DeCoT: Debiasing Chain-of-Thought for Knowledge-Intensive Tasks in Large Language Models via Causal Intervention

Anonymous ACL submission

Abstract

Large language models (LLMs) often require task-relevant knowledge to augment their internal knowledge through prompts. However, simply injecting external knowledge into prompts does not guarantee that LLMs can identify and use relevant information in the prompts to conduct chain-of-thought reasoning, especially when the LLM's internal knowledge is derived from the biased information on the pretraining data. In this paper, we propose a novel causal view to formally explain the internal knowledge bias of LLMs via a Structural Causal Model (SCM). We review the chain-of-thought (CoT) prompting from a causal perspective, and discover that the biased information from pretrained models can impair LLMs' reasoning abilities. When the CoT reasoning paths are misled by irrelevant information from prompts and are logically incorrect, simply editing factual information is insufficient to reach the correct answer. To estimate the confounding effect on CoT reasoning in LLMs, we use external knowledge as an instrumental variable. We further introduce CoT as a mediator to conduct front-door adjustment and generate logically correct CoTs where the spurious correlation between LLMs' pretrained knowledge and task queries is reduced. With extensive experiments, we validate that our approach enables more accurate CoT reasoning and enhances LLM generation on knowledge-intensive tasks.

1 Introduction

For knowledge-intensive tasks (Petroni et al., 2021; Hu et al., 2023; Sun et al., 2023b), specific knowledge is required in order to obtain an accurate response, which can be out of the distribution of LLMs' internal knowledge (Yao et al., 2023; Yuan et al., 2023c). Since frequently fine-tuning LLMs can be highly expensive and inefficient (Zhai et al., 2023), the LLM's internal knowledge can also be outdated and cause knowledge bias problems in LLMs (Zhang et al., 2023b; Wu et al., 2023; Zhang et al., 2023c). In order to efficiently incorporate external knowledge, prompt engineering methods try to retrieve task-relevant language evidence into prompts (Liu et al., 2023; He et al., 2022; Zhu et al., 2023b; Shao et al., 2023; Trivedi et al., 2022a). Additionally, external knowledge bases can also directly augment and edit the knowledge-injected prompts (Wen et al., 2023; Sun et al., 2023a; Baek et al., 2023; Zhao et al., 2023b). However, simply injecting external knowledge in prompts does not guarantee that LLMs can identify and use relevant information in the prompts (Shi et al., 2023a; Weston and Sukhbaatar, 2023), especially when the LLM acquires biased information on the pretraining data (Zhang et al., 2023b), such that the LLM may use the irrelevant information from

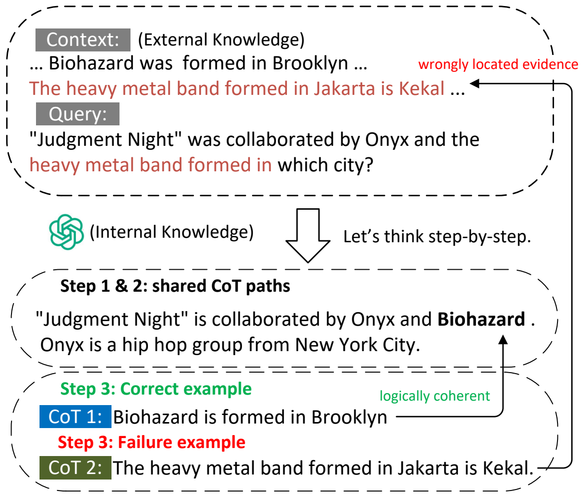


Figure 1: The LLM internal knowledge bias can trigger the usage of irrelevant information in the prompts, generate incoherent reasoning chains, and impair model's logical reasoning ability. This example is derived from the experiments by GPT3.5 on HotpotQA. Please note that 'The heavy metal band formed in Jakarta is Kekal' refers to a heavy meta band which is different from Biohazard. However, GPT3.5 incorrectly assume that 'The heavy metal band' refers to Kekal, and provides incorrect information in the step 3 of chain-of-thought.

prompts. The knowledge bias in LLMs can further cause knowledge conflict or misunderstanding between the external knowledge and the model's internal knowledge (Mallen et al., 2023; Wang et al., 2023g,a). In such cases, LLMs may generate incorrect and unexpected responses (Li et al., 2023c; Xie et al., 2023).

When the LLM relies on chain-of-thought (CoT) reasoning for complex tasks, the biased information from pretrained models further impairs LLMs' reasoning abilities. To eliminate the factual errors in the generated CoT paths, many works propose to verify and post-edit the generated reasoning paths before prompt again (Zhao et al., 2023b; Peng et al., 2023; Wang et al., 2023c). However, the logical reasoning errors can not be easily detected or corrected, as the effectiveness of factual verification and post-editing reasoning chains can be limited to simply injecting more knowledge. For example in Figure 1, given the query (*e.g.*, ""Judgment Night" was collaborated by Onyx and the heavy metal band formed in which city?"), the LLM may generate logically incorrect CoT (*e.g.*, CoT 1), in which the last chain is deviated from the reasoning paths (*e.g.*, instead of the origin of "Biohazard", some arbitrary band mentioned). Such logical incoherence can be caused by the spurious correlation between the query (*e.g.*, the concept "the heavy mental band formed in") and the LLM's internal knowledge understanding. Thus, even the LLM is prompted with the gold-truth context, the spurious correlation can lead the LLM to find some arbitrary evidence regardless of its logical connection to the previous chain, as long as it contains the exact phrase. In such cases, factual verification methods are not capable of detecting logical reasoning errors, and the answer can still be incorrect even with the facts verified as correct.

In this work, we propose a novel causal view via a Structural Causal Model (SCM) (Pearl et al., 2016) to formally explain the internal knowledge bias of LLMs. To measure spurious correlation, we propose to use external knowledge as an instrumental variable (Morgan and Winship, 2015) to estimate the Average Causal Effect (ACE) of CoT reasoning paths in LLMs through causal intervention. Based on the measurement of ACE, we can further introduce a CoT sampling method to find the best CoT as a mediator and conduct front-door adjustment (Pearl, 2009). In this approach, the spurious correlation between LLMs' internal knowledge and task queries can be reduced, which ensures correct CoT reasoning and LLM generation. We summarize our contributions as follows:

- We discover that the bias from pretrained LLMs can trigger the usage of irrelevant information in the prompts, generate incoherent reasoning chains, and impair model's logical reasoning ability.
- To formally understand the bias affecting CoT reasoning abilities, we propose a novel causal view introducing the external knowledge in prompts as an instrumental variable. This causal view uncovers the spurious correlation between queries and LLMs' internal knowledge understanding.
- To alleviate the bias and ensure correct CoT reasoning, we estimate the average causal effect (ACE) between the CoT and the answer, and further propose a CoT sampling method to conduct the front-door adjustment.
- We conduct multiple experiments on various knowledge-intensive tasks as well as numbers of LLM backbone models, which demonstrates the effectiveness of our method.

2 Related Work

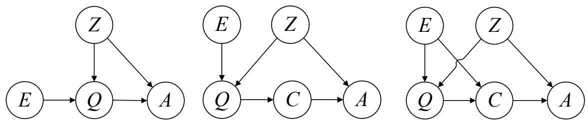
LLMs in Knowledge-intensive Tasks. In knowledge-intensive tasks (Petroni et al., 2021; Hu et al., 2023; Yang et al., 2018; Welbl et al., 2018) the LLM is asked to provide responses based on multiple pieces of evidence located in the context or from LLMs' intrinsic knowledge. Retrieval-augmented prompting methods focus on how to identify accurate and comprehensive evidence from support documents (Hoshi et al., 2023; Qian et al., 2023), in-context examples (Press et al., 2022; Khattab et al., 2022), knowledge bases (Trivedi et al., 2022a; Xu et al., 2023; Wang et al., 2023c; Feng et al., 2023; Zhu et al., 2023a), knowledge graphs (Wen et al., 2023; Sun et al., 2023a; Salnikov et al., 2023; Zhang et al., 2023a) and human feedback (Zhang et al., 2023b). However, extensive knowledge-injected prompts can introduce irrelevant information to distract LLMs (Shi et al., 2023a; Wang et al., 2023h) and cause LLMs' unpredictable behaviours (Li et al., 2023c; Chen et al., 2023b). In this case, the issue of spurious correlations between the context and the LLM can become more significant. Based on the motivation, instead of focusing on how to identify the best knowledge evidence, we investigate how to find logically correct CoT reasoning paths.

Chain-of-thought Prompting. Chain-of-thought prompting has shown great potential in explaining LLMs' thinking process (Yuan et al., 2023a; Li and Du, 2023) and answering multi-hop questions (Wang et al., 2023d; Ma et al., 2022) in several complex reasoning tasks (Fu et al., 2023). However, further works also mention issues of faithfulness and self-consistency in LLMs (Lanham et al., 2023; Turpin et al., 2023). In order to improve the faithfulness of intermediate chains, several works propose to verify and edit (He et al., 2022; Wang et al., 2023e; Zhao et al., 2023b) the factual errors in unfaithful chains, supervised by an external knowledge base. (Radhakrishnan et al., 2023; Zhu et al., 2023a) propose to decompose complex questions and answer them individually. The self-consistency principle is also considered in chain-of-thought distillation (Wang et al., 2023f). While factual verification and post-editing methods can fix factual errors and alleviate inconsistency problems in reasoning paths, how to detect logically incorrect chain-of-thoughts is less explored. We argue that such logically incorrect reasoning paths can lead the LLM astray from the right direction of finding the answer, even with the chains factually correct. Similar to previous works, our method also requires the LLM to first generate some candidate chain-of-thought reasoning paths, which makes our method using similar numbers of API calls (or inference times) per sample.

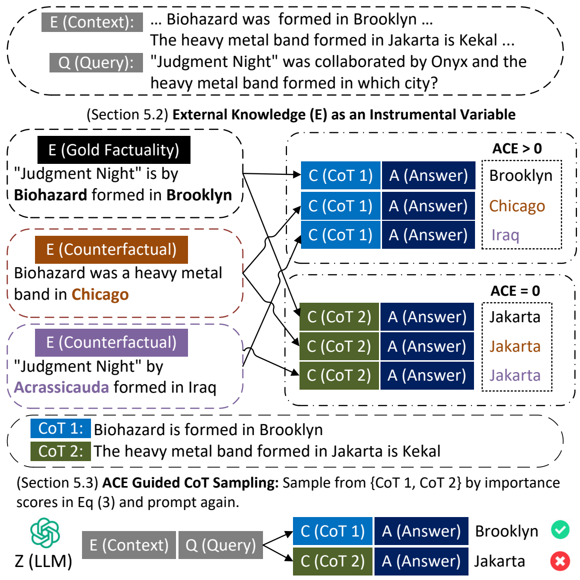
Causal Intervention in Language Models. Previous causal intervention methods in language models are focusing on entity-level spurious correlation (Wang et al., 2023b; Zeng et al., 2020; Yang et al., 2023) or sentence-level selection bias (McMilin, 2022). Since LLMs are black-box models (Gat et al., 2023; Cheng et al., 2023), direct methods of causal-aware model reparameterization methods are limited in usage. With causal explanability extracted from the models (Wu et al., 2021; Zhao et al., 2023a), further studies introduce human-in-the-loop debiasing methods (Zhang et al., 2023b; Wu et al., 2022, 2021). Human feedback can serve as the unbiased intermediate verification source. However, human effort is normally more expensive and involving humans in the loop may reduce the efficiency of the system. Instead, our method uses counterfactual context to automatically measure the causal effect. Without additional human effort or data resources for finetuning, our method can automatically find better CoT and improve the performance of LLMs.

3 A Causal View

To understand the causal relationships in knowledge-intensive tasks, we introduce a Structural Causal Model (SCM) (Pearl et al., 2000) and identify the internal knowledge understanding of the LLMs (Z) as the confounder. In Figure 2a and Figure 2b, we formulate two types of conventional knowledge injection methods as two SCMs respectively. With the SCMs, we explain the effectiveness of conventional knowledge injection methods as well as their limitations. We further present the SCM of our method, debiasing chain-of-thought (**DeCoT**), in Figure 2c. The formulation of our method **DeCoT** is illustrated in Figure 2d and explained in Section 5.



(a) SCM for External Knowledge (b) SCM for Chain-of-thought (c) SCM for DeCoT



(d) The illustration of our approach (detailed in Section 5).

Figure 2: Structural causal graphs for (a) injection of external knowledge (*i.e.,* context), (b) chain-of-thought prompting and (c) our approach using external knowledge as an instrumental variable (detailed in Section 5.2). Our proposed debiasing chain-of-thought method DeCoT is illustrated in (d).

3.1 SCM for External Knowledge

In Figure 2a, the causal path $E \rightarrow Q \leftarrow Z$ represents the knowledge injection process, in which E denotes the injected external knowledge, Q denotes the queries in the inference stage and Z denotes the LLM's internal knowledge. Ideally, since in this causal path, the query (Q) is the collider influenced by the other two variables, the external knowledge (E) and LLM's internal knowledge (Z), the spurious correlation can be alleviated as long as E and Z are causally irrelevant (Pearl et al., 2000). However, most conventional knowledge injection techniques (Baek et al., 2023; Li et al., 2023b) simply incorporate the external knowledge as the context which is prefixed to the input prompt. Thus, the causal influence of the external knowledge on the query is also determined by the LLM, which makes E and Z unable to be regarded as independent variables and the spurious correlation between Q and Z remains.

3.2 SCM for Chain-of-thought

Chain-of-thought (as in Figure 2b) is introduced (Li et al., 2023a; Wei et al., 2022; Fu et al., 2023) to make the LLM explain and follow the reasoning path before giving the final answer. The causal path $Q \rightarrow C \rightarrow A$ shows that the chain-of-thought reasoning path (C) can serve as the mediator between the query (Q) and the answer (A). However, since the chain-of-thought reasoning path is also prompted from the LLM (Wei et al., 2022; Fu et al., 2023), it can also be causally dependent to LLM's internal knowledge, and thus forms the spurious correlation between C and Z. Notably, knowledge editing methods (Zhao et al., 2023b; Li et al., 2023a; Peng et al., 2023) can correct the factual errors in the context and the reasoning paths, while the reasoning logic remains incorrect.

4 Preliminaries

4.1 Task Formulation

For knowledge-intensive question-answering tasks, the model is prompted with a query $Q = [q_1, q_2, \dots, q_n]$ and a passage of context $E = [e_1, e_2, \dots, e_l]$ which contains the supportive information to the query. Given the query Q and the context E, the model θ is prompted to recurrently generate the response Y by sampling from the conditional probability distribution,

$$y_t \sim p_\theta \left(y|E, Q, y_{<t}\right).$$

As illustrated and explained in Figure 2a of Section 3.1, the model directly generates the answer $A = [a_1, a_2, \dots, a_m]$ without providing the intermediate reasoning process (*i.e.*, $A = Y$).

Chain-of-thought Prompting. Following (Wei et al., 2022), we add the additional instruction to ask the model to generate its reasoning paths C by explaining step-by-step, before generating the final answer A (*i.e.*, $Y = [C, A]$). By sampling N different chain-of-thought reasoning paths $C = [C_1, C_2, \dots, C_N]$ conditioned on the query Q and the context E, we can further condition the generation process of the final answer A,

$$C_i \sim p_\theta \left(C|E, Q\right), \quad (1)$$

$$A_{i,r} \sim p_\theta \left(A|E, Q, C_i\right). \quad (2)$$

In Equation 1, since conventional chain-of-thought reasoning paths C are also generated from the LLM in which the pretrained internal knowledge Z can also confound on the chain-of-thought C generation process. As explained in Section 3.2, the confounding effect can not only affect the factual accuracy of the generated chain-of-thought, but also lead to incorrect reasoning logic. Thus knowledge editing and verification methods (Zhao et al., 2023b; Li et al., 2023a; Peng et al., 2023) which solve the former problem, are limited in correcting the reasoning logic of the chain-of-thought.

5 DeCoT: Debiasing Chain-of-thought

5.1 SCM for DeCoT

In Figure 2c, we propose to use the chain-of-thought reasoning path (C) which serves as the mediator between the query (Q) and the answer (A), to enable front-door adjustment. Based on the front-door criterion (Pearl et al., 2000), the mediator (C) should be causally independent to the confounder (Z). However, in practice, the chain-of-thought reasoning paths are also prompted from LLMs, which suggests potential spurious correlations between the chain-of-thought reasoning path (C) and LLM's internal knowledge (Z). Thus, to track the bias from the unobserved confounder Z, we introduce the external knowledge as the instrumental variable (IV) (Kawakami et al., 2023; Kilbertus et al., 2020; Yuan et al., 2023b). By changing the instrumental variable E (*i.e.*, the external knowledge), we can estimate the true causal relationship between C and A (Yuan et al., 2023b). For example, in Figure 2d, two pieces of counterfactual external knowledge (*e.g.* *"Biohazard formed in Chicago"* and *"Judgement Night was by Acrassicauda formed in Iraq"*) are introduced in the same example of Figure 1 and the average causal effect (ACE) can be calculated by the average difference of the former response distribution. Due to the spurious correlation in the third chain of thoughts of "CoT 1", the response generated from "CoT 1" remains unchanged (*i.e.*, ACE $= 0$), regardless of the counterfactual evidence. While for the correct reasoning path "CoT 2", the generated response changes corresponding to the different counterfactual evidence (*i.e.*, ACE > 0).

5.2 External Knowledge as an Instrumental Variable

We model the external knowledge as an instrumental variable E, such that we can change the instrumental variable E and measure the ACE metric, to understand the causal relationship between the CoT C and the answer A (Yuan et al., 2023b). Due to the limitation of directly controlling the generation process of chain-of-thought reasoning paths, we perform the causal treatment by including counterfactual knowledge through the instrumental variable E to understand the causal relationship between the chain-of-thought C and the answer A (Kawakami et al., 2023; Yuan et al., 2023b). Specifically, we query the LLM to extract T factual entities $V = [v_1, v_2, \ldots, v_T]$ which correspond to T counterfactual context $E_1^*, E_2^*, \ldots, E_T^*$ (prompt design explained in Appendix B). In each sample

$$E_j^* = [e_1, e_2, \ldots, v_j, \ldots, e_l] \,,$$

the corresponding factual entity v_j is to be replaced by counterfactual entities. Then, the LLM is further prompted to propose P counterfactual entities $V_j^* = [v_{j,1}^*, v_{j,2}^*, \ldots, v_{j,P}^*]$ to each extracted entity $v_j \in V$ (prompt design explained in Appendix B). P counterfactual context samples

$$E_{j,k}^*(v_{j,k}^*) = [e_1, e_2, \ldots, v_{j,k}^*, \ldots, e_l], k \leq P \quad (3)$$

are constructed, by replacing the corresponding factual entity v_j in each sample E_j^*. In this approach, we can estimate the average causal effect (ACE) corresponding to each chain-of-thought reasoning path C_i,

$$\begin{aligned} \text{ACE}(C_i, v_j) &= \mathbb{E}\left(A|do(E), Q, C_i\right) - \quad (4) \\ &\quad \mathbb{E}\left(A|do(E_j^*), Q, C_i\right) \\ &= \mathbb{E}_{v_{j,k}^* \in V_j^*}[p_\theta\left(A|E, Q, C_i\right) - \\ &\quad p_\theta\left(A|E_{j,k}^*(v_{j,k}^*), Q, C_i\right)], \end{aligned}$$

in which the average causal effect measures the decreased confidence of the answer with counterfactual context as the evidence. We observe that the average causal effect of different factual entities can be various to the context, queries and chain-of-thoughts, which we further conduct analysis experiments in Section 6.4. In order to consider the overall causal effect of the external context on each chain-of-thought reasoning path, we propose to measure the average causal effect of all the intervened entities,

$$\text{ACE}(C_i) = \mathbb{E}_{v_j \in V} \, \text{ACE}(C_i, v_j), \quad (5)$$

in which we assume the intervened entities v_j are sampled from a uniform distribution of the external context E.

5.3 Average Causal Effect Guided Chain-of-thought Sampling

With the measured ACE scores, we develop an efficient sampling approach to obtain high-quality CoTs with more coherent reasoning chains. Since LLMs are black-box models (Gat et al., 2023; Cheng et al., 2023), direct causal intervention methods on the parameterization of the input query Q and the context E are limited. In addition, inserting additional deconfounding layers (Zhang et al., 2020; Wu et al., 2022) to finetune the LLM requires considerable computation and data resources, which makes these methods less efficient. We propose to use the sampled chain-of-thought reasoning paths C as the mediator variable to conduct the front-door adjustment.

Based on the measured average causal effect (ACE), we construct the importance scores in terms of how the final answer A reacts to the different chain-of-thought reasoning paths C that intervened by the context E,

$$C^* \sim \text{softmax}\left[p_\theta\left(C_i|E, Q\right) \cdot \text{ACE}(C_i)\right], \quad (6)$$

and the front-door adjustment can be realized by introducing the mediator C^* sampled with the largest average causal effect in the reasoning path,

$$\begin{aligned} A^* &\sim P\left(A|E, Q, do(C)\right) \quad (7) \\ &\propto p_\theta\left(A|E, Q, C^*\right). \end{aligned}$$

The causal effect on the sampled answer A^* is mediated by the sampled CoT reasoning path C^*, whose mediator-outcome confounding effect is controlled and alleviated.

We summarize our algorithm DeCoT in Algorithm 1 (Appendix D).

Model	Decoding	HotpotQA		MuSiQue		SciQ		WikiHop		Average	
		EM↑	F1↑	EM↑	F1↑	EM↑	F1↑	EM↑	F1↑	EM↑	F1↑
Flan-T5	CoT w/o ctx	7.41	17.99	2.57	8.50	11.09	17.80	4.12	6.88	6.30	12.79
	CoT	9.48	23.70	19.53	27.61	51.75	63.79	15.02	21.79	23.95	34.22
	CAD	9.65	24.77	**20.56**	28.57	59.69	69.94	17.28	24.31	26.80	36.90
	DeCoT	**11.72**	**28.70**	**20.56**	**30.54**	**63.55**	**75.64**	**22.34**	**28.41**	**29.54**	**40.82**
LlaMA-2	CoT w/o ctx	1.67	3.04	0.56	1.44	4.08	5.45	1.19	1.64	1.88	2.89
	CoT	8.86	26.79	20.22	27.46	30.64	39.59	23.10	28.23	20.71	30.52
	CAD	**10.53**	30.98	21.62	28.10	33.93	41.35	23.50	29.81	22.40	32.56
	DeCoT	10.03	**31.48**	**22.75**	**30.99**	**48.58**	**57.95**	**27.35**	**34.46**	**27.18**	**38.72**
GPT-3.5	CoT w/o ctx	5.60	30.97	2.09	7.96	29.82	43.18	11.62	19.31	12.28	25.36
	CoT	5.10	32.55	22.30	34.22	54.53	68.50	25.40	35.25	26.83	42.63
	CAD	5.43	35.24	24.14	36.81	57.74	70.73	28.37	38.45	28.92	45.31
	DeCoT	**10.21**	**40.19**	**31.28**	**44.14**	**64.61**	**78.10**	**31.89**	**43.45**	**34.50**	**51.47**

Table 1: The comparison results of DeCoT based on different backbone LLMs on four knowledge-intensive tasks. The best results for each backbone model and each dataset are highlighted in a **bold font**.

6 Experiments

6.1 Dataset and Evaluation

Knowledge-intensive tasks commonly require each question to be paired with a paragraph of context as support evidence (Li et al., 2023a; Zhu et al., 2023b; Su et al., 2023; Jang et al., 2023). We follow the evaluation protocols in (Yang et al., 2018) and conduct our experiments on datasets as follows:

HotpotQA (Yang et al., 2018) contains questions which require multi-step reasoning over multiple support contexts. For each question, support documents are provided in the dataset, which are used as the context in our experiments.

MuSiQue (Trivedi et al., 2022b) is another multi-step question answering dataset. Similar to previous work (Ramesh et al., 2023), we conduct our experiment on the challenging part of the dataset, in which questions are annotated as ≥ 4 *hops*.

WikiHop (Welbl et al., 2018) is a multi-choice multi-hop reasoning dataset. We use the queries in the dataset as the questions (Tu et al., 2019) in our setting. For baselines and our method, the models are prompted to generate the answers free-form instead of retrieving from the candidate list.

SciQ (Welbl et al., 2017) is a domain-specific question answering task which contains only scientific questions. We evaluate baselines and our method on test samples with support evidence.

For datasets which lack of test labels, we follow the same evaluation protocol as (Press et al., 2022; Shao et al., 2023; Chen et al., 2023a) and use the development sets as our test set. We use the Exact Match (EM) and F1 proposed in (Yang et al., 2018) as our evaluation metrics.

6.2 Baseline and Backbone Model

Following (Shi et al., 2023b; Su et al., 2023; Trivedi et al., 2022a), we have applied our method to different pretrained LLMs: Flan-T5-XXL (Chung et al., 2022) which has 11B model parameters, LlaMA-2-7B (Touvron et al., 2023) and GPT-3.5 Turbo (Brown et al., 2020). For LlaMA-2-7B model, we choose the finetuned versions from human feedback (Christiano et al., 2017), which can generally yield more stable chain-of-thought reasoning paths.

For baselines, we compare our method with a conventional chain-of-thought prompting method (CoT) (Wei et al., 2022) and context-aware contrastive decoding method CAD (Shi et al., 2023b). We also include the baseline which devises conventional chain-of-thought prompting methods without context (CoT w/o ctx) (Wei et al., 2022), in order to investigate the effect of context in different datasets. The implementation details are described in Appendix A.

6.3 Main Results

Table 1 presents evaluation results on the four datasets with three LLM backbone models.

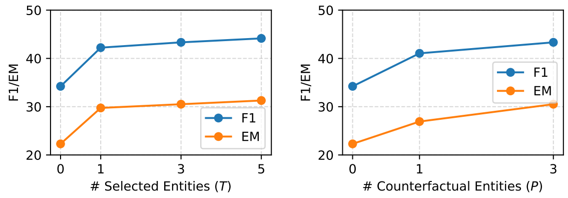
Comparison with Baselines. As we expected, for all LLMs the performance are significantly lower when the context of supporting evidence is absent. Because of the poor performance of the direct prompting method, the context-aware contrastive decoding (CAD) baseline is able to use its answer distribution as the negative penalty on the positive distribution which is obtained by prompting with both the query and context. However, with the gold-truth context provided, the negative answer distribution which is unsupported by either internal or external knowledge can be more random, which limits the effectiveness of contrastive decoding methods. On the other hand, our method DeCoT achieves generally higher improvements on regular CoT comparing with CAD by detecting logically incorrect CoTs and penalizing on them. Instead of simply contrasting the distributions of positive and negative answers, we use counterfactual context to examine the answer distribution changes, which provides a more fine-grained measurement of the causal effect on LLMs' internal knowledge bias. The consistent performance improvements suggest DeCoT can more accurately detect logically incorrect CoTs by the measurement, and perform more targeted causal intervention.

Logical Reasoning Performance Understanding. We also observe that DeCoT gains relatively better F1 improvements on the SciQ dataset, which reach to 18.58%, 46.37% and 14.01% for Flan-T5-xxl, LlaMA-2-7B and GPT-3.5 models respectively. This is because for scientific questions, accurate logical reasoning paths are more strictly required, which makes the correctness of the generated CoTs more crucial. Thus, DeCoT's better performance on the SciQ dataset suggests DeCoT is more effective in debiasing LLMs' logical reasoning ability in CoT.

6.4 Impact of the Selected Entities

We evaluate the impact of the number of the selected entities on the MuSiQue dataset based on the backbone model GPT-3.5. Since annotations in MuSiQue guarantee the minimal number of chains of thought is 4 hops (Ramesh et al., 2023), more factual evidence is required to support the final answer, which makes the impact of the selected entities higher in this case.

To illustrate the trend, we only conduct experiments on number of the selected entities T with these representative values considering the expensive GPT API costs. In Figure 3a, we show the F1 and EM performance w.r.t. the different number of selected entities $T = 0, 1, 3, 5$. We include the result of $T = 0$ which indicate the regular CoT prompting method. With a larger T, it has a higher probability for DeCoT to find more important entities to perform causal intervention on. However,



(a) Impact of T.　(b) Impact of P.

Figure 3: Sensitivity studies on impact of the number of the selected entities and alternative entities. We also include the $T = 0$ and $P = 0$ data points indicating the performance of regular CoT prompting methods. The experiments are conducted on the GPT-3.5 backbone model on the MuSiQue dataset.

including more causal intervention experiments requires more counterfactual prompting, which is at the expense of more API calls or more inference time. We observe that for context annotated as the gold truth, we can accurately find good factual entities by selecting the most popular entities.

6.5 Impact of the Alternative Entities

In Figure 3b, we also show the performance w.r.t. the different number of alternative entities $P = 0, 1, 3$. For the same reason of computational cost, we only choose a small number of alternative entities. Since the alternative entities are counterfactual, normally the negative causal effect should be similar among the alternative entities. However, practically the alternative entities may further introduce unexpected causal relationship to the context. Thus, by calculating the average causal effect of multiple alternative entities, such unexpected causal effect can be averaged out.

6.6 Case Study

We present case study results to understand the impact of gold truth to the quality of CoTs. Besides, we further examine the performance of CoT w/ Post-Editing, in the setting when the gold truth is not available. We further illustrate that our approach can overcome the limitations of these different CoT variants. In Table 2, we show the CoT reasoning paths prompted from DeCoT and following methods:

- **CoTs w/ Gold Truth**: regular CoT prompting with gold-truth context (Wei et al., 2022) (*i.e.,* CoT in Section 6.2.).
- **CoTs w/o Gold Truth**: regular CoT prompting with only the task query (Wei et al., 2022) (*i.e.,* CoT w/o ctx in Section 6.2.).

Dataset	Question	CoTs w/ Gold Truth	CoTs w/o Gold Truth	CoT w/ Post-Editing	DeCoT
HotpotQA	The Pluralist school is said to have included what creator of the theory of atomism?	The Pluralist school was a school of pre-Socratic philosophers. The Pluralist school included Leucippus and Democritus. Therefore, the final answer is [Democritus ✗].	One of the key pluralist figures was Greek philosopher Aristotle. Aristotle is known for his theory that nature contains no vacuums. Therefore, the final answer is [Aristotle ✗].	One of the key pluralist figures was Greek philosopher Aristotle. Aristotle is known for his theory of atomism. Therefore, the final answer is [Aristotle ✗].	The Pluralist school is a philosophical movement that emerged in Ancient Greece. It is said to have included Leucippus, the creator of the theory of atomism. Therefore, the final answer is [Leucippus ✓].
MuSiQue	What county contains the city with a radio station that broadcasts to the capital city of the state where the Peace center is located?	The state where the Peace center is located is Greenville, South Carolina. The county containing the city with a radio station broadcasting to Greenville is likely Greenville County. Therefore, the final answer is [Greenville County ✗].	The Peace Center is located in Greenville, South Carolina. The capital city of South Carolina is Columbia. There are two radio stations in Greenville that broadcast to Columbia - WJMZ-FM and WMYI. WMYI, which is located in Greenville, broadcasts to the capital city of Columbia. Therefore, the final answer is [Greenville County ✗].	The Peace Center is located in Greenville, South Carolina. The capital city of South Carolina is Columbia. There are two radio stations in Greenville that broadcast to Columbia - WJMZ-FM and WMYI. WMYI is located in Hendersonville, North Carolina. Therefore, the final answer is [Hendersonville ✗].	The Peace Center is located in Greenville, South Carolina. The radio station WWNQ is licensed to Forest Acres and broadcasts to Columbia, which is the capital city of South Carolina. Forest Acres is a city in Richland County, South Carolina. Therefore, the final answer is [Richland County ✓].
SciQ	By exploding what the space probes get separated from their launchers?	The context is discussing how space probes use exploding bolts to separate from their launchers. The question is asking what the space probes are separated from during this process. Therefore, the final answer is [their launchers ✗].	When a space probe is launched, it is attached to a rocket or launcher. Once the rocket has reached a certain point in space, it explodes, releasing the space probe. Therefore, the final answer is [the rocket or launcher ✗].	When a space probe is launched, it is attached to a rocket or launcher. ~~Once the rocket has reached a certain point in space, it explodes, releasing the space probe.~~ Therefore, the final answer is [the explosion or detonation ✗].	The context states that space probes may be separated from their launchers by exploding bolts. Exploding bolts are bolts that explode or break apart in order to separate two objects. Therefore, the final answer is [bolts ✓].
WikiHop	What's the genre of Terra Brasilis?	"Genre" refers to a category of artistic work. Terra Brasilis is a famous album by Brazilian musician Milton Nascimento. Therefore, the final answer is [Brazilian music ✗].	"Terra Brasilis" is a song by Brazilian composer Caetano Veloso. The song is from his 1978 album "Muito (Dentro Da Estrela Azulada)". Therefore, the final answer is [Brazilian ✗].	"Terra Brasilis" is an album by Brazilian composer Antônio Carlos Jobim. ~~The song is from his 1978 album "Muito (Dentro Da Estrela Azulada)".~~ Therefore, the final answer is [Brazilian ✗].	The album Terra Brasilis was released in 1980 by Brazilian composer Antônio Carlos Jobim. Jobim is known for his contributions to bossa nova music. Therefore, the final answer is [bossa nova ✓].

Table 2: Examples of failure CoTs generated from regular CoT prompting, CoT prompting without gold-truth context and knowledge post-edited CoT prompting, as well as DeCoT sampled successful CoTs, from four datasets with the GPT3.5 model. In the examples, we highlight factual and logical errors with a red font, while the correct reasoning evidence is in a green font. The edited factuality is also highlighted with a blue font.

- **CoT w/ Post-Editing**: CoT with knowledge verification and post-edit (Zhao et al., 2023b).

For all the methods, the CoTs are prompted from the backbone GPT-3.5 model.

Failure Cases by CoTs with Post-Editing. We observe that CoTs generated without the gold truth are very likely to contain incorrect knowledge, which can further mislead the reasoning paths. For example, the directly generated CoTs of the question in the WikiHop dataset say "Terra Brasilis is a song by Caetano Veloso", which is factually incorrect (highlighted in a red font). Due to this incorrect assumption made from LLMs' hallucination, the following reasoning paths are misled to talk about irrelevant information (*e.g.*, "Caetano Veloso's album") and thus the answer is wrong even with factual edit (highlighted in a blue font).

Failure Cases by CoTs with Gold Truth. Compared to CoTs generated without the gold truth, we observe that the CoTs prompted with the gold-truth context can be factually more faithful. However, the logical reasoning of these CoTs can still be wrong. For example, the CoTs generated with the gold truth in the HotpotQA dataset are correctly locating "Leucippus and Democritus" as the two "Pluralist school" members (highlighted in a red font). However, instead of answering with the one who creates the school, the LLM mistakenly chooses the wrong answer "Democritus". We highlight the correct reasoning paths in a green font to show that the key point in answering this question is by identifying the "creator".

7 Conclusion

In this paper, we formally examine the LLM's internal knowledge bias and identify it as the confounder by a structural causal model (SCM). We discover the spurious correlation between the LLM and task queries, which can further impairs the LLM's CoT reasoning abilities. Then, we propose DeCoT, a debiasing chain-of-thought prompting method in knowledge-intensive tasks, which alleviates the spurious correlation and enables the LLM to find more accurate and logically sound responses. Extensive experimental results and case studies validate the effectiveness of our method DeCoT.

8 Limitations

Since DeCoT is an inference-stage causal intervention method, the improvement on LLMs' reasoning abilities is attributed to alleviating the bias, but can be limited to the upper bound of the LLM's capacity. To alleviate the causal effect of knowledge bias on LLMs' reasoning abilities, future works can incorporate unbiased causal learning methods in the model pretraining or instruction tuning stage, which may enable more robust CoT reasoning. It is also interesting to study the theoretical causal foundation of CoT prompting's mediator role in LLMs, which can be beneficial to better interpretability of black-box LLMs.

9 Ethics Statement

Our study on mitigating bias in Large Language Models (LLMs) recognizes the ethical implications of data-driven biases in AI, specifically addressing how these biases affect reasoning processes. We propose a novel approach to reduce bias impact, emphasizing the responsible and ethical advancement of AI technology. The datasets we used in our experiments are all publicly available. No personal information was gathered from our human participants, and they were not exposed to any harmful model outputs.

References

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136*.

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.

Mingda Chen, Xilun Chen, and Wen-tau Yih. 2023a. Efficient open domain multi-hop question answering with few-shot data synthesis. *arXiv preprint arXiv:2305.13691*.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2023b. Unveiling the siren's song: Towards reliable fact-conflicting hallucination detection. *arXiv preprint arXiv:2310.12086*.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Black-box prompt optimization: Aligning large language models without model training. *arXiv preprint arXiv:2311.04155*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Chao Feng, Xinyu Zhang, and Zichu Fei. 2023. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv:2309.03118*.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306*.

Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2023. Faithful explanations of black-box nlp models using llm-generated counterfactuals. *arXiv preprint arXiv:2310.00603*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii, and Jun Deguchi. 2023. Ralle: A framework for developing and evaluating retrieval-augmented large language models. *arXiv preprint arXiv:2308.10633*.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*.

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2023. Exploring the benefits of training expert language models over instruction tuning. *arXiv preprint arXiv:2302.03202*.

Yuta Kawakami, Manabu Kuroki, and Jin Tian. 2023. Instrumental variable estimation of average partial causal effects. In *International Conference on Machine Learning*, pages 16097–16130. PMLR.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Niki Kilbertus, Matt J Kusner, and Ricardo Silva. 2020. A class of algorithms for general instrumental variable models. *Advances in Neural Information Processing Systems*, 33:20108–20119.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Ruosen Li and Xinya Du. 2023. Leveraging structured information for explainable multi-hop question answering and reasoning. *arXiv preprint arXiv:2311.03734*.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023a. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

Zhifeng Li, Bowei Zou, Yifan Fan, and Yu Hong. 2023b. Ufo: Unified fact obtaining for commonsense question answering. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2023c. Unveiling the pitfalls of knowledge editing for large language models. *arXiv preprint arXiv:2310.02129*.

Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. 2023. Reta-llm: A retrieval-augmented large language model toolkit. *arXiv preprint arXiv:2306.05212*.

Kaixin Ma, Hao Cheng, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. 2022. Open-domain question answering via chain of reasoning over heterogeneous knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5360–5374, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Emily McMilin. 2022. Selection bias induced spurious correlations in large language models. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*.

Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Judea Pearl. 2009. Causal inference in statistics: An overview.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus. *arXiv preprint arXiv:2304.04358*.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*.

Gowtham Ramesh, Makesh Sreedhar, and Junjie Hu. 2023. Single sequence prediction over reasoning graphs for multi-hop qa. *arXiv preprint arXiv:2307.00335*.

Mikhail Salnikov, Hai Le, Prateek Rajput, Irina Nikishina, Pavel Braslavski, Valentin Malykh, and Alexander Panchenko. 2023. Large language models meet knowledge graphs to answer factoid questions. *arXiv preprint arXiv:2310.02166*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023b. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Xin Su, Tiep Le, Steven Bethard, and Phillip Howard. 2023. Semi-structured chain-of-thought: Integrating multiple sources of knowledge for improved language model reasoning. *arXiv preprint arXiv:2311.08505*.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023a. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. *arXiv preprint arXiv:2307.07697*.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023b. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022a. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022b. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. *arXiv preprint arXiv:1905.07374*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023b. A causal view of entity bias in (large) language models. *arXiv preprint arXiv:2305.14695*.

Jianing Wang, Qiushi Sun, Nuo Chen, Xiang Li, and Ming Gao. 2023c. Boosting language models reasoning with chain-of-knowledge prompting. *arXiv preprint arXiv:2306.06427*.

Jinyuan Wang, Junlong Li, and Hai Zhao. 2023d. Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning. *arXiv preprint arXiv:2310.13552*.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023e. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023f. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023g. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023h. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

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.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*.

Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 attention (is something you might need too). *ArXiv*, abs/2311.11829.

Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao, and Ani Nenkova. 2022. Context-aware information-theoretic causal de-biasing for interactive sequence labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3436–3448.

Junda Wu, Tong Yu, and Shuai Li. 2021. Deconfounded and explainable interactive vision-language retrieval of complex scenes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2103–2111.

Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. 2023. Eva-kellm: A new benchmark for evaluating knowledge editing of llms. *arXiv preprint arXiv:2308.09954*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-seng Chua. 2023. Search-in-the-chain: Towards the accurate, credible and traceable content generation for complex knowledge-intensive tasks. *arXiv preprint arXiv:2304.14732*.

Zhen Yang, Yongbin Liu, and Chunping Ouyang. 2023. Causal interventions-based few-shot named entity recognition. *arXiv preprint arXiv:2305.01914*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2023a. Back to the future: Towards explainable temporal reasoning with large language models. *arXiv preprint arXiv:2310.01074*.

Junkun Yuan, Xu Ma, Ruoxuan Xiong, Mingming Gong, Xiangyu Liu, Fei Wu, Lanfen Lin, and Kun Kuang. 2023b. Instrumental variable-driven domain generalization with unobserved confounders. *ACM Transactions on Knowledge Discovery from Data*.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023c. Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations. *arXiv preprint arXiv:2306.04618*.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, Online. Association for Computational Linguistics.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023a. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023b. Mitigating language model hallucination with interactive question-knowledge alignment. *arXiv preprint arXiv:2305.13669*.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023c. How do large language models capture the ever-changing world knowledge? a review of recent advances. *arXiv preprint arXiv:2310.07343*.

Ruochen Zhao, Shafiq Joty, Yongjie Wang, and Prathyusha Jwalapuram. 2023a. Towards causal concepts for explaining language models.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023b. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada. Association for Computational Linguistics.

Wang Zhu, Jesse Thomason, and Robin Jia. 2023a. Chain-of-questions training with latent answers for robust multistep question answering. *arXiv preprint arXiv:2305.14901*.

Yin Zhu, Zhiling Luo, and Gong Cheng. 2023b. Furthest reasoning with plan assessment: Stable reasoning path with retrieval-augmented large language models. *arXiv preprint arXiv:2309.12767*.

A Implementation Details

To obtain diversified chain-of-thought reasoning paths, we sample $N = 5$ chains-of-thought with the temperature set to 1.0 for all the backbone models. For **DeCoT**, we let the LLM extract the top $T = 5$ most frequently appearing entities in the context as to be intervened. The LLM will be further prompted to provide $P = 3$ alternative counterfactual entities to each of the extracted entities.

As for the open-sourced LLMs (*i.e.*, Flan-T5-XXL and LlaMA-2-7B), we use the official Hugging Face implementations. The experiments are conducted using 4 NVIDIA RTX A6000 GPUs with 48GBs. For GPT-3.5 Turbo, we use the OpenAI API to conduct the experiments.

To prompt the LLMs to generate more robust chain-of-thought results and also follow a unified answer format, we have included 3 few-shot in-context learning examples. The in-context learning examples are from a separate set of data which provide no extra knowledge to the evaluated tasks. In addition, we have also included 3 in-context learning examples for both the entity extraction and the alternative entity proposal prompts. Detailed designs of these in-context examples and prompts are explained in Appendix C.

B Prompt Design

B.1 Factual Entity Extraction

To extract the most relevant factual entities V in the context E (Section 5.2),

$$v_j \sim p_\theta\left(V|E, Instruct_{ent}\right), \quad (8)$$
$$E_j^* = \left[e_1, e_2, \ldots, v_j, \ldots, e_l\right],$$

in which $Instruct_{ent}$ is the explicit prompt instruction shown in following.

Context Example 1:

The Ritz-Carlton Jakarta is a hotel and skyscraper in Jakarta, Indonesia and 14th Tallest building in Jakarta. It is located in city center of Jakarta, near Mega Kuningan, adjacent to the sister JW Marriott Hotel. It is operated by The Ritz-Carlton Hotel Company. The complex has two towers that comprises a hotel and the Airlangga Apartment respectively. Nakuul Mehta, Kunal Jaisingh and Leenesh Mattoo respectively portray Shivaay, Omkara and Rudra, the three heirs of the Oberoi family.

Instruction Example 1:

Extract the top 5 most frequently appeared entities in the context and provide in the format of a list: [Ritz-Carlton, Jakarta, Indonesia, Airlangga Apartment, Nakuul Mehta]

Context Example 2:

Lisa Marie Simpson is a fictional character in the animated television series "The Simpsons". She is the middle child and most intelligent of the Simpson family. Voiced by Yeardley Smith, Lisa first appeared on television in "The Tracey Ullman Show" short "Good Night" on April 19, 1987. Cartoonist Matt Groening created and designed her while waiting to meet James L. Brooks. Groening had been invited to pitch a series of shorts based on his comic "Life in Hell", but instead decided to create a new set of characters. He named the elder Simpson daughter after his younger sister Lisa Groening.

Instruction Example 2:

Extract the top 5 most frequently appeared entities in the context and provide in the format of a list: [Lisa Marie Simpson, The Simpsons, Yeardley Smith, The Tracey Ullman Show, Lisa Groening]

B.2 Alternative Entity Proposal

To ask the LLM to propose P counterfactual entities $E_{j,1}^*, E_{j,2}^*, \ldots, E_{j,P}^*$ to the extracted entity v_j (Section 5.2),

$$v_{j,k}^* \sim p_\theta\left(V|v_j, Instruct_{alt}\right), \quad (9)$$
$$E_{j,k}^*(v_{j,k}^*) = \left[e_1, e_2, \ldots, v_{j,k}^*, \ldots, e_l\right],$$

in which $Instruct_{alt}$ is the explicit prompt instruction shown in following.

Example 1

Provide 3 most similar entities to "America", which are from the same entity type and have similar meanings. Provide the answers as a list: [Britain, France, Germany].

Example 2

Provide 3 most similar entities to "Florida", which are from the same entity type and have similar meanings. Provide the answers as a list: [New York, California, Arizona].

Example 3

Provide 3 most similar entities to "Beethoven", which are from the same entity type and have similar meanings. Provide the answers as a list: [Mozart, Mahler, Wagner].

C In-context Learning Examples

We design several CoT prompting examples to guide the LLM to generate in an expected format as following:

Example 1

Context: The Ritz-Carlton Jakarta is a hotel and skyscraper in Jakarta, Indonesia and 14th Tallest building in Jakarta.
Question: The Oberoi family is part of a hotel company that has a head office in what city?

Based on the context, answer the question step by step and provide the final answer in the end.
1. The Oberoi family is an Indian family known for their involvement in hotel management
2. The Oberoi Group is the hotel company founded by the family
Therefore, the final answer is [Delhi].

Example 2

Context: Lisa Marie Simpson is a fictional character in the animated television series "The Simpsons".
Question: Musician Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who?

Based on the context, answer the question step by step and provide the final answer in the end.
1. The character Milhouse from the TV show "The Simpsons" was named by Matt Groening
2. Groening named the character after President Richard Nixon's middle name
Therefore, the final answer is [Richard Nixon].

Example 3

Context: Moloch: or, This Gentile World is a semi-autobiographical novel written by Henry Miller in 1927-28, initially under the guise of a novel written by his wife, June.
Question: What nationality was James Henry Miller's wife?

Based on the context, answer the question step by step and provide the final answer in the end.
1. James Henry Miller (25 January 1915 – 22 October 1989) was an English folk singer and songwriter
2. His second wife was Margaret "Peggy" Seeger (born June 17, 1935), who is an American folksinger
Therefore, the final answer is [American].

D Pseudo Code for DeCoT

Algorithm 1: DeCoT

Input: Task query Q, Context E, the LLM with model parameters θ

Init: Sample chain-of- thought reasoning paths $C = [C_1, C_2, \ldots, C_N]$ by Eq. 1.
Extract T factual entities $V = [v_1, v_2, \ldots, v_T]$ from E by Eq. 8 ;
while $i < N$ **do**
    while $j < T$ **do**
        Propose P counterfactual entities $\{v^*_{j,1}, v^*_{j,2}, \ldots, v^*_{j,P}\}$ by Eq. 9 ;
        while $k < P$ **do**
            Construct counterfactual context $E^*_{j,k}(v^*_{j,k})$ by Eq. 3 ;
        end
        Estimate ACE$(C_i, v_j), i < N$ for each entity v_j by Eq. 4 ;
    end
    Estimate ACE(C_i) by Eq. 5 ;
end
Sample CoT by Eq. 6 ;
Sample the answer by Eq. 7 ;