A Reasoning-Based Evaluation Framework for Knowledge Graph-Augmented LLMs

Ativ Joshi
atjoshi@umass.edu

Isheeta Sinha
issinha@umass.edu

Sravanthi Machcha
smachcha@umass.edu

Ayush Gupta
ayushanilgup@umass.edu

1 Introduction

Knowledge graphs are increasingly becoming popular to reduce hallucinations in LLMs as they are proven to be effective. However, the underlying reasoning that the LLM uses to arrive at the correct conclusions is still questionable. Techniques such as Chain-of-Thought (CoT) prompting (Wei et al., 2023) have been employed to reveal the reasoning capabilities. Our work aims to build an evaluation framework to determine how correctly the LLMs utilize the knowledge graph using the Chain-of-Thought reasoning which reflects how the model arrives at the final answer, and introduce a metric that reflects the faithfulness of the reasoning process to the input information.

As the usage of LLMs in sensitive domains like medicine and biomedical research increases, the reliability and faithfulness of reasoning in LLMs becomes crucial for ensuring accuracy, trustworthiness, and ethical integrity (Agarwal et al., 2024). It enhances user confidence and mitigates risks of misinformation or biased outputs. Reliable reasoning is essential for maintaining coherence and accuracy in complex problem-solving scenarios.

Limitations/Assumptions: The knowledge graph can be incomplete. To address this, we will use only questions from the dataset in which all entities involved are present in the knowledge graph. Our aim to is to study the effectiveness in the ideal complete knowledge case - in real world settings, getting a complete knowledge graph is a separate challenge by itself, our work does not attempt to solve this problem.

2 Related work

2.1 Background on Knowledge Graph

A knowledge graph (KG) is a structured representation of facts, consisting of entities, relationships, and semantic descriptions (Ji et al., 2021). Entities represent objects in the real world or abstract concepts, and relationships define the connections between these entities. These relationships, along with entities, are often captured in the form of triples, following the `<subject, predicate, object>` format. In a more general sense, these are usually shown as `<entity1, relation, entity2>` triples. For example, a triple like `<Albert Einstein, WinnerOf, Nobel Prize>` represents a fact about Albert Einstein. Knowledge graphs can be visualized as directed graphs, where nodes represent entities and edges denote relationships. They are widely used for knowledge representation and reasoning, enabling applications in search engines, recommendation systems, and question-answering tasks.

2.2 KG Augmented LLM

There are several works that show how knowledge graph-augmented LLMs have improved performance (Agrawal et al., 2023). There are various stages of the pipeline where the knowledge graph can be integrated to enhance LLMs. One of the ways is to integrate it into a prompt. Researchers integrate KGs for structured symbolic knowledge, primarily incorporating them at the input level to enhance contextual understanding (Agrawal et al., 2023). (Soman et al., 2024a) is an example of this work, where integrating RAG at prompt level has reduced hallucination in the domain of BioNLP with the given prompt. Despite the efforts, the issue of hallucination may continue to persist in the realm of KG integrated LLMs for the

foreseeable future (Pan et al., 2024). The actual reasoning behind the correctness of the answer is not evident, nor is it evident how correctly the KG was used.

There is also work done to evaluate the effectiveness of the knowledge graph (Jiang et al., 2024), which uses the output of the LLM to generate a graph, and compare it with the original graph. However, (Jiang et al., 2024) they don't explain/extract any reasoning metrics to evaluate the faithfulness. Instead, they feed the generated graph back to the model for fine-tuning model parameters.

2.3 LLM Reasoning

Many studies have proposed prompt based solutions to harness the reasoning ability of LLMs to handle complex tasks. However, the problem of hallucinations and lack of knowledge affect the faithfulness of LLMs' reasoning (Luo et al., 2023). Providing LLMs with relevant information from KG should mitigate hallucinations caused because of lack of information. However, it still does not address hallucinations due to limitations in reasoning capabilities. Chain-of-Thought (CoT) prompting (Wei et al., 2023) has become a popular way to interact with LLMs and understand its workings. It is important to note that validating the actual computations performed by the model against what the CoT presents is out of scope of this paper. Therefore, we focus on faithfulness of CoT reasoning to the input knowledge provided.

Faithfulness is a crucial attribute of large language models (LLMs), particularly in high-stakes domains such as healthcare and law, where the reliability of model outputs is critical (Agarwal et al., 2024). However, there is no universally agreed-upon definition or accepted metric for faithfulness. For the purposes of our work, we adopt the definition of faithfulness as an explanation that accurately reflects the reasoning process behind the model's prediction (Lyu et al., 2024)

In contrast to faithfulness, plausibility refers to how convincing the interpretation is to humans (Jacovi and Goldberg, 2020). Current LLMs inherently overemphasize plausibility over faithfulness (Agarwal et al., 2024). While LLMs can produce coherent and logical explanations which enhance their performance on complex tasks and decision-making processes through CoT, recent study (Si et al., 2024) suggests that users may over-rely on these CoT explanations, even when they are incorrect. This over-reliance is a critical risk associated with CoT reasoning, and it becomes particularly problematic when the LLM's reasoning seems highly plausible but is based on an incorrect line of reasoning.

2.4 Faithfulness metrics

There is no unique definition of faithfulness. There are many metrics used to capture different aspects of faithfulness. There is a need for developing reliable metrics to characterize the faithfulness of explanations (Agarwal et al., 2024). There are automated frameworks like ROUGE, BLEU, BERTScore, however they assess aspects such as fluency, semantic and lexicographic similarity (Malin et al., 2024). Firstly, we do not want to make comparisons based on the language, as there is a chance of ambiguity. Secondly, none of these methods target the faithfulness from the perspective of reasoning-chain and effectiveness knowledge graph utilization. Human evaluation is also a common method in evaluating reasoning effectiveness, however, not only is it resource-intensive and time-consuming, but it is also highly prone to subjective variation (Malin et al., 2024). Hence, it is not a desirable approach (Jacovi and Goldberg, 2020). Our work focuses more on developing a metric that captures the reasoning capabilities, regardless of how plausible the answer may seem. Other works, such as (Lyu et al., 2023), focus on evaluating faithfulness based on final answer accuracy, with an emphasis on improving the faithfulness of CoT reasoning generated by LLMs. Our work, however, centers on developing a metric that specifically quantifies the faithfulness of the reasoning process to the input information.

3 Approach

We have identified a need to understand the effectiveness of introducing knowledge graphs in order to mitigate hallucinations, specifically in terms of reasoning. We propose an evaluation framework to address this. The investigations will be restricted to one domain - we choose biomedical question answering - because generic knowledge graphs are known to be sparse and incom-

plete as they attempt to encapsulate broad knowledge across all domains (Demir et al., 2023; Waagmeester et al., 2020) and thus are not well-suited for evaluating reasoning capabilities. We aim to overcome this problem by narrowing the scope of our experiments to a specific domain.

3.1 Preprocess the QA dataset

The two main datasets we will be using is SPOKE (Morris et al., 2023) which is the knowledge graph and then the BiomixQA (Soman et al., 2024b) question answering (QA) dataset. In order to ensure that we don't evaluate the reasoning on a question that doesn't have the relevant information in the knowledge graph, we need to preprocess the QA dataset. This is done by generating entities from all the questions and using only questions where all the entities are present in the KG. The other questions are discarded. We plan to use a method similar to the one used in KG-RAG, i.e, utilize a system prompt to extract entities from the input questions and perform entity matching to align the extracted entities with their corresponding concept names in the KG. Entity matching is achieved by precomputing embeddings for all nodes in SPOKE using a sentence transformer model.

3.2 Integrate domain knowledge graph into LLM

As discussed in the previous section, there are multiple ways to augment knowledge graphs into LLMs at different stages. For our biomedical domain, we will be using the already generated knowledge graph SPOKE, which is accessible through APIs. We will integrate this knowledge in the LLM at the inference stage. Since the full knowledge graph may not be useful for a specific question, a relevant subgraph will be extracted. We use the terms input KG and sub-KG interchangeably to refer to the extracted relevant subgraph. This process involves identifying the essential entities in the question and then selecting nodes within an empirically determined diameter from that central node. The subgraph is then converted to natural language which is then combined with the question prompt, to produce an enriched prompt. This enriched prompt is combined with Chain-of-Thought (CoT) prompt template and fed into the model to produce an answer along with the reasoning behind it.

3.3 Construct a knowledge graph from reasoning

The CoT reasoning in the output will contain relations between entities as understood by the LLM. In order to verify the factuality of these claims in a structured manner, this reasoning text is converted into a knowledge graph, where the nodes represent the entities in the text and the edges represent the relations between the entities. The conversion process is carried out by the use of a refined prompt which strictly entails all the steps the LLM should take in creating the knowledge graph along with relevant examples for reference.

3.4 Calculate the evaluation metric

The output knowledge graph can be understood intuitively as a list of triples (also known as links) in the following format: `<entity1, relation, entity2>`. The corresponding triples should exist in the input KG for the LLM reasoning to be coherent. If we observe incorrect links or fabricated links in the CoT KG, it shows that the LLM has hallucinated. The first step is searching the relevant link in the input KG for each link in the output KG. This is done by converting the entities and relations into embeddings using the same model which was used in 3.2 and finding the edge which has the same two entities. Once we have both the links, we compute the cosine similarity for each pair of input-output links in the output KG, which can be interpreted as the coherence score for that link. The final score is calculated by averaging over the coherence scores for all links in the output KG. If any link is found to have a coherence score below a certain threshold, the final score is reported as zero for that query. This follows the assumption that a wrong logic in an intermediate step falsifies all subsequent logic. We can further provide a score for the entire question-answer dataset by averaging over all the queries in the dataset.

4 Schedule

The following subtasks are identified for project completion, with the expected time required to complete. We plan on working on all the subtasks together. The whole project spans till the end of the Spring 2025 semester. It is in total taking 7 weeks, and buffer of 2 weeks is added for unforeseen challenges and pivots.

1. Implement algorithm to pre-process QA dataset - 1 week
2. Integrate SPOKE into LLMs - 2 weeks
   (a) Investigate use of SPOKE API - 0.5 week
   (b) Implement algorithm to extract subgraph - 1 week
   (c) Integrate and develop question with CoT- 0.5 week
3. Extract knowledge graph from CoT output - 1 week
4. Experiment with different types of KG generation - 0.5 week
5. Implement algorithm for faithfulness metric - 1 week
6. Test and validate the metrics - 0.5 week
7. Work on final reports - 1 week

5 Data

The input knowledge graph that we are using for augmenting data into the LLM is SPOKE which combines over 40 biomedical knowledge sources, containing over 40 million nodes for diseases, genes, symptoms and many more. The question-answering dataset is BioMixQA (Soman et al., 2024b) which contains multiple choice questions. Each question contains 5 options and one correct answer.

6 Tools

We will primarily be using standard deep learning libraries like PyTorch and existing pre-trained models from Huggingface. Google Colab Pro would be sufficient for our GPU usage. If we need more compute, we plan to use Unity.

7 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - We primarily used Deep Research mode on Perplexity (which seems to be using Deepseek R1)

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.

  We mainly used LLMs to brainstorm the ideas and find references for background readings.

  - Here's an idea to verify that the output of an LLM is factually correct. First, we take an LLM and augment it with a knowledge graph. We pass the LLM a particular query and the LLM generates an output along with a chain of thought which represents the LLM's internal reasoning. We take the chain of thought output and convert it into a list of facts. Then we cross check these basic facts individually with the knowledge graph. Is this a feasible method? Has this been done before? Can you suggest any better ideas.
  - Can you suggest some project ideas along this direction?
  - Suggest some research project with core technical contributions, not just the ones where existing libraries are used to build a tool.
  - How important is it that the reasoning of LLM is reliable?
  - How sparse is the wikidata knowledge graph?

- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?

  - The deep research functionality was very useful for finding out relevant papers and cross verifying our rough ideas.

References

Agarwal, C., Tanneru, S. H., and Lakkaraju, H. (2024). Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models.

Agrawal, G., Kumarage, T., Alghamdi, Z., and Liu, H. (2023). Can knowledge graphs reduce hallucinations in llms?: A survey. *arXiv preprint arXiv:2311.07914*.

Demir, C., Wiebesiek, M., Lu, R., Ngomo, A.-C. N., and Heindorf, S. (2023). Litcqd: Multi-hop reasoning in incomplete knowledge graphs with numeric literals.

Jacovi, A. and Goldberg, Y. (2020). Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Ji, S., Pan, S., Cambria, E., Marttinen, P., and Philip, S. Y. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514.

Jiang, Z., Zhong, L., Sun, M., Xu, J., Sun, R., Cai, H., Luo, S., and Zhang, Z. (2024). Efficient knowledge infusion via kg-llm alignment. *arXiv preprint arXiv:2406.03746*.

Luo, L., Li, Y.-F., Haffari, G., and Pan, S. (2023). Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*.

Lyu, Q., Apidianaki, M., and Callison-Burch, C. (2024). Towards faithful model explanation in nlp: A survey.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. (2023). Faithful chain-of-thought reasoning.

Malin, B., Kalganova, T., and Boulgouris, N. (2024). A review of faithfulness metrics for hallucination assessment in large language models. *arXiv preprint arXiv:2501.00269*.

Morris, J. H., Soman, K., Akbas, R. E., Zhou, X., Smith, B., Meng, E. C., Huang, C. C., Cerono, G., Schenk, G., Rizk-Jackson, A., Harroud, A., Sanders, L., Costes, S. V., Bharat, K., Chakraborty, A., Pico, A. R., Mardirossian, T., Keiser, M., Tang, A., Hardi, J., Shi, Y., Musen, M., Israni, S., Huang, S., Rose, P. W., Nelson, C. A., and Baranzini, S. E. (2023). The scalable precision medicine open knowledge engine (spoke): a massive knowledge graph of biomedical information. *Bioinformatics*, 39(2):btad080.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Si, C., Goyal, N., Wu, S. T., Zhao, C., Feng, S., III, H. D., and Boyd-Graber, J. (2024). Large language models help humans verify truthfulness – except when they are convincingly wrong.

Soman, K., Rose, P. W., Morris, J. H., Akbas, R. E., Smith, B., Peetoom, B., Villouta-Reyes, C., Cerono, G., Shi, Y., Rizk-Jackson, A., et al. (2024a). Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9):btae560.

Soman, K., Rose, P. W., Morris, J. H., Akbas, R. E., Smith, B., Peetoom, B., Villouta-Reyes, C., Cerono, G., Shi, Y., Rizk-Jackson, A., Israni, S., Nelson, C. A., Huang, S., and Baranzini, S. E. (2024b). Biomedical knowledge graph-optimized prompt generation for large language models.

Waagmeester, A., Stupp, G., Burgstaller-Muehlbacher, S., Good, B. M., Griffith, M., Griffith, O. L., Hanspers, K., Hermjakob, H., Hudson, T. S., Hybiske, K., et al. (2020). Wikidata as a knowledge graph for the life sciences. *Elife*, 9:e52614.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.