Direct Evaluation of Chain-of-Thought in Multi-hop Reasoning with Knowledge Graphs

Minh-Vuong Nguyen[*♡] **Linhao Luo**[*♢] **Fatemeh Shiri**[♢] **Dinh Phung**[♢,♡]
Yuan-Fang Li[♢] **Thuy-Trang Vu**[♢] **Gholamreza Haffari**[♢]
[♢]Department of Data Science & AI, Monash University
[♡]VinAI Research, Vietnam
{trang.vu1,first.last}@monash.edu

Abstract

Large language models (LLMs) have demonstrated strong reasoning abilities when prompted to generate chain-of-thought (CoT) explanations alongside answers. However, previous research on evaluating LLMs has solely focused on answer accuracy, neglecting the correctness of the generated CoT. In this paper, we delve deeper into the CoT reasoning capabilities of LLMs in multi-hop question answering by utilizing knowledge graphs (KGs). We propose a novel discriminative and generative CoT evaluation paradigm to assess LLMs' knowledge of reasoning and the accuracy of the generated CoT. Through experiments conducted on 5 different families of LLMs across 2 multi-hop question-answering datasets, we find that LLMs possess sufficient knowledge to perform reasoning. However, there exists a significant disparity between answer accuracy and faithfulness of the CoT generated by LLMs, indicating that they often arrive at correct answers through incorrect reasoning. [1]

1 Introduction

While large language models (LLMs) have shown great potential as general-purpose task solvers, they tend to be unreliable reasoners (Bang et al., 2023). Prior research suggests that LLMs demonstrate reasoning-like behaviors as the number of parameters increases (Wei et al., 2022). Notably, Chain-of-Thought (CoT) prompting, where LLMs are explicitly instructed to decompose questions into a sequence of logical steps before generating answers, has achieved impressive performance in various reasoning tasks (Wei et al., 2022; Kojima et al., 2022). However, as LLMs function as black-box models, the mechanism behind their reasoning processes remains largely unknown.

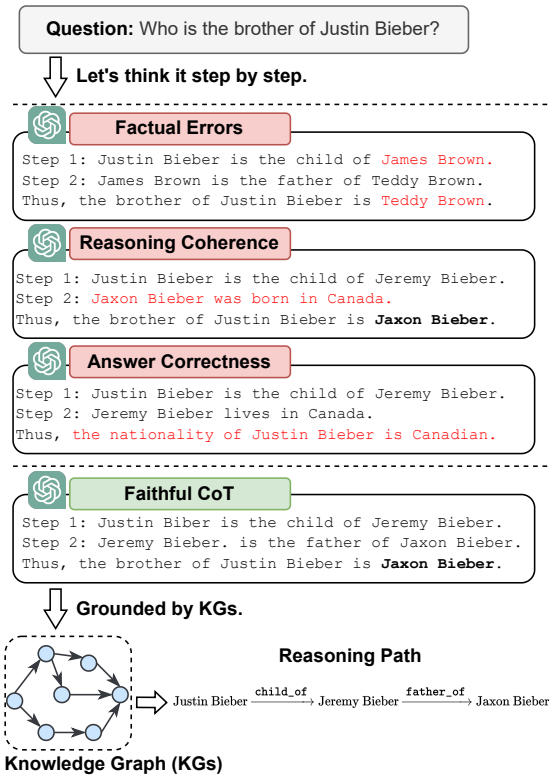

Figure 1: Examples of different reasoning errors and a faithful CoT grounded by knowledge graph.

Previous research measures the reasoning ability of LLMs by reporting their performance, e.g. accuracy, on the downstream tasks that require reasoning (Huang and Chang, 2023). This evaluation strategy cannot provide a direct assessment of the reasoning steps. Hence, it remains unclear whether their strong performance is the result of true reasoning ability or simple heuristics. Recent studies on analyzing CoT reasoning introduce perturbations to prompts, including the injection of invalid reasoning paths, incorrect facts, or the addition of arbitrary symbols to the few-shot examples (Madaan et al., 2023; Wang et al., 2023a; Ye et al., 2023). These studies show that various aspects of prompts, such as query relevance, style patterns, and the correct ordering of reasoning steps,

[*]Equal contribution
[1]Code and data are available at: https://github.com/MinhVuong2000/LLMReasonCert

are more important than the validity of reasoning in the in-context demonstrations. While revealing interesting insights into the reasoning process of LLMs, prompt perturbation-based methods still cannot directly evaluate the correctness of reasoning steps. Automatically verifying CoT reasoning steps is still an open challenge due to the unstructured nature of its freeform rationales.

In this paper, we go beyond evaluating only the final answers to *directly* analyzing the intermediate reasoning steps generated by CoT prompting in multi-hop question-answering (QA) tasks. To tackle the unstructured nature of CoT, we introduce a novel evaluation framework that grounds LLMs' responses in knowledge graphs (KGs) and verify whether it forms a faithful path to the given KGs. Before evaluating more open-ended generative reasoning skills, we design *discriminative* tests to assess whether LLMs can identify faithful reasoning paths when perturbed with factual errors, incoherent, and misguided reasoning steps. Our discriminative evaluation results reveal that LLMs possess certain knowledge of valid reasoning under sufficient knowledge conditions. Building on this observation, we further propose the *generative* evaluation to measure the reasoning ability of LLMs and detect fine-grained reasoning errors (see Figure 1). In the generative evaluation, we instruct LLMs to generate CoT in a structured format, enabling us to parse their responses into a structured reasoning path and validate against KGs. Our ablation study with human experts shows that our framework achieves good accuracy in reasoning path retrieval and evaluation.

We use the proposed evaluation framework to understand the CoT reasoning process of five state-of-the-art LLMs on two complex QA tasks, which require performing multi-step reasoning to answer the questions. Our study reveals that

- LLMs contain sufficient knowledge to conduct reasoning. However, they are still limited in considering the coherence of the reasoning and hallucinations during CoT generation.
- The correct final answer may not necessarily follow from faithful reasoning. We observe a significant gap between answer accuracy and reasoning faithfulness. It highlights the necessity of directly evaluating the reasoning steps rather than solely scoring the final answers.
- The performance gap between the final answer and reasoning worsens as the model size increases. As the answer accuracy also increases with the model size, it suggests that the bigger models may have the knowledge of the final answer without the need to perform reasoning.
- Better prompting strategies such as self-consistency or instructing LLMs with planning can further improve both the final answer and reasoning faithfulness.

2 Preliminaries

Chain-of-thought (CoT) Reasoning Chain-of-thought (CoT) (Wei et al., 2022) is a reasoning framework that prompts LLMs to generate a step-by-step reasoning process $S = \{s_1, s_2, \ldots, s_n\}$ to a question q, where s_i is a natural language sentence describing a step in the reasoning process.

Faithful CoT A faithful CoT should satisfy the following properties (Creswell and Shanahan, 2022): (i) there are no *factual* errors, (ii) the reasoning process is *coherent* (i.e., the conclusion of previous step s_{i-1} should be the prerequisite of the current step s_i), (iii) the reasoning process leads to the *correct* answers. Examples of violations of these properties are shown in Figure 1.

Knowledge Graphs (KGs) Knowledge graphs (KGs) are structured representations of knowledge that contain abundant facts in the form of triples $\mathcal{G} = \{(e_h, r, e_t) \mid e_h, e_t \in \mathcal{E}, r \in \mathcal{R}\}$, where e_h and e_t are head and tail entities, and r is the relation between them; $\mathcal{E}$ and $\mathcal{R}$ are the entity and relation sets respectively. A path in KGs is a sequence of triples: $P = e_0 \xrightarrow{r_1} e_1 \xrightarrow{r_2} \ldots \xrightarrow{r_l} e_l$, connecting the entity e_0 to the entity e_l.

Reasoning Paths Given a question q and the answer a, a *valid* reasoning path $P^* = e_q \xrightarrow{r_1} e_1 \xrightarrow{r_2} \ldots \xrightarrow{r_l} e_a$ is a path that connects the topic entity e_q of q to the answer entity e_a of a in KGs. The reasoning path P^* expresses a valid reasoning process for answering the question according to the KG.

Example 1. Given a question "Who is the brother of Justin Biber?", we can find a valid reasoning path P^* in KGs as: `Justin Bieber` $\xrightarrow{\texttt{child\_of}}$ `Jeremy Bieber` $\xrightarrow{\texttt{father\_of}}$ `Jaxon Bieber`. It indicates: (i) Justin Bieber is the child of Jeremy Bieber, and (ii) Jeremy Bieber is the father of Jaxon Bieber. Thus, the brother of Justin Bieber is Jaxon Bieber.

Faithful CoT Grounded by KGs We verify the faithfulness of the LLMs' CoT reasoning by grounding it with KGs. By treating each reasoning step as a triple in KGs, we convert the CoT into a reasoning path. If a reasoning path starting from the question and ending at the answers exists in KGs, we deem the CoT of LLMs faithful. A grounded example is shown at the bottom of Figure 1.

3 Evaluating the CoT Reasoning of LLMs

We propose a framework to evaluate the CoT reasoning process of LLMs with the help of KGs. Specifically, we propose two evaluation modules: *discriminative evaluation* and *generative evaluation*. The discriminative evaluation investigates whether LLMs possess enough knowledge for conducting faithful reasoning and the generative evaluation further analyzes whether LLMs can provide faithful reasoning process during CoT generation. The overall framework is shown in Figure 3.

3.1 Discriminative Evaluation

The discriminative evaluation aims to analyze whether the LLMs possess enough knowledge to conduct faithful reasoning. i.e. whether it can recognize certain properties of faithful reasoning, including no factual error, coherence and leading to correct answers. We hypothesize that if the LLMs possess sufficient knowledge for faithful reasoning, they should be able to distinguish valid reasoning paths from invalid ones given the question and answer. Following previous studies that evaluate the factual knowledge inside LLMs (Luo et al., 2023b), we feed both the valid and invalid reasoning paths to the LLMs and ask them to predict the validity of these paths. This allows us to assess the reasoning knowledge inside LLMs by analyzing their prediction accuracy. We carefully design prompts to describe the task and instruct LLMs to provide the prediction. Figure 2 shows an example of the zero-shot prompt template.

A *valid* reasoning path is a sequence of triples that can be used to derive the answer of given question. The valid reasoning paths are extracted from the ground-truth reasoning paths[2] $P^* \in \mathcal{P}^*$. We generate three types of *invalid* reasoning paths P' by breaking specific properties of a faithful CoT:

- *Factual error reasoning path:* we construct the invalid paths with factual errors by randomly corrupting entities within the valid reasoning path. This would result in some factual errors in the reasoning path, which are not valid for answering the question.

- *Incoherent reasoning path:* we shuffle the triples of valid paths to construct an incoherent reasoning path. Even though the facts within the paths are accurate, the overall coherence of the path is compromised.

- *Misguided reasoning path:* we randomly sample the paths starting from other questions in KGs. These paths are factually correct and coherent, but they are not related to the questions and lead to incorrect answers.

[2]The ground-truth reasoning paths are constructed from the SPARQL queries provided in the datasets. The detailed construction is shown in Appendix A.

Zero-shot Discriminative Evaluation Prompt

A reasoning path is a sequence of triples that can be used to derive the answer of given question. Given this reasoning path, do you think this is a valid path to derive the answer of given question? If yes please answer "YES", otherwise please answer "NO

Question:
`<Question>`

Answer:
`<Answer>`

Reasoning path:
`<Reasoning Path>`

Figure 2: Discriminative Evaluation Prompt. `<Question>` indicates the question, `<Answer>` denotes the corresponding answer, and `<Reasoning Path>` denotes the input reasoning path, which is verbalized as a structured sentence

To thoroughly assess the reasoning abilities of LLMs, in addition to the zero-shot prompt, we have also developed few-shot, zero-shot CoT, and few-shot CoT prompts. The details of these prompts are shown in Appendix F.1.

Findings The results of the discriminative assessment are shown in §5.1. From the results, we can conclude that LLMs possess enough knowledge to identify factual errors as well as reasoning path relatedness, but have limitations in considering the coherence of reasoning paths and CoT generation. Therefore, we propose the generative evaluation to further assess the faithfulness of CoT reasoning in LLMs' generation.

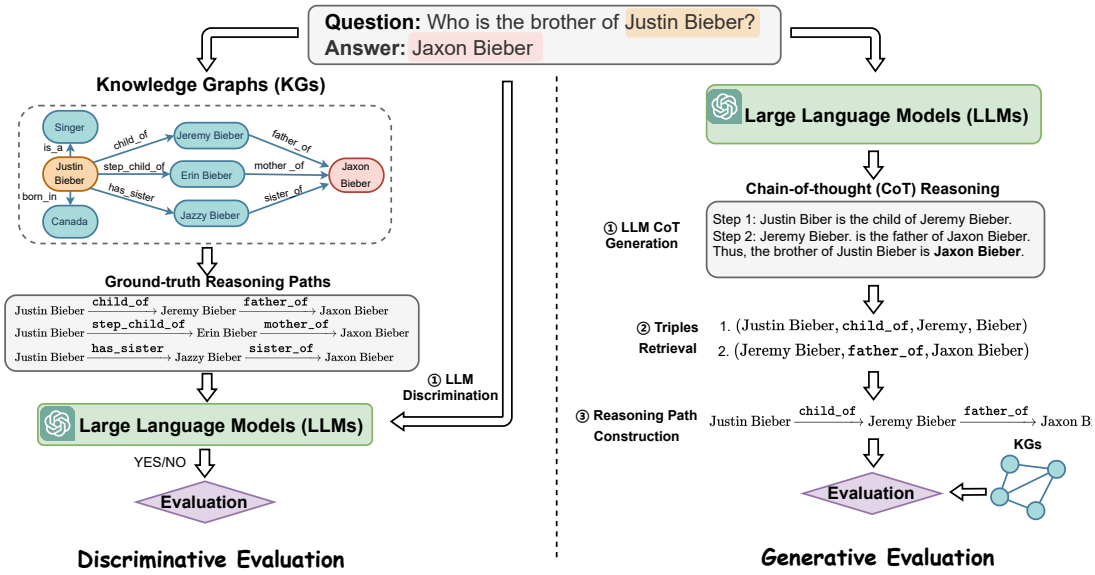

Figure 3: The overall framework of evaluating the CoT reasoning of LLMs, which contains two evaluation modules: discriminative evaluation and generative evaluation. The orange and red rectangles denote the entities mentioned in the question and answer, respectively.

3.2 Generative Evaluation

The generative evaluation aims to assess the faithfulness of the CoT reasoning process *generated* by LLMs. Our main idea is to *ground* LLMs' CoT into KG and verify whether it forms a valid path. To address the challenge of evaluating unstructured CoT, we carefully design a prompting strategy to instruct LLMs to output the CoT in a structured format. This enables us to parse LLM's response into a structured reasoning path, which can then be validated against KG. The example prompts and structured CoT output are provided in Appendix F.2.

Specifically, given a generated CoT response S of question q, we first construct a reasoning path $\hat{P}$ by retrieving triples from the KGs. Then, we evaluate the validity of the reasoning path by checking whether the path coherently connects the question and answer entities in the KGs. The details of these steps are explained in the following subsections.

3.2.1 Reasoning Path Construction

Given a CoT response $S = \{s_1, s_2, \ldots, s_n\}$, we first retrieve a triple[3] $T = (e_h, r, e_t)$ for each step s_i in the CoT response. The retrieved triple is the structural representation of each reasoning step, which can be used to construct the reasoning path for evaluation.

Previous works usually retrieve triples by identifying the entities and relations mentioned in the sentence and linking them to the KGs (Lan et al., 2021; Wang et al., 2021). However, this process is not scalable to KGs. Inspired by the recent fact retrieval method (Baek et al., 2023), we represent the reasoning step s and triples in a unified embedding space and retrieve the triple T based on their embedding similarity.

For all the triples in a KG $\mathcal{G}$, we verbalize each triplet into a sentence by concatenating the entities and relation $x = "e_h\ r\ e_t."$. Then, we use the Sentence-BERT model (Reimers and Gurevych, 2019) to obtain its embedding $h_T = E(x)$. These embeddings are constructed in advance and saved in a vector database for efficient retrieval. Similarly, the embedding of a given reasoning step s is computed as $h_s = E(s)$. Then we retrieve the top-K triples from KG by calculating the embedding similarity between h_s and h_T as:

$$\tau_i = f(h_s, h_{T_i}), \quad T_i = (e_h, r, e_t) \in \mathcal{G}, \qquad (1)$$

where τ_i denotes the similarity score of triple T_i, and $f(\cdot, \cdot)$ is a non-parametric scoring function that measures the similarity between two embeddings. We adopt cosine similarity as the scoring function.

The embedding-based retrieval method may lead to the omission of entities mentioned in the reasoning step. To solve this problem, we also take into account the presence of head and tail entities in the reasoning step in the scoring function. The final

[3]We noticed in almost all cases in our experiments, a CoT step corresponds to one KG triplet. The extension to multiple triplets per CoT step is left for future work.

score for each retrieved triple is calculated as,

$$\tilde{\tau}_i = \frac{\tau_i + \epsilon_h + \epsilon_t}{3}, \quad (2)$$

where ϵ_h and ϵ_t represent the fuzzy-match ratio of head and tail entities in the reasoning step, which are range from 0 to 1, where 0 denotes no existence, 1 denotes a complete match. The overall retrieval process is presented in the Algorithm 1 in the Appendix B.

Thus, we could obtain a set of triples $\mathcal{T} = \{T_1, T_2, \ldots, T_n\}$ for the CoT response S. Then, we construct the reasoning path $\hat{P}$ by connecting the triples in $\mathcal{T}$.

3.2.2 Reasoning Path Evaluation

By evaluating the validity of constructed reasoning paths, we can assess the faithfulness of the CoT reasoning process generated by LLMs. Specifically, we evaluate the validity of the constructed path $\hat{P}$ from three aspects:

- *Factual correctness:* $\hat{P}$ contains factual error if the similarity score $\tilde{\tau}_i$ of any retrieved triples are below a factual threshold σ.
- *Coherence:* given a factually correct path, it is incoherent if there exists a step where its premise is not the conclusion of the previous step.
- *Final answer correctness:* given a factually correct and coherent path, whether the final answer is correct, i.e. matched with ground-truths.

Validity of Reasoning Path The prerequisite and conclusion at each reasoning step are considered head and tail entities, respectively. If the reasoning path $\hat{P}$ can connect the question and answer entities in the KG, we can conclude that it is a valid path. The detailed algorithm is shown in the Algorithm 2 in the Appendix B.

Fine-grained Assessment In addition to the binary evaluation, we also report the minimum edit distance between the constructed reasoning path $\hat{P}$ and the ground-truth path P^*. This serves as a fine-grained assessment of CoT reasoning capability. We adopt a widely used sequence alignment algorithm - Needleman Wunsch algorithm (Needleman and Wunsch, 1970) to obtain continuous alignment scores (i.e., edit distance), which indicate how close the constructed reasoning path is to the ground-truth reasoning paths. If multiple ground-truth paths exist, we report the score against one with the highest match rate. The detailed algorithm is shown in the Algorithm 3 in the Appendix B.

Dataset	**#Test**	**#2hop**	**#≥3hop**
CWQ	1421	1386	35
GrailQA	1813	1528	285

Table 1: Statistic of datasets.

4 Experiment Settings

We use the proposed evaluations to understand the CoT reasoning process of the state-of-the-art LLMs on complex question-answering (QA) tasks which requires performing multi-step reasoning to answer the questions. Through analysis, we seek to answer the following research questions (**RQs**)

- **RQ1**: *Do LLMs have the knowledge of faithful reasoning?* We leverage the discriminative evaluation to test whether LLMs can identify valid reasoning paths. This evaluation focuses on assessing LLMs' knowledge about the properties of faithful reasoning described in Section 2.
- **RQ2**: *Can LLMs express such knowledge to generate faithful reasoning?* Utilizing our generative evaluation framework, we assess the capacity of LLMs to produce coherent and correct reasoning. We also investigate various factors, such as model size and prompting strategies, to understand their impact on reasoning capability.

Dataset We conduct experiments on two QA datasets: Complex WebQuestions (CWQ) (Talmor and Berant, 2018) and GrailQA (Gu et al., 2021) which contain up to 4-hop questions. To evaluate multi-step reasoning capability, we filter out single-hop questions in the test set. Table 1 shows the statistics of the filtered test set. The generated reasoning paths are validated against Freebase (Bollacker et al., 2008) - an open knowledge graph containing around 88M entities, 20K relations, and 126M triples. More details can be found at Appendix C.1.

Large Language Models We evaluate the reasoning capability of several LLMs with instruction-following capability at different sizes, including `Mistral` (7B) (Jiang et al., 2023), `Qwen` (7B, 14B)

(Bai et al., 2023), Vicuna (33B) (Chiang et al., 2023), LLaMA2-Chat (70B) (Touvron et al., 2023) and ChatGPT (175B)[4](OpenAI, 2023). The details of model versions are available in Appendix C.2. We set temperature as 0.7 and top p as 0.9 for generation in all models.

Prompting Strategies We experiment with multiple CoT prompting strategies, including

- **Few-shot CoT** Five examples with structured CoT followed by the answer are added to the prompt (Figure 12 in Appendix F.2).
- **Few-shot CoT with planning (CoT-Plan)** We also explore the ability of LLMs to plan and decompose the relations required to reach the answer before verbalizing the CoT reasoning. In particular, we add the ground-truth *plan* (Luo et al., 2023a) (i.e., a relation path pointing to the answers) into each example. An example prompt is given in Figure 13 in Appendix F.2.
- **Few-shot CoT with self-consistency (CoT-SC)** Beyond the conventional CoT prompting, we also experiment with Self-Consistency (Wang et al., 2023c), a more sophisticated method designed to mitigate the inconsistencies in CoT reasoning by aggregating the final answer through majority votes. In our evaluation, we sample four outputs, and report the maximum performance across all the outputs.

Evaluation Framework Implementation Given a question from the benchmark, in discriminative evaluation, we construct the invalid paths by randomly perturbing the ground-truth paths extracted from SPARQL (Kumar et al., 2019). The implementation detail is described in Appendix A. In generative mode, we use FAISS (Johnson et al., 2019) as the vector database, Sentence-BERT (Reimers and Gurevych, 2019) as the employed embedding model and partial ratio fuzzy matching[5] as the entity scoring function. We retrieve top-10 triples and set the factual threshold σ of 0.7.

Evaluation Metrics For discriminative evaluation, we report the accuracy of detecting valid reasoning paths from invalid ones. For generative evaluation, we report CoT reasoning performance of LLMs with the following metrics: (i) final answer accuracy, (ii) faithful reasoning score, and (iii) minimum edit distance between the generated and ground truth paths. As different LLMs vary in instruction-following capabilities and guardrail implementations, we may encounter responses with unstructured format or abstained answers (Luo et al., 2023b). Therefore, we classify LLMs' responses into four groups: abstained (A), unstructured (U), faithful reasoning (FR), and unfaithful reasoning (UR). We use the F1 score to measure the faithfulness of CoT reasoning where precision and recall are calculated as $P = \frac{\text{FR}}{\text{FR+UR}}$ and $R = \frac{\text{FR}}{\text{FR+UR+A+U}}$. Detailed implementations are described in Appendix C.3. Results of precision and recall are presented in Appendices D.2 and D.3.

LLMs	Size	Zero-shot	Zero-shot CoT	Few-shot	Few-shot CoT
Mistral	7B	87.59	89.88	56.91	69.98
Qwen	7B	74.76	76.13	79.64	73.23
Qwen	14B	88.59	88.86	**88.81**	75.87
Vicuna-1.5	33B	**92.79**	**92.88**	84.91	67.05
LLaMA2-Chat	70B	77.96	80.71	56.99	47.76
ChatGPT	175B	89.86	90.17	87.09	**80.15**

Table 2: Discriminative evaluation results of different LLMs on CWQ. We use binary accuracy as the metric. The best results of each column and row are highlighted in **bold** and underlined.

5 Main Results

5.1 Discriminate Evaluation

Finding 1: LLMs possess knowledge of valid reasoning The overall discriminative evaluation results are shown in Table 2. Based on the results, it is evident that all LLMs achieve a high level of accuracy in distinguishing valid reasoning paths. This indicates that LLMs, which are pre-trained on large-scale corpora, already possess certain knowledge to perform reasoning tasks effectively. However, when using few-shot prompts, there is a noticeable decrease in performance for Mistral and LLaMA2. This could be attributed to the sensitivity of these particular LLMs towards the provided few-shot examples. The detailed results of each perturbation type are illustrated in Appendix D.1, where the accuracy of incoherent paths is lower than other types. We speculate that LLMs cannot capture structural information in the context (Guo et al., 2023). Moreover, the few-shot CoT fails to improve the accuracy in identifying valid paths. We speculate that LLMs are prone to hallucination during CoT generation, resulting in

[4]Previous works mentioned ChatGPT having 175B parameters (Meyer et al., 2023). However, OpenAI still doesn't give any official news about ChatGPT's model size.

[5]https://github.com/seatgeek/thefuzz

LLMs	Size	CWQ				GrailQA			
		Answer↑	Reasoning↑	Gap↓	Edit Dist.↓	Answer↑	Reasoning↑	Gap↓	Edit Dist.↓
		Fewshot CoT							
Mistral	7B	36.45	25.18	**11.27**	69.86	16.35	2.12	14.23	94.03
Qwen	7B	32.52	19.38	13.14	76.78	13.35	1.63	**11.72**	94.69
Qwen	14B	40.39	27.38	13.01	74.49	18.83	2.13	16.70	92.90
Vicuna	33B	44.50	15.92	28.58	74.60	18.26	0.95	17.31	95.39
LLaMA2	70B	49.80	33.98	15.82	62.23	22.05	2.88	19.17	92.58
ChatGPT	175B	**49.85**	**37.13**	12.72	**57.94**	**23.69**	**4.17**	19.52	**90.13**
		Fewshot CoT - Plan							
Mistral	7B	37.14$_{+0.69}$	25.69$_{+0.51}$	**11.45**	70.01	17.30$_{+0.95}$	3.36$_{+1.24}$	13.94	94.46
Qwen	7B	35.35$_{+2.91}$	21.57$_{+2.19}$	13.86	74.74	13.74$_{+0.39}$	2.06$_{+0.43}$	**11.68**	94.61
Qwen	14B	40.86$_{+0.47}$	27.97$_{+0.59}$	12.02	73.68	19.00$_{+0.17}$	2.48$_{+0.35}$	15.43	92.58
Vicuna	33B	48.80$_{+4.30}$	20.24$_{+4.32}$	28.56	63.93	20.84$_{+2.58}$	2.09$_{+1.14}$	18.75	92.12
LLaMA2	70B	50.26$_{+0.46}$	37.08$_{+3.10}$	13.18	57.81	22.35$_{+0.30}$	3.29$_{+0.41}$	19.06	89.61
ChatGPT	175B	**51.74**$_{+1.89}$	**38.60**$_{+1.47}$	13.14	**56.61**	**24.21**$_{+0.52}$	**4.32**$_{+0.15}$	19.11	**89.84**
		Fewshot CoT - SC							
Mistral	7B	40.86$_{+4.41}$	30.38$_{+5.20}$	**10.48**	65.21	16.70$_{+0.35}$	2.60$_{+0.48}$	14.10	94.10
Qwen	7B	34.75$_{+6.08}$	23.21$_{+3.83}$	15.39	74.24	14.00$_{+0.65}$	2.32$_{+0.69}$	**11.68**	94.35
Qwen	14B	41.01$_{+0.62}$	29.26$_{+1.88}$	11.75	73.21	21.00$_{+2.17}$	3.24$_{+1.11}$	17.76	92.50
Vicuna	33B	45.43$_{+2.18}$	21.32$_{+5.40}$	25.36	66.17	18.92$_{+0.66}$	1.88$_{+0.93}$	17.04	94.23
LLaMA2	70B	50.42$_{+0.62}$	37.00$_{+3.02}$	13.42	58.55	22.35$_{+0.30}$	3.29$_{+0.41}$	19.06	91.50
ChatGPT	175B	**51.74**$_{+1.89}$	**40.73**$_{+3.60}$	11.01	**52.57**	**24.97**$_{+1.28}$	**4.86**$_{+0.69}$	20.11	**89.22**

Table 3: Generative evaluation performance of different LLMs on CWQ and GrailQA datasets. F1-scores of the final answer and reasoning accuracy are reported in `Answer` `Reasoning` respectively. The `Gap` column denotes the differences between `Answer` and `Reasoning`. The `Edit Dist.` denotes the edit distance metric described in Appendix C.3. `+x.xx` denotes the improvement in comparison to few-shot CoT.

incorrect predictions. We can conclude that despite having enough reasoning knowledge, LLMs still face limitations in conducting faithful reasoning during CoT generation.

5.2 Generative Evaluation

Table 3 shows the performance of LLMs in generative evaluation mode. Overall, `ChatGPT` demonstrate superior performance in terms of both final answer accuracy and faithfulness of the reasoning. Surprisingly, `Mistral 7B`, despite being the smallest model, exhibits competitive performance comparable to larger models within the <50B range. Furthermore, enhancing prompting strategies with planning (CoT-Plan) and self-consistency (CoT-SC) results in substantial improvements across all LLMs, especially for smaller models.

Finding 2: The correct final answer may not necessarily result from faithful reasoning As shown in Table 3, there is a notable discrepancy between the accuracy of the final answer and the reasoning process. The average gap is 15.76% for CWQ, and 16.44% for GrailQA. While advanced prompting may improve answer and reasoning accuracy, this performance gap mostly stays consistent. Interestingly, `Vicuna` achieves reasonable answer accuracy but has the lowest reasoning performance of all the models, suggesting its reasoning ability is inferior, even when compared to small models like `Mistral` and `Qwen-7B`. This finding highlights the inadequacy of relying on final answer accuracy as a proxy to gauge reasoning ability.

Finding 3: The reasoning gap worsens as the model size increases It can be seen that the reasoning performance increases gradually with model size, proving the reasoning ability of bigger models. However, the gap between answer and reasoning performance also gradually increases with model size and the correctness of the answer. While `LLaMA2-70B` and `ChatGPT` rank first in performance, their gaps are also the highest. Meanwhile, the smallest-size models, including `Mistral-7B` and `Qwen-7B`, hold the lowest gap on CWQ and GrailQA, respectively. We speculate that larger LLMs may grasp the question context better or have more knowledge to provide the correct answer directly without performing reasoning.

Finding 4: Better prompting strategy can improve both the answer and reasoning accuracy

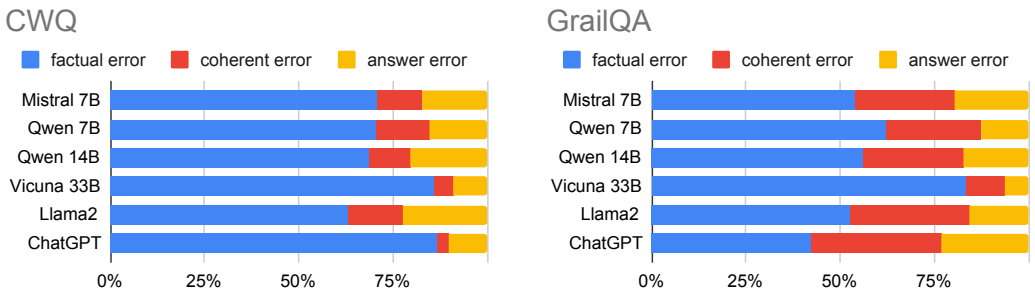

Figure 4: The breakdown of reasoning error types in CWQ and GrailQA.

Reasoning Types	Precision	Recall	F1
Faithful Reasoning	95.42	83.88	89.28
Unfaithful Reasoning	86.12	96.01	90.80

Table 4: Precision, Recall, and F1 score of the framework on human annotated datasets.

without worsening the reasoning gap The use of enhanced prompting strategies such as CoT-Plan and CoT-SC lead to improvement of both the answer and reasoning accuracy across most LLMs. However, the gap between them remains consistent, regardless of prompt strategy.

5.3 Analysis

Reasoning Errors We present a detailed breakdown of the reasoning errors in Figure 4. The results reveal that factual errors account for the majority of errors, indicating that LLMs tend to generate incorrect information during reasoning. As GrailQA is a more complex dataset, LLMs have a higher percentage of coherence errors on GrailQA than CWQ. Interestingly, even when the generated CoT paths are free from factual and coherent errors, LLMs may fail to produce correct answers, evidenced by a substantial amount of answer errors. Error case examples are shown in Appendix D.4.

Ablation Study To ensure the effectiveness of our generative evaluation framework, we randomly select 100 CoT responses generated by `ChatGPT` in CWQ dataset and asked two human experts to evaluate the constructed reasoning path. The detail of the human evaluation study is described in the appendix E. The results presented in Table 4 demonstrate that our method can accurately detect both faithful and unfaithful reasoning paths. This further confirms the efficacy of our approach in evaluating CoT reasoning.

Parsing Error While we carefully design prompts to instruct LLMs to generate a structured CoT, there are still corner cases where LLMs generate unstructured and abstention responses due to their unpredicted behaviors. As reported in Appendix D.2, the unstructured and abstention rates are less than 20% in CWQ dataset and can be mitigated with CoT-Plan and CoT-SC.

6 Related Work

Reasoning with LLMs While LLMs have proven to offer a variety of reasoning abilities, they still tend to hallucinate facts, making them unreliable and imperfect (Qiao et al., 2022). Several studies have concentrated on improving their reasoning capacity through prompting (Wang et al., 2023c; Ye and Durrett, 2022; Wiegreffe et al., 2022). CoT (Wei et al., 2022) is a prompting approach that has demonstrated notable improvements in reasoning performance. A significant enhancement compared to CoT, self-consistency (Wang et al., 2023c), is a scheme where multiple CoTs are generated and the most consistent self-generated answer is selected. Recently, self-consistency was extended with Tree of Thoughts (ToT) (Yao et al., 2023), which models the reasoning process with a tree. ToT allows LLMs to interactively backtrack and explore alternate chains of reasoning, avoiding getting stuck on a single line of incorrect reasoning. Ye and Durrett (2022) mitigate the effect of unreliable rationales by calibrating the prediction probability based on the factuality of CoT. Wiegreffe et al. (2022) train a Seq2Seq model to filter out unacceptable rationale. Liu et al. (2021) utilize GPT-3 (Brown et al., 2020) with few-shot prompting to generate knowledge and prompts the downstream language models.

Reasoning Evaluation Evaluation of the reasoning ability of LLMs has been undertaken for two main purposes: to enhance the reasoning ability of LLMs (Lyu et al., 2023; Li et al., 2023; Tyen et al.,

2023; Chen et al., 2023) and to quantify the reasoning ability of LLMs (Wang et al., 2023b; Atanasova et al., 2023). For instance, Huang and Chang (2023) gauges the reasoning ability of LLMs by assessing their performance on reasoning benchmarks such as GSM8K and BIG-bench for downstream tasks. However, this evaluation strategy is unable to offer a direct assessment of the reasoning steps. Tyen et al. (2023) release the BIG-Bench Mistake dataset, which includes logical errors in CoT reasoning steps. Using this benchmark, Tyen et al. (2023); Chen et al. (2023) illustrate the inability of state-of-the-art LLMs to identify mistakes and reasoning errors, even in unequivocal cases.

7 Conclusion

We propose an evaluation framework to understand the CoT reasoning capability of LLMs beyond the sole assessment of final answer accuracy. With the help of a KG and a careful prompting strategy, we can turn the unstructured CoT into a structured format for automatic evaluation. Our framework consists of two evaluation modules: (i) a discriminative module that isolates the effects of different reasoning errors to verify LLMs' knowledge about reasoning, and (ii) a generative module to assess the generated CoT reasoning. While LLMs showcase remarkable capabilities in generating correct answers, our study emphasizes the need for more nuanced evaluations of their reasoning processes. Addressing the gap between the final answer and reasoning accuracy remains a critical area for further exploration in enhancing the true reasoning capabilities and interpretability of LLMs.

Limitation

The limitation of our work includes:

- We consider a CoT step corresponding to one KG triple and a single correct answer for each question. However, LLMs may generate a sentence containing more than two relations. This can be tackled by returning top-K candidates from Algorithm 1 and a dynamic program algorithm expanded from Algorithm 2.
- We assume the availability of completed knowledge graphs (KGs) for factual retrieval. We leave the incorporation of the knowledge graph completion methods to improve the comprehension of the retrieval algorithm as future works.
- This study mainly focuses multi-hop reasoning questions over knowledge graphs (KGs). Nevertheless, there exist intricate reasoning inquiries, such as those in mathematics or logical reasoning, which involve unstructured replies that are not easily resolvable through KGs. The establishment of verification frameworks for diverse forms of reasoning queries plays a significant role in enhancing the reliability and utility of responses generated by LLMs. This aspect is still an open and challenging problem, requiring extensive explorations within the research community.

Acknowledgments

The authors are grateful to the anonymous reviewers for their helpful comments. The computational resources of this work are supported by the Multimodal Australian ScienceS Imaging and Visualisation Environment (MASSIVE)[6]. This material is based on research partially sponsored by the DARPA Assured Neuro Symbolic Learning and Reasoning (ANSR) program under award number FA8750-23-2-1016 and the DARPA Knowledge Management at Scale and Speed (KMASS) program under award number HR00112220047. Minh-Vuong Nguyen was supported by VinAI Research.

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Jinheon Baek, Alham Fikri Aji, Jens Lehmann, and Sung Ju Hwang. 2023. Direct fact retrieval from knowledge graphs without entity linking. In *EMNLP*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang

[6] www.massive.org.au

Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023. Felm: Benchmarking factuality evaluation of large language models. *arXiv preprint arXiv:2310.00741*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488. ACM.

Jiayan Guo, Lun Du, and Hengyu Liu. 2023. Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. *arXiv preprint arXiv:2305.15066*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Vishwajeet Kumar, Yuncheng Hua, Ganesh Ramakrishnan, Guilin Qi, Lianli Gao, and Yuan-Fang Li. 2019. Difficulty-controllable multi-hop question generation from knowledge graphs. In *International Workshop on the Semantic Web*.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A survey on complex knowledge base question answering: Methods, challenges and solutions. In *IJCAI*, pages 4483–4491.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023a. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arxiv:2310.01061*.

Linhao Luo, Trang Vu, Dinh Phung, and Reza Haffari. 2023b. Systematic assessment of factual knowledge in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13272–13286, Singapore. Association for Computational Linguistics.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes chain-of-thought prompting effective? a counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535, Singapore. Association for Computational Linguistics.

Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. 2023. Chatgpt and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

OpenAI. 2023. Gpt-4 technical report.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Gladys Tyen, Hassan Mansoor, Peter Chen, Tony Mak, and Victor Cărbune. 2023. Llms cannot find reasoning errors, but can correct them!

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Jianing Wang, Qiushi Sun, Nuo Chen, Xiang Li, and Ming Gao. 2023b. Boosting language models reasoning with chain-of-knowledge prompting.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2021. Retrieval, re-ranking and multi-task learning for knowledge-base question answering. In *EACL*, pages 347–357.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Mengyu Ye, Tatsuki Kuribayashi, Jun Suzuki, Goro Kobayashi, and Hiroaki Funayama. 2023. Assessing step-by-step reasoning against lexical negation: A case study on syllogism. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14753–14773, Singapore. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

A Ground-truth Reasoning Path Construction

To obtain graph and ground truth reasoning paths for each question, we utilize SPARQL query, topic entities, answer entities provided in datasets to construct subgraphs, then find paths from topic entities to answer entities as the ground-truth reasoning paths. In particular,

- We replace `SELECT` query in each SPARQL query into a `CONSTRUCT` query to return the corresponding graph. An example `CONSTRUCT` query and its returned graph from the CWQ dataset is shown in example 2.
- In the returned graph, we use NetworkX[7] library to find the paths starting at one of the topic entities and ending at one of the answer entities provided in original dataset. As a result, we obtain a list of ground-truth reasoning paths used for both discriminative and generative evaluation.

Example 2. An example `CONSTRUCT` query and the corresponding returned graph.

```
CONSTRUCT query
PREFIX ns: <http://rdf.freebase.com/ns/>
CONSTRUCT WHERE { FILTER (?x != ?c)
FILTER (!isLiteral(?x) OR lang(?x) = '' OR
langMatches(lang(?x), 'en'))
?c ns:location.country.administrative_divisions
ns:m.02g__4 .
?c ns:location.country.languages_spoken ?x .
}
Corresponding returned graph
m.03gyl location.country.administrative_divisions
m.02g__4 .
m.03gyl location.country.languages_spoken
m.02dhwl .
m.03gyl location.country.languages_spoken
m.064_8sq .
```

B Generative Evaluation

The details of triple retrieval, reasoning path evaluation, and fine-grained path evaluation are shown in the Algorithms 1 to 3, respectively.

Algorithm 1: Triple Retrieval Algorithm

Input: Step s; Top-K; embedding model E; knowledge graph $\mathcal{G}$.
Output: Triple T

$h_s \leftarrow E(s)$
$[T_i, \tau_i]_{i=1}^{K} \leftarrow \mathcal{G}.\text{search}(h_s, k)$
$\mathcal{C} = []$
for $i \leftarrow 1$ **to** K **do**
  $\epsilon_h, \epsilon_t \leftarrow$ Fuzzy-match(T_i, s)
  $\tilde{\tau}_i \leftarrow \frac{\tau_i + \epsilon_h + \epsilon_t}{3}$
  $\mathcal{C}.\text{append}((\tilde{\tau}_i, T_i))$
end
$T = \arg\max_{T_i \in \mathcal{C}} \tilde{\tau}_i$

C Experiment Settings

C.1 Datasets

We adopt two benchmark KGQA datasets: Complex WebQuestions (CWQ)[8](Talmor and Berant, 2018) and GrailQA[9](Gu et al., 2021) in this work, especially we use the test split of CWQ, and validation split of GrailQA. We only keep questions requiring more than 2-hop reasoning. The number of questions is shown in Table 1. Both CWQ and GrailQA can be reasoned based on Freebase KGs[10](Bollacker et al., 2008). To reduce the KG size, we combine the subgraphs obtained from SPARQL queries as the final KG. The detail of subgraphs extraction is shown in Appendix A.

C.2 Large Language Models

The LLMs used in experiments are shown in Table 5. We utilize available checkpoints from HuggingFace[11].

C.3 Evaluation Metrics

Faithful Reasoning Score. To enable robust and truthful reasoning, many LLMs adopt the guardrail techniques to *abstain* from providing answers when they are uncertain (Liu et al., 2023; Luo et al., 2023b). Also, LLMs exhibit different instruction-following capabilities, which may result in the CoT response in *unstructured* format. Thus, we need to consider both the *abstained* and *unstructured* responses in the evaluation. We define the precision and recall as,

$$\text{Precision} = \frac{\#\text{correct}}{\#\text{correct}+\#\text{incorrect}}, \tag{3}$$

$$\text{Recall} = \frac{\#\text{correct}}{\#\text{correct}+\#\text{incorrect}+\#\text{abstained}+\#\text{unstructured}}. \tag{4}$$

The F1 score is then the harmonic average of the precision and recall. The details of abstained

[7] https://networkx.org/
[8] https://www.tau-nlp.sites.tau.ac.il/compwebq
[9] https://huggingface.co/datasets/grail_qa
[10] https://github.com/microsoft/FastRDFStore
[11] https://huggingface.co/

Algorithm 2: Reasoning Path Evaluation

Input: Reasoning path $\hat{P}$; threshold σ; knowledge graph $\mathcal{G}$; ground answer entity $\mathcal{A}$.
Output: Validity v.

factual_error $\leftarrow False$
$order\_error \leftarrow False$
$answer\_error \leftarrow answer_{\hat{P}} != \mathcal{A}$
for $T_i \in \hat{P}$ **do**
  if $\tau_i < \sigma$ **then**
    factual_error $\leftarrow True$
  else
    ~~continue~~
    if $head_{T_i} != tail_{T_{i-1}}$ **then**
      $order\_error \leftarrow True$
    end
  end
end
$coherent\_error \leftarrow \neg factual\_error \wedge order\_error$
$answer\_error \leftarrow \neg coherent\_error \wedge answer\_error$
if $factual\_error \vee coherent\_error \vee answer\_error$ **then**
  $v \leftarrow False$
else
  $v \leftarrow True$
end
return v

Algorithm 3: Fine-grained Path Evaluation

Input: Reasoning path $\hat{P}$; Ground-truth paths $\mathcal{P}^*$
Output: Edit distance u

$m \leftarrow 0$
$u \leftarrow 0$
for $P^* \in \mathcal{P}^*$ **do**
  $u' \leftarrow \text{NeedlemanWunsch}(\hat{P}, P^*)$
  if $u' < u$ **then**
    $u \leftarrow u'$ ▷ Get minimum edit distance.
  end
end
return u

LLM	Model Implementation
Mistral 7B	mistralai/Mistral-7B-Instruct-v0.1
QWen 7B	Qwen/Qwen-7B-Chat
QWen 14B	Qwen/Qwen-14B-Chat
Vicuna 33B	lmsys/vicuna-33b-v1.3
LlaMA2 70B	meta-llama/Llama-2-70b-chat
ChatGPT	GPT-3.5-turbo

Table 5: Details of used LLMs.

and unstructured response detection are in Appendices C.4 and C.5.

Answer Accuracy The answers generated by LLM can have a different length than the ground-truth, so we utilize fuzzy matching to check the correctness of answer. It checks whether the generated answer appears in the ground-truth answers, and vice versa.

Edit Distance We further look into the fine-grained evaluation of the reasoning path by calculating the edit distance (Needleman and Wunsch, 1970): The minimum number of edits (add/remove the reasoning steps) required to convert the CoT path to the ground-truth path.

C.4 Abstained Answer Detection

We detect abstained answers through following specific keywords in LLMs' responses:

List of Abstention keywords

not have knowledge
more information
need more
unknown
cannot
sorry
impossible
not possible
unable
unclear

C.5 Unstructured Answer Detection

As the instruction shown in prompts 12 and 13, a structured CoT should follow the following format:

A structured answer

1. <step1>
2. <step2>
...
(T|t)he answer(to the question)? is ?
?(.*?)?.?

Therefore, the CoT responses that do not match the pattern are identified as unstructured.

D Additional Results

D.1 Detailed Results of Discriminative Evaluation

We illustrate the discriminative evaluation results of different types of reasoning paths in Table 6. For each error types, LLMs reach the best performance under factual errors and misguided reasoning path. However, the performance under incoherent paths is lower, which could due to the limits of LLMs in understanding structural context. Besides, the performance of valid reasoning path is slightly lower than invalid reasoning path. Because it requires both three properties (i.e., factual errors, incoherence, misguidance) satisfied at the same time.

D.2 Detailed Results of Precision

Based on the definition in eq. (3), precision ignores the abstained and unstructured responses. The detailed precision results as well as the abstained and unstructured ratio are shown in Figures 5 to 7.

The precision of the answer gradually improves with the size of the model. However, Mistral 7B and Qwen 14B are competitive in terms of reasoning performance compared to models with ten times more parameters. While Vicuna 33B has an intermediate model size, its performance is low. It is also worth noting that Qwen variants exhibit a high rate of abstention. This leads to abstaining from answering uncertain questions and achieving higher precision, but at the cost of reduced recall for correct answers and reasoning.

D.3 Detailed Results of Recall

The recall is calculated on all types of responses. Thus, we directly report the recall in Table 7.

D.4 Case Studies

We present the detailed cases of different error types in Table 8.

E Human Evaluation

This section aims to be two-fold. Firstly, it provides a sanity check of the proposed prompt in instructing LLMs to generate CoT in a structured format. Secondly, we aim to construct a small test set to evaluate the capability of the proposed framework in detecting the faithful CoT.

E.1 Annotation

Firstly, a sample of 100 questions is chosen randomly from the CWQ dataset along with the responses generated by ChatGPT using proposed prompts. Each question in the sample is matched with reference answers, which include the final answer provided in the datasets, as well as all the ground-truth reasoning paths extracted during the preprocessing phase outlined in appendix A. This data is supplied to annotators for classification purposes.

The annotators are two PhD students who are familiar with the related works and have experience working with knowledge graphs. To prepare them for the task, we provide them the definition of faithful reasoning and error types. Given sample cases that consist of the question, answer, the ground-truth reasoning paths (the desired reasoning steps), and the model's initial CoT response, two human experts performed a trial run to align their understanding and coding criteria for faithful CoT responses by answering two following questions:

- `Q1: If LLM response is an incorrect final answer?` True labeling if LLM final answer does not hit the reference answers; False otherwise.
- `Q2: If LLM CoT is an incorrect reasoning?` True labeling if there exists any step its relation, subject/object is in ground-truth reasoning paths but there is no link between them or the object/subject/relation is not in the ground-truth reasoning paths; False otherwise.

The human annotation dataset acquired ultimately comprises 100 samples, with each sample consisting of the question, ground truth paths, LLM answers, a column indicating the annotated incorrect answer (1 for True labeling of Q1, 0 otherwise), and another column indicating the annotated incorrect reasoning (1 for True labeling of Q2, 0 otherwise).

E.2 Framework Evaluation

Given the human annotation data provided above, our generative evaluation framework is assessed to ensure its quality prior to its application in all experiments. Initially, the responses from the LLM are inputted into the generative evaluation mode to obtain predictions for Q1 (predicted incorrect answer) and Q2 (predicted incorrect reasoning). Consequently, these two outcomes are compared with the annotated dataset to assess the precision, recall, and f1_score for the final score. The metrics are shown in table 4.

LLMs	Size	Valid Reasoning Path				Factual Errors Reasoning Path			
		Zero-shot	Zero-shot-Cot	Few-shot	Few-shot CoT	Zero-shot	Zero-shot-Cot	Few-shot	Few-shot CoT
Mistral	7B	79.01	72.32	98.78	74.54	98.68	99.46	74.72	90.67
Qwen	7B	94.92	91.84	85.05	58.60	94.71	94.45	95.25	92.10
Qwen	14B	62.41	61.97	75.74	21.69	100.00	99.96	99.78	99.31
Vicuna	33B	83.54	79.32	91.80	19.02	99.37	99.77	98.96	98.40
LLaMA2	70B	18.79	27.84	92.50	57.52	99.54	99.50	59.44	72.48
ChatGPT	175B	70.27	74.57	86.56	71.45	99.72	99.72	99.70	95.86
LLMs	Size	Incoherent Reasoning Path				Misguided Reasoning Path			
		Zero-shot	Zero-shot-Cot	Few-shot	Few-shot CoT	Zero-shot	Zero-shot-Cot	Few-shot	Few-shot CoT
Mistral	7B	90.11	92.10	39.92	53.08	82.55	95.62	14.20	61.61
Qwen	7B	61.37	61.82	73.06	66.37	48.02	56.40	65.18	75.85
Qwen	14B	93.56	94.91	88.56	85.49	98.38	98.61	91.14	97.00
Vicuna	33B	90.21	93.62	65.65	56.56	98.04	98.82	83.21	94.23
LLaMA2	70B	95.92	97.43	44.41	18.27	97.58	98.05	31.60	42.78
ChatGPT	175B	93.36	90.57	76.76	59.01	96.07	95.82	85.34	94.28

Table 6: Discriminative evaluation results of different LLMs on CWQ dataset. We use the binary accuracy as metric.

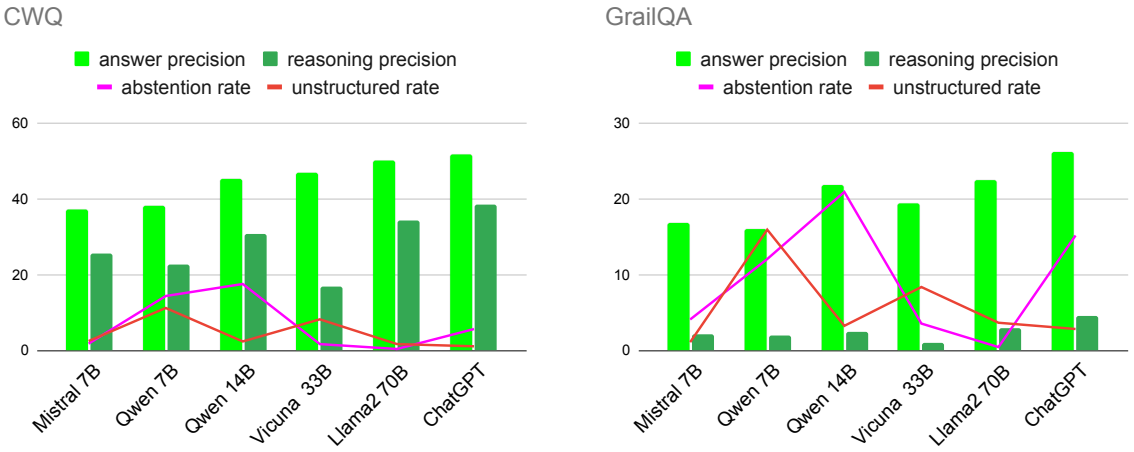

Figure 5: The precision of LLMs using few-shot CoT prompt.

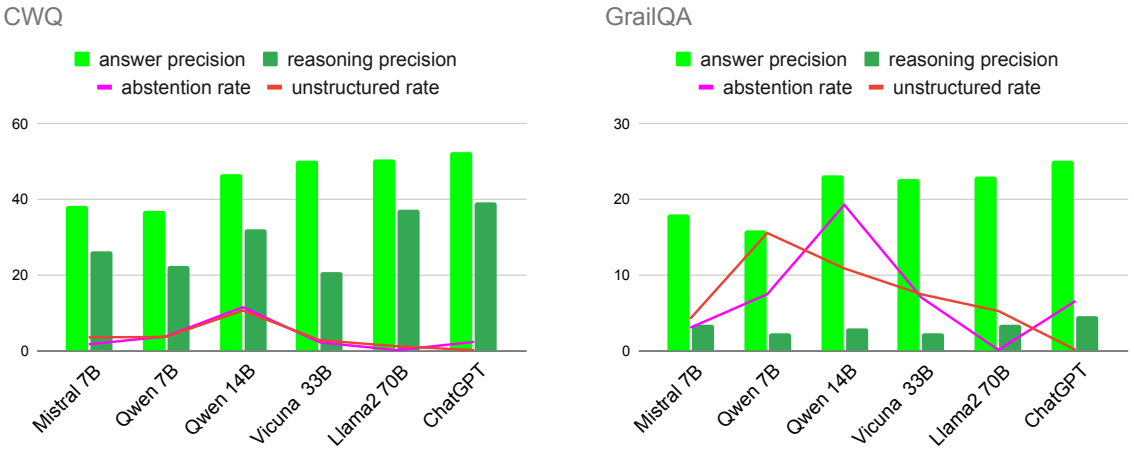

Figure 6: The precision of LLMs using few-shot CoT - Plan prompt.

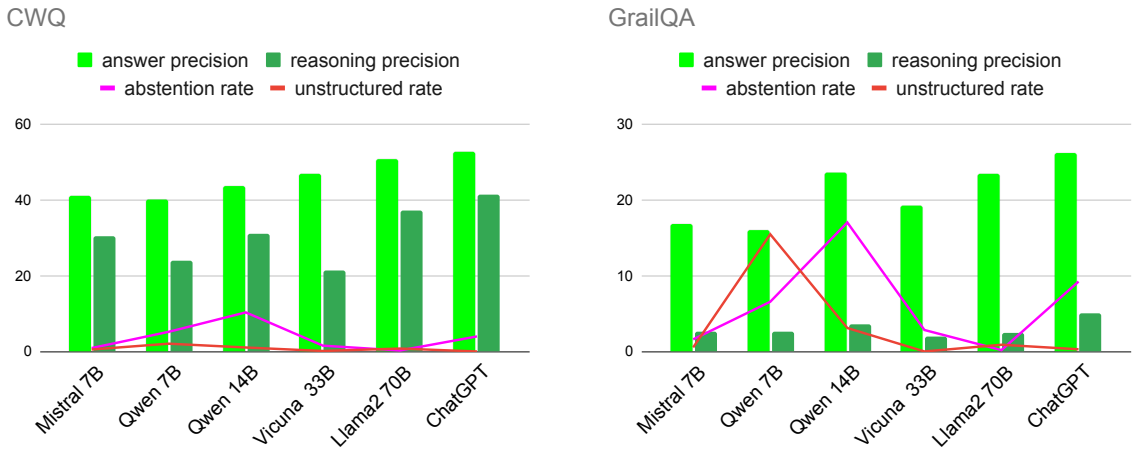

Figure 7: The precision of LLMs using few-shot CoT - SC prompt.

LLMs	Size	CWQ		GrailQA	
		Answer	Reasoning	Answer	Reasoning
		Fewshot CoT			
Mistral	7B	35.67	24.64	15.93	2.07
Qwen	7B	28.36	16.9	11.48	1.4
Qwen	14B	36.37	24.65	16.55	1.87
Vicuna	33B	42.3	15.13	17.17	0.89
LLaMA2	70B	49.29	33.63	21.6	2.82
ChatGPT	175B	48.17	35.88	21.55	3.8
		Fewshot CoT - Plan			
Mistral	7B	36.18	25.02	16.67	3.24
Qwen	7B	34.13	20.78	12.16	1.82
Qwen	14B	36.35	24.88	16.14	2.11
Vicuna	33B	47.6	19.74	19.33	1.94
LLaMA2	70B	49.94	36.84	21.76	3.2
ChatGPT	175B	51.11	38.13	23.4	4.18
		Fewshot CoT - SC			
Mistral	7B	40.57	30.17	16.52	2.57
Qwen	7B	37.2	22.37	12.45	2.07
Qwen	14B	38.68	27.6	18.88	2.91
Vicuna	33B	46.29	21.14	18.65	1.86
LLaMA2	70B	50.16	36.81	21.76	3.2
ChatGPT	175B	50.71	39.92	23.78	4.63

Table 7: The recall of LLMs on CWQ and GrailQA datasets.

F Prompts

F.1 Discriminative Evaluation Prompts

We present the zero-shot, zero-shot CoT, few-shot, and few-shot CoT prompts in Figures 8 to 11.

F.2 Generative Evaluation Prompts

We present the few-shot, few-shot plan prompts in Figures 12 and 13.

Correct Answer & Correct Reasoning	Question: People from the country that contains Corfu speak what language? Groundtruth: [["Corfu","location.country.administrative_divisions","Greece"], ["Greece","location.country.languages_spoken","Greek Language"]] LLM CoT Answer: 1. Corfu is an island located in Greece. 2. Greece is a country that contains Corfu. 3. The primary language spoken in Greece is Greek. So the answer is (Greek). Reasoning path from our framework: 1. [["Corfu","location.country.administrative_divisions","Greece"] 2. "Greece","location.country.languages_spoken","Greek Language"]]
Correct Answer & Incorrect Reasoning	Question: What is the currency in the country where the Brazilian cruzado is also used? Groundtruth: [["Brazilian cruzado","location.country.currency_formerly_used","Brazil"], ["Brazil","location.country.currency_used","Brazilian real"]] LLM CoT Answer: 1. The Brazilian cruzado was the currency of Brazil from 1986 to 1989. 2. Brazil is the country where the Brazilian cruzado was used. So the answer is (Brazilian real). Reasoning path from our framework: 1. ["Brazilian cruzado","location.country.currency_formerly_used","Brazil"] 2. ["Brazilian cruzado","location.country.currency_formerly_used","Brazil"]
Incorrect Answer & Correct Reasoning	Question: Where was the main artist featured in the Rihanna: Live in Concert Tour raised? Groundtruth: [["Rihanna: Live in Concert Tour","music.artist.concert_tours","Rihanna"], ["Rihanna","people.person.place_of_birth","Saint Michael Parish"]] LLM CoT Answer: 1. The main artist featured in the Rihanna: Live in Concert Tour is Rihanna. 2. Rihanna was raised in Barbados. So the answer is (Barbados). Reasoning path from our framework: 1. ["Rihanna: Live in Concert Tour","music.artist.concert_tours","Rihanna"] 2. ["Rihanna","people.person.nationality","Barbados"]
Incorrect Answer & Incorrect Reasoning	Question: Where did the author who published "The Story of My Life" die? Groundtruth: [["The story of my life","book.author.book_editions_published","Helen Keller"], ["Helen Keller","people.deceased_person.place_of_death","Easton"]] LLM CoT Answer: 1. "The Story of My Life" is a book written by Helen Keller. 2. Helen Keller died in Connecticut, USA. So the answer is (Connecticut, USA). Reasoning path from our framework: 1. ["The story of my life","book.author.book_editions_published","Helen Keller"], 2. ["Helen Keller","people.deceased_person.place_of_death","Easton"],p=0.57

Table 8: Examples of answer and reasoning accuracy of generative evaluation framework.

Zero-shot Discriminative Evaluation Prompt

A reasoning path is a sequence of triples that can be used to derive the answer of given question. A valid reasoning path should follow these rules:

1. No factual errors: Each triple in the reasoning path should adhere to real-world factual knowledge.
2. Coherence: The tail entity of the previous triple should be the head entity of the next triple.
3. Correctness: The reasoning path should lead to the correct answer at the last tail entity.

Given this reasoning path, do you think this is a valid path to derive the answer of given question? If yes please answer "YES", otherwise please answer "NO".

Question:
`<Question>`

Answer:
`<Answer>`

Reasoning path:
`<Reasoning Path>`

Figure 8: The zero-shot prompt used for discriminative evaluation

Zero-shot CoT Discriminative Evaluation Prompt

A reasoning path is a sequence of triples that can be used to derive the answer of given question. A valid reasoning path should follow these rules:

1. No factual errors: Each triple in the reasoning path should adhere to real-world factual knowledge.
2. Coherence: The tail entity of the previous triple should be the head entity of the next triple.
3. Correctness: The reasoning path should lead to the correct answer at the last tail entity.

Given this reasoning path, do you think this is a valid path to answer the question? If yes please answer "YES", otherwise please answer "NO". Let's think it step by step.

Question:
`<Question>`

Answer:
`<Answer>`

Reasoning path:
`<Reasoning Path>`

Figure 9: The zero-shot CoT prompt used for discriminative evaluation.

Few-shot Discriminative Evaluation Prompt

A reasoning path is a sequence of triples that can be used to derive the answer of given question. A valid reasoning path should follow these rules:

1. No factual errors: Each triple in the reasoning path should adhere to real-world factual knowledge.
2. Coherence: The tail entity of the previous triple should be the head entity of the next triple.
3. Correctness: The reasoning path should lead to the correct answer at the last tail entity.

Given this reasoning path, do you think this is a valid path to answer the question? If yes please answer "YES", otherwise please answer "NO". Here are some examples:

Input:
Question:
What type of government is used in the country with Northern District?

Answer:
Parliamentary system

Reasoning Paths:
Step 1: Northern District -> location.administrative_division.first_level_division_of -> Israel
Step 2: Israel -> government.form_of_government.countries -> Parliamentary system

Output:
YES

Input:
Question:
Where is the home stadium of the team who won the 1946 World Series championship?

Answer:
Busch Stadium

Reasoning Paths:
Step 1: 1946 World Series -> sports.sports_team.championships -> St. Louis Cardinals
Step 2: St. Louis Cardinals -> sports.sports_team.arena_stadium -> Roger Dean Stadium

Output:
NO

Input:
Question:
In which American Southern City did the ""Downs"" composer die?

Answer:
New Orleans

Reasoning Paths:
Step 1: Alex Chilton -> people.deceased_person.place_of_death -> New Orleans
Step 2: Downs -> music.composer.compositions -> Alex Chilton

Output:
NO

Input:
Question:
Where was the main artist featured in the Rihanna: Live in Concert Tour raised?

Answer:
Saint Michael Parish

Reasoning Paths:
Step 1: Rihanna: Live in Concert Tour -> music.concert_tour.artist -> Rihanna
Step 2: Rihanna -> music.track_contribution.role -> Vocals

Output: NO

Input:
Question:
`<Question>`

Answer:
`<Answer>`

Reasoning path:
`<Reasoning Path>`

Output:

Figure 10: The few-shot prompt used for discriminative evaluation.

Few-shot CoT Discriminative Evaluation Prompt

A reasoning path is a sequence of triples that can be used to derive the answer of given question. A valid reasoning path should follow these rules:

1. No factual errors: Each triple in the reasoning path should adhere to real-world factual knowledge.
2. Coherence: The tail entity of the previous triple should be the head entity of the next triple.
3. Correctness: The reasoning path should lead to the correct answer at the last tail entity.

Given this reasoning path, do you think this is a valid path to answer the question? If yes please answer "YES", otherwise please answer "NO". Here are some examples:

Input:
Question:
What type of government is used in the country with Northern District?

Answer:
Parliamentary system

Reasoning Paths:
Step 1: Northern District -> location.administrative_division.first_level_division_of -> Israel
Step 2: Israel -> government.form_of_government.countries -> Parliamentary system

Output:
This reasoning path indicates that:
1. "Northern District" is a location within some country.
2. The reasoning path mentions "Northern District -> location.administrative_division.first_level_division_of -> Israel," indicating that the Northern District is part of Israel.
3. It further states "Israel -> government.form_of_government.countries," suggesting that Israel's form of government is being discussed.
4. The last part of the reasoning path indicates that Israel has a "Parliamentary system."
Based on the provided reasoning paths, it can be concluded that the type of government used in the country with the Northern District (Israel) is a Parliamentary system. Therefore, the path adheres to the definition of valid reasoning path. The answer is "YES"

Input:
Question:
Where is the home stadium of the team who won the 1946 World Series championship?

Answer:
Busch Stadium

Reasoning Paths:
Step 1: 1946 World Series -> sports.sports_team.championships -> St. Louis Cardinals
Step 2: St. Louis Cardinals -> sports.sports_team.arena_stadium -> Roger Dean Stadium

Output:
This reasoning path indicates that:
1. St. Louis Cardinals as the team that won the 1946 World Series
2. Roger Dean Stadium is the stadium associated with the St. Louis Cardinals.
However, the home stadium of St. Louis Cardinals is Busch Stadium, not Roger Dean Stadium. Therefore, this path contains factual errors. The answer is "NO".

Input:
Question:
In which American Southern City did the ""Downs"" composer die?

Answer:
New Orleans

Reasoning Paths:
Step 1: Alex Chilton -> people.deceased_person.place_of_death -> New Orleans
Step 2: Downs -> music.composer.compositions -> Alex Chilton

Output:
This reasoning path indicates that:
1. Alex Chilton was dead in New Orleans
2. The composition of Downs is Alex Chilton

Even through the first step of reasoning path leads to the correct answer, which is New Orleans, the reasoning path is not coherent. To answer the question, we need to first the composition of Downs which is Alex Chilton, then we find the death place of Alex Chilton which is New Orleans. Therefore, the answer is "NO".

Input:
Question:
Where was the main artist featured in the Rihanna: Live in Concert Tour raised?

Answer:
Saint Michael Parish

Reasoning Paths:
Step 1: Rihanna: Live in Concert Tour -> music.concert_tour.artist -> Rihanna
Step 2: Rihanna -> music.track_contribution.role -> Vocals

Output: This reasoning path indicates that:
1. The artist of Rihanna: Live in Concert Tour is Rihanna
2. Rihanna is a vocal artist

Even through there are no factual erros and the reasoning path is coherent, the reasoning path does not lead to the correct answer. The question asks for the birth place of the main artist, not the role of the artist. Therefore, the answer is "NO".

Input:
Question:
`<Question>`

Answer:
`<Answer>`

Reasoning path:
`<Reasoning Path>`

Output:

Figure 11: The few-shot CoT prompt used for discriminative evaluation.

Few-shot CoT prompt for Generative Evaluation

1. <step1>
2. <step2>
...
So the answer is (<answer>).
Make sure that the answer uses the above format and answers the question step by step.

Q: when Lou Seal is the mascot for the team that last won the World Series?
A: Let's work this out in a step by step way to be sure we have the right answer.
1. Lou Seal is the mascot for the San Francisco Giants.
2. The San Francisco Giants are associated with the sports championship event, the 2014 World Series.
So the answer is (2014 World Series).

Q: What nation has an army or more than 713480 people and borders the country of Bolivia?
A: Let's work this out in a step by step way to be sure we have the right answer.
1. Bolivia is a landlocked country located in South America.
2. Bolivia shares its borders with several countries, including Argentina, Brazil, Chile, Paraguay, and Peru.
So the answer is (Brazil).

Q: What movie was displayed at the 2012 Refugee Film Festival and had Angelia Jolie directing it?
A: Let's work this out in a step by step way to be sure we have the right answer.
1. Angelia Jolie whose first major film as a director which named "In the Land of Blood and Honey".
2. "In the Land of Blood and Honey" was shown at the 2012 Refugee Film Festival.
So the answer is (In the Land of Blood and Honey).

Q: How many Mary Mary sisters?
A: Let's work this out in a step by step way to be sure we have the right answer.
1. Mary Mary is a group which has a member named Tina Campbell
2. Mary Mary is a group which has a member named Erica Campbell
So the answer is (Erica Campbell, Tina Campbell).

Q: Which languages are used in the location that the breed Egyptian Mau started in?
A: Let's work this out in a step by step way to be sure we have the right answer.
1. The Egyptian Mau is a breed of domestic cat that is believed to have originated in Egypt.
2. In Egypt, the primary language spoken is Arabic, besides Domari or Nobiin.
So the answer is (Arabic, Domari, Nobiin).

Q: `{Question}`
A: Let's work this out in a step by step way to be sure we have the right answer.

Figure 12: The few-shot CoT prompt used for generative evaluation.

Few-shot CoT - Plan prompt used for Generative Evaluation

Relation path is a sequence relation that describes each step of the reasoning process.
You first give a relation path as a HINT, then reason the answer step-by-step based on it.
HINT:
1. <step1>
2. <step2>
...
So the answer is (<answer>).
Make sure that the answer uses the above format and answers the question step by step.

Q: when Lou Seal is the mascot for the team that last won the World Series?
A: Let's work this out in a step by step way to be sure we have the right answer.
HINT: sports.sports_team.team_mascot -> sports.sports_team.championships
1. Lou Seal is the mascot for the San Francisco Giants.
2. The San Francisco Giants are associated with the sports championship event, the 2014 World Series.
So the answer is (2014 World Series).

Q: What nation has an army or more than 713480 people and borders the country of Bolivia?
A: Let's work this out in a step by step way to be sure we have the right answer.
HINT: geography.river.basin_countries -> location.location.partially_contains
1. Bolivia is a landlocked country located in South America.
2. Bolivia shares its borders with several countries, including Argentina, Brazil, Chile, Paraguay, and Peru.
So the answer is (Brazil).

Q: What movie was displayed at the 2012 Refugee Film Festival and had Angelia Jolie directing it?
A: Let's work this out in a step by step way to be sure we have the right answer.
HINT: film.director.film -> film.film_regional_release_date.film_regional_debut_venue
1. Angelia Jolie whose first major film as a director which named "In the Land of Blood and Honey".
2. "In the Land of Blood and Honey" was shown at the 2012 Refugee Film Festival.
So the answer is (In the Land of Blood and Honey).

Q: How many Mary Mary sisters?
A: Let's work this out in a step by step way to be sure we have the right answer.
HINT: music.group_membership.member -> music.group_membership.member
1. Mary Mary is a group which has a member named Tina Campbell
2. Mary Mary is a group which has a member named Erica Campbell
So the answer is (Erica Campbell, Tina Campbell).

Q: Which languages are used in the location that the breed Egyptian Mau started in?
A: Let's work this out in a step by step way to be sure we have the right answer.
HINT: biology.breed_origin.breeds_originating_here -> location.country.languages_spoken
1. The Egyptian Mau is a breed of domestic cat that is believed to have originated in Egypt.
2. In Egypt, the primary language spoken is Arabic, besides Domari or Nobiin.
So the answer is (Arabic, Domari, Nobiin).

Q: `{Question}`
A: Let's work this out in a step by step way to be sure we have the right answer.

Figure 13: The few-shot CoT - Plan prompt used for generative evaluation.