, 40 questions were generated. These questions involved entities from the knowledge graph, but the answers were not present in the knowledge graph. For such questions, the model should Reject to Answer (RA).

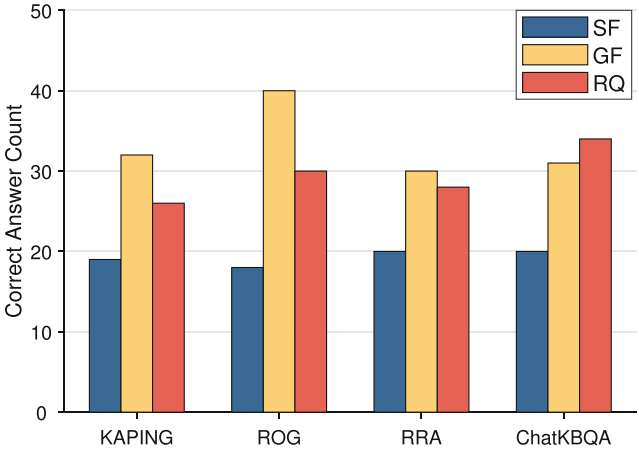

Fig. 3. Robustness analysis of KG-LLM Framework.

SE represents spelling errors, GE represents grammatical errors, and RA represents questions that should be refused to answer. Analysis of the results in Fig. 3 reveals that most spelling errors that the model cannot answer correctly are related to misspelled entities. The inability to correctly locate the right entities and relation paths in KGs due to misspelled entities leads to the generation of incorrect answers. Grammar errors have the least impact on the KG-LLM framework, as the LLM, leveraging its advanced semantic understanding capabilities, can still comprehend questions accurately. For questions that involve entities in KGs but are not present in them, noise is introduced during KG retrieval, increasing the difficulty of LLM reasoning. In general, current KG-LLM frameworks face challenges in effectively handling questions with entity spelling errors.

4.5 Knowledge Update Ability Analysis of KG-LLM Framework

In order to assess the KG-LLM framework's ability to handle new knowledge, this section employed the following approach: in the WebQSP, half of the information in the knowledge graph was removed, leaving only the remaining half. This was done to compare the performance of the KG-LLM framework in answering questions under these conditions.

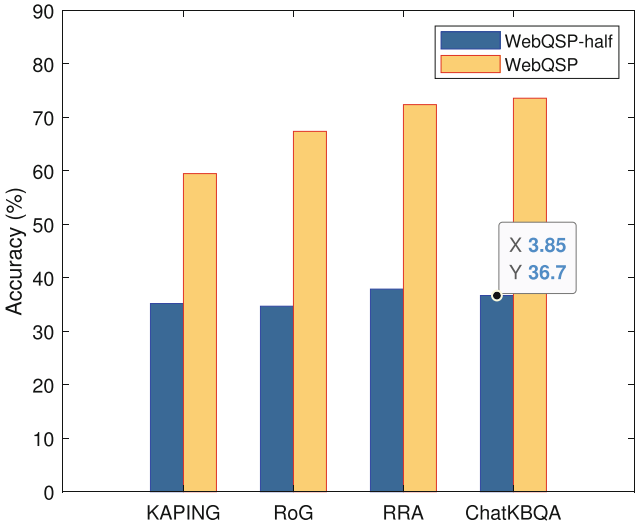

Fig. 4. The knowledge update ability on WebQSP.

Through the analysis of the results in Fig. 4, it is evident that when only half of the information in the knowledge graph remains, the KG-LLM framework often struggles to deduce correct answers through reasoning. This underscores the crucial importance of the adequacy of information within the knowledge graph for the successful inference of the KG-LLM framework. Effective reasoning and obtaining accurate results only occur when the information within the knowledge graph is extended to cover the content of the questions. It is worth noting that the emphasis here goes beyond the quantity of knowledge, encompassing the coverage and quality of knowledge. The evaluated knowledge update capabilities of the four KG-LLM frameworks in this study have been validated.

5 Conclusion

This paper introduces a Comprehensive Evaluation Method for KG-Augmented Methods, aiming to assess the effectiveness of KG-LLM framework from various perspectives. The proposed method defines six metrics for quantitative evaluation, providing a method for comparing KG-LLM frameworks. Extensive experiments conducted on the WebQSP and CWQ datasets reveal that current KG-LLM frameworks still exhibit certain limitations and challenges.

References

1. OPENAI, O.: GPT-4 Technical Report (2023)
2. Yang, A., Xiao, B., et al.: Baichuan 2: open large-scale language models (2023)
3. Zeng, A., Liu, X., et al.: Glm-130b: an open bilingual pre-trained model (2022)

4. Lewis, P., Perez, E., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)
5. Pan, S., Luo, L., et al.: Glm-130b: unifying large language models and knowledge graphs: a Roadmap(2023)
6. Wu, Y., Hu, N., et al.: Retrieve-rewrite-answer: a KG-to-text enhanced LLMs framework for knowledge graph question answering (2023)
7. Guan, X., Liu, Y., et al.: Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting (2023)
8. Baek, J., Aji, A., et al.: Knowledge-augmented language model prompting for zero-shot knowledge graph question answering (2023)
9. Jiang, J., Zhou, K., et al.: Structgpt: a general framework for large language model to reason over structured data (2023)
10. Wang, J., Huang, W., et al.: Knowledge prompting in pre-trained language model for natural language understanding (2022)
11. Yih, W., Richardson, M.: The value of semantic parse labeling for knowledge base question answering. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 201–206 (2016)
12. Zhang, Y., Dai, H.: Variational reasoning for question answering with knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
13. Luo, L., Li, Y., et al.: Reasoning on graphs: faithful and interpretable large language model reasoning (2023)
14. Luo, H., Tang, Z., et al.: Chatkbqa: a generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models (2023)