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

(/pdf?id=k8znTbs8t8N)

Anonymous

16 Feb 2024 ACL ARR 2024 February Blind Submission Readers: February, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers, Paper2816 Authors Show Revisions (/revisions?id=k8znTbs8t8N)

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 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.

Paper Type: long

Research Area: Generation

Contribution Types: Model analysis & interpretability, NLP engineering experiment, Theory

Languages Studied: English

Revealed to Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan A. Rossi, Sungchul Kim, Anup Rao, Julian McAuley

15 Feb 2024 (modified: 16 Feb 2024) ACL ARR 2024 February Submission

Authors: *Junda Wu (/profile?id=~Junda_Wu1), Tong Yu (/profile?id=~Tong_Yu3), Xiang Chen (/profile?id=~Xiang_Chen9), Haoliang Wang (/profile?id=~Haoliang_Wang1), Ryan A. Rossi (/profile?id=~Ryan_A._Rossi2), Sungchul Kim (/profile?id=~Sungchul_Kim1), Anup Rao (/profile?id=~Anup_Rao1), Julian McAuley (/profile?id=~Julian_McAuley1)*

Previous URL: /forum?id=JGVgR5L0iDq (/forum?id=JGVgR5L0iDq)

Previous PDF: pdf (/attachment?id=V1CHKymxWA3&name=previous_PDF)

Response PDF: pdf (/attachment?id=V1CHKymxWA3&name=response_PDF)

Reassignment Request Action Editor: No, I want the same action editor from our previous submission and understand that a new action editor may be assigned if the previous one is unavailable

Reassignment Request Reviewers: No, I want the same set of reviewers from our previous submission and understand that new reviewers may be assigned if any of the previous ones are unavailable

Preprint: no

Preprint Status: We are considering releasing a non-anonymous preprint in the next two months (i.e., during the reviewing process).
Preferred Venue: ACL 2024
Consent To Share Data: yes
Consent To Review: yes
Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.
A1: yes
A1 Elaboration For Yes Or No: Section 8
A2: n/a
A3: yes
A3 Elaboration For Yes Or No: Abstract and Section 1
B: no
B1: n/a
B2: n/a
B3: n/a
B4: n/a
B5: n/a
B6: yes
B6 Elaboration For Yes Or No: Section 6
C: yes
C1: yes
C1 Elaboration For Yes Or No: Section 6
C2: yes
C2 Elaboration For Yes Or No: Section 6
C3: yes
C3 Elaboration For Yes Or No: Section 6
C4: yes
C4 Elaboration For Yes Or No: Section 6
D: no
D1: n/a
D2: n/a
D3: n/a
D4: n/a
D5: n/a
E: no
E1: n/a

Add **Author-Editors Confidential Comment** **Withdraw**

Reply Type: all Author: everybody **13 Replies**

Visible To: Paper2816 Authors, Paper2816 Reviewers, ... Hidden From: nobody

[–] **Meta Review of Paper2816 by Area Chair Ey32**

ACL ARR 2024 February Paper2816 Area Chair Ey32

07 Apr 2024, 08:50 ACL ARR 2024 February Paper2816 Meta Review Readers: Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Authors, Paper2816 Reviewers Submitted, Program Chairs Show Revisions (/revisions?id=lh1d64A9JX)

Paper Summary:
The paper presents an approach, DeCoT, to selecting the final reasoning path and answer from multiple chain-of-thought (CoT) generations produced by large language models (LLMs). Specifically, the method prompts LLMs to generate multi-hop reasoning paths, estimates the Average Causal Effect (ACE) of reasoning paths and answers by constructing counterfactual contexts, and ultimately samples the final path and answer. DeCoT can help LLMs find more accurate and logically sound responses in knowledge-intensive tasks.

Summary Of Strengths:
- The causal perspective adopted for analysis and the methodology proposed are novel and offer a fresh angle to approach CoT in LLMs.
- The method outperforms the vanilla CoT approaches across multiple datasets, demonstrating strong empirical results.

Summary Of Weaknesses:
- In section 6.4, CoT/CAD (w/ ReAct) can not work well, while the proposed DeCoT, a further extension of CoT, brings great improvement, which requires some explanation.
- Some writing issues mentioned by reviewers should be considered.

Overall Assessment: 4 = There are minor points that may be revised

Best Paper Ae: No

Information Regarding The New ACL Policy On Deanonymized Preprints: I confirm I have read the information above about changes to the anonymity policy.

Add **Author-Editors Confidential Comment**

[–] # Official Review of Paper2816 by Reviewer 5jLu

ACL ARR 2024 February Paper2816 Reviewer 5jLu

22 Mar 2024, 07:31 (modified: 02 Apr 2024, 21:50) ACL ARR 2024 February Paper2816 Official Review Readers: Program Chairs, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers Submitted, Paper2816 Authors Show Revisions (/revisions?id=kk4Obkr-D4)

Recommended Process Of Reviewing: I have read the instructions above

Paper Summary:
The work performs a causal analysis into the large language model by considering the LLM's internal state as a random variable and considering the external knowledge as an instrument variable. The authors consider SCM settings for both standard as well as chain-of-thought modeling, and guided by their analysis they show the impact of different chains of thought and external knowledge through average causal effect analysis. Motivated by their findings, they provide a debiasing strategy for CoT models and test it on QA datasets, highlighting their superior performance over the baselines considered.

Summary Of Strengths:
- The direction of research is quite interesting, and I believe leveraging causal estimation strategies in the context of large language modeling is an under-explored but important direction of research.
- The authors conduct experiments on QA and show that their proposed debiasing technique does indeed lead to better performance.

Summary Of Weaknesses:
- It is unclear what the authors mean when they consider the variable Z? Do they mean the hidden representation in the LLM or the weights of the LLM itself? In either case, why would the query be causally dependent on Z since the query Q is provided by the user and is dependent on nothing? How is C not dependent on Z? (in reference to Fig. 2 a-c)

- What is the benefit of considering $v^*_{k,j}$ in Equation (3), as opposed to just v_j?
- It was unclear throughout the paper whether the authors wanted to study the effect of interventions on external knowledge or chain of thoughts to the answer?
- In Equation (4), when the authors consider interventions on the external information $do(E)$, should that not change the CoT C, as it is causally dependent on E?
- The authors do not provide detailed analysis of the inference costs when using DeCoT? Essentially, it looks like computing ACE(C_i) would require a lot of inference over an LLM?

Comments, Suggestions And Typos:

- Did the authors mean to write Biohazard instead of Kekal in the second to last line in the caption for Figure 1?

Soundness: 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.

Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.

Confidence: 2 = Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.

Best Paper: No

Limitations And Societal Impact:
The authors discuss the limitations of the proposed work.

Ethical Concerns:
N/A

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Add **Author-Editors Confidential Comment**

[–] Response to Reviewer

ACL ARR 2024 February Paper2816 Authors *Junda Wu (/profile?id=~Junda_Wu1) (privately revealed to you)*

30 Mar 2024, 02:32 ACL ARR 2024 February Paper2816 Official Comment Readers: Program Chairs, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers Submitted, Paper2816 Authors Show Revisions (/revisions?id=u2rwXnWoCY)

Comment:

We sincerely appreciate your time in reviewing our paper. Your comments and suggestions are very insightful and helpful to our work.

Response to Weaknesses 1:

By considering the variable Z, we identify and leverage this confounder in causal debiasing to improve the generated results. Without parameterizing the confounders (e.g., parameterization as the representation or weights), other techniques such as counterfactual/human-in-the-loop intervention and instrumental variables, were also investigated for causal debiasing in previous works [2, 4, 6, 7]. In previous works of causal reasoning in LLMs, the confounder can be identified as pre-trained knowledge [2, 6] in LLMs. Similar to [2, 4, 6, 7], our approach is based on the counterfactual intervention and instrumental variable, and the confounder Z is the

LLM's internal knowledge (e.g., ``Jakarta is Kekal'' in Figure 1). As these LLMs are usually pre-trained on the prompts or instructions provided by the human, unavoidably there will be some dependency between the user input query/prompt and the pre-trained knowledge in LLMs [2, 8, 9, 10] (e.g., as discussed in Fig. 4(a) of [2]). Thanks for pointing out that C depends on Z. We missed the arrowed line from Z to C when drawing Figure 2. We will fix this in an updated version. Our approach is actually designed based on the fact that C depends on Z.

Response to Weaknesses 2:

We would like to clarify that $v^*_{j,k}$ and v_j are two possible values of the variable V, where k is the index of the k-th counterfactual entity. By considering counterfactual entities $v^*_{j,k}$, we can conduct the counterfactual intervention on the reasoning paths, which enables causal effect estimation [4].

Response to Weaknesses 3:

In our work, we investigate the causal effect of the chain of thoughts on the answer, as described in Section 5.2. To estimate the causal effect, we introduce external knowledge as the instrumental variable [5]. Based on the estimated causal effect, we propose a causal intervention method on the chain of thoughts.

Response to Weaknesses 4:

By changing the external knowledge E (instrumental variable) and keeping the CoT variable unchanged, we can enable the instrumental variable for causal effect estimation. Similar techniques are also investigated in [5, 6, 11]. Intuitively, such counterfactual intervention enables us to see how the counterfactual knowledge (E), together with the unchanged CoT variable, will lead to the change of the final answers.

References: (Most of the references listed below are already mentioned in the original paper. We reindexed the references for easy presentation in the rebuttal.)

[1] D'Amour, Alexander. "On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples." Impossibility, and Alternatives (2019).

[2] Knowledge, Grounding. "The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models."

[3] Simon, Herbert A. "Spurious correlation: A causal interpretation." Journal of the American statistical Association 49.267 (1954): 467-479.

[4] Zeng, Xiangji, et al. "Counterfactual generator: A weakly-supervised method for named entity recognition." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[5] Kawakami, Yuta, Manabu Kuroki, and Jin Tian. "Instrumental variable estimation of average partial causal effects." International Conference on Machine Learning. PMLR, 2023.

[6] Yuan, Junkun, et al. "Instrumental variable-driven domain generalization with unobserved confounders." ACM Transactions on Knowledge Discovery from Data 17.8 (2023): 1-21.

[7] Hu, Zhiting, and Li Erran Li. "A causal lens for controllable text generation." Advances in Neural Information Processing Systems 34 (2021): 24941-24955.

[8] Zhang, Shengyu, et al. "Devlbert: Learning deconfounded visio-linguistic representations." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

[9] Cao, Boxi, et al. "Can prompt probe pretrained language models? understanding the invisible risks from a causal view." arXiv preprint arXiv:2203.12258 (2022).

[10] Wang, Siyin, et al. "Causal intervention improves implicit sentiment analysis." arXiv preprint arXiv:2208.09329 (2022).

[11] Yue, Zhongqi, et al. "Interventional few-shot learning." Advances in neural information processing systems 33 (2020): 2734-2746.

Response to Reviewer

ACL ARR 2024 February Paper2816 Authors Junda Wu (/profile?id=~Junda_Wu1) (privately revealed to you)

30 Mar 2024, 02:33 ACL ARR 2024 February Paper2816 Official Comment Readers: Program Chairs, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers Submitted, Paper2816 Authors Show Revisions (/revisions?id=265ApjPBb1)

Comment:

Response to Weaknesses 5:

We have analyzed the possible inference overhead in Section 6.5 and Section 6.6, which suggests that our method can still improve on CoT prompting without a large overhead. Based on the observations in Figure 3, our approach can already achieve clear improvement compared to the baselines, with inference cost that is one time higher (i.e., with twice the inference cost). Besides, in practice counterfactual inference can be parallelly done since generating more counterfactual examples improves the robustness of causal effect estimation but does not block the inference pipeline, such that our approach does not require additional latency.

Response to Comments:

Thanks for the comment. 'Kekal' is what we intend to mention in the second to last line in the caption for Figure 1. In this example, we mean to explain that the spurious correlation, in this case, could be GPT3.5's spurious correlation on an arbitrary sentence ''The heavy metal band formed in Jakarta is Kekal.'' in the given context, which is irrelevant to the real user query in this example. Such spurious correlation could lead to incorrect final answers (i.e., the failure case in Figure 1).

Add **Author-Editors Confidential Comment**

[–] ## Response to Authors

ACL ARR 2024 February Paper2816 Reviewer 5jLu

30 Mar 2024, 12:35 ACL ARR 2024 February Paper2816 Official Comment Readers: Program Chairs, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers Submitted, Paper2816 Authors Show Revisions (/revisions?id=Mey9OeMivq)

Comment:

Thanks to the authors for providing clarifications. I have raised my score a bit based on the authors' comments, but I am not satisfied with their reasoning behind a causal connection from Z to Q. In particular, an intervention on Z should not affect Q in any form. If the authors can provide some justification behind this, I would be amenable to raising my score.

Add **Author-Editors Confidential Comment**

[–] ### Response to Reviewer

ACL ARR 2024 February Paper2816 Authors Junda Wu (/profile?id=~Junda_Wu1) (privately revealed to you)

31 Mar 2024, 00:50 (modified: 31 Mar 2024, 05:32) ACL ARR 2024 February Paper2816 Official Comment Readers: Program Chairs, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers Submitted, Paper2816 Authors Show Revisions (/revisions?id=8WYdgv3i4Y)

Comment:

Thanks for your consideration and the additional comment. We greatly appreciate your comments and the opportunity to clarify any questions you may have.

The connection from Z to Q (i.e., some confounder as the parent node of the user's input) has been similarly proposed in causal debiasing for various applications in NLP or CV [2, 4, 8, 9, 10, 12, 13, 14, 15, 16, 17]. In Fig. 4(a) of [2], Fig. 2 of [10], Fig. 2 of [14], Fig. 2 of [16], Fig. 1 of [17], Fig. 2 of [4], Fig. 1 of [8], Fig. 2 of [9], Fig. 2 of [12], Fig. 2 of [13], and Fig. 3 of [15], it is widely adopted that there are some confounder variables (e.g., pre-trained knowledge, sentiment words, noun words in the queries) which are parent nodes of the user's input of query/prompt/sentence/image.

Besides, we would like to clarify that in our approach, instead of intervention on Z as you mentioned, our approach only intervenes on the variable C (i.e., do(C) in Eq. 7), and observes the outcome variable A.

[2] Knowledge, Grounding. "The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models."

[4] Zeng, Xiangji, et al. "Counterfactual generator: A weakly-supervised method for named entity recognition." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[8] Zhang, Shengyu, et al. "Devlbert: Learning deconfounded visio-linguistic representations." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

[9] Cao, Boxi, et al. "Can prompt probe pretrained language models? understanding the invisible risks from a causal view." arXiv preprint arXiv:2203.12258 (2022).

[10] Wang, Siyin, et al. "Causal intervention improves implicit sentiment analysis." arXiv preprint arXiv:2208.09329 (2022).

[12] Yue, Zhongqi, et al. "Transporting causal mechanisms for unsupervised domain adaptation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[13] Chen, Yuedong, et al. "Towards unbiased visual emotion recognition via causal intervention." Proceedings of the 30th ACM International Conference on Multimedia. 2022.

[14] Lu, Yujie, et al. "Neuro-Symbolic Procedural Planning with Commonsense Prompting." The Eleventh International Conference on Learning Representations. 2022.

[15] Wang, Tan, et al. "Causal attention for unbiased visual recognition." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[16] Wang, Fei, et al. "A causal view of entity bias in (large) language models." arXiv preprint arXiv:2305.14695 (2023).

[17] Zhao, Haiteng, et al. "Certified robustness against natural language attacks by causal intervention." International Conference on Machine Learning. PMLR, 2022.

Add **Author-Editors Confidential Comment**

View 1 More Reply →

[–] ## Official Review of Paper2816 by Reviewer 9iMi

ACL ARR 2024 February Paper2816 Reviewer 9iMi

21 Mar 2024, 01:43 ACL ARR 2024 February Paper2816 Official Review Readers: Program Chairs, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers Submitted, Paper2816 Authors Show Revisions (/revisions?id=rtnBkJN_y9)

Paper Summary:
The paper presents an approach to selecting the final reasoning path and answer from multiple chain-of-thought (CoT) generations produced by large language models (LLMs). Specifically, the method prompts LLMs to generate multi-hop reasoning paths, estimates the Average Causal Effect (ACE) of reasoning paths and answers by constructing counterfactual contexts, and ultimately samples the final path and answer.

Summary Of Strengths:
- The causal perspective adopted for analysis and the methodology proposed are novel and offer a fresh angle to approach CoT in LLMs.
- The method outperforms the vanilla CoT approaches across multiple datasets, demonstrating strong empirical results.

Summary Of Weaknesses:
I was the reviewer for this paper in the previous submission round. Below are the weaknesses I pointed out earlier. In this round of submission, the description of their method has alleviated my concerns regarding its generalization ability/ground-truth assumption. Additionally, the inclusion of extra experiments (React) has also mitigated my worries about the baselines in this paper.

> 1. A key assumption of the method is the need for a gold-truth context, which could limit its generalizability. The absence of gold-truth context or the presence of extensive irrelevant contexts could lead to failure or high costs due to the necessity of constructing counterfactuals for each fact. The authors should include more experimental results using non-gold-truth contexts to validate the robustness of their approach.
> 2. The paper's essence is to utilize context better for generating more accurate answers. However, it only compares with the most basic CoT methods and lacks comparison with recent, stronger methods of context usage and model reasoning, such as React and DSP. Recent methods have been presented in this paper: https://arxiv.org/abs/2309.15402 (https://arxiv.org/abs/2309.15402). The authors should include stronger baseline models to establish the superiority of their method.

Comments, Suggestions And Typos:
- Please provide specific details regarding the size of the test sets and the evaluation setup.
- Please provide the implementation details of the React method.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.

Overall Assessment: 3.5

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Best Paper: No

Ethical Concerns:
None

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: 9iMi

Add **Author-Editors Confidential Comment**

[-] Response to Reviewer

ACL ARR 2024 February Paper2816 Authors Junda Wu (/profile?id=~Junda_Wu1) (privately revealed to you)

30 Mar 2024, 02:36 ACL ARR 2024 February Paper2816 Official Comment Readers: Program Chairs, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers Submitted, Paper2816 Authors Show Revisions (/revisions?id=F_6cdCEjIf)

Comment:

We sincerely appreciate your time in reviewing our paper. Your comments and suggestions are very insightful and helpful to our work.

Response to Comment 1: We have some details regarding the size of the test sets and the evaluation setup in Section 6.1. We will add more details in our updated version.

Response to Comment 2: We use the same codebase from React (https://react-lm.github.io/ (https://react-lm.github.io/)) to get the retrieved context as the external knowledge in our experiment. Then, during question-answering, we replace the context with React-retrieved external knowledge. We will add more details in Section 6.4 in our updated version.

Thanks for your recognition of our work and your valuable suggestions.

Add **Author-Editors Confidential Comment**

[–] # Official Review of Paper2816 by Reviewer DEvz

ACL ARR 2024 February Paper2816 Reviewer DEvz

18 Mar 2024, 20:27 ACL ARR 2024 February Paper2816 Official Review Readers: Program Chairs, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers Submitted, Paper2816 Authors Show Revisions (/revisions?id=vjICnUci0g)

Recommended Process Of Reviewing: I have read the instructions above

Paper Summary:

This work points out the latent internal knowledge bias of LLMs, especially when performing CoT, with the help of Structural Causal Model (SCM). And the authors proposed Debiasing Chain-of-Thought (DeCoT), a CoT resampling method to conduct the front-door adjustment on the probabilities, with the estimated average causal effect (ACE) of each CoTs as reweighing parameters. DeCoT can help LLMs find more accurate and logically sound responses in knowledge-intensive tasks.

Summary Of Strengths:

1. The authors proposed an innovative approach of utilizing causal intervention for debiasing the CoT reasoning in LLMs, marking a significant advancement in this domain.
2. The authors describe the formalized derivation of ACE in details. The main sampling method, DeCoT, is further constructed under the guidance of theoretical basis of ACEs.
3. Relatively sufficient experiments, including serval datasets, models and reasoning path, as well as the different hyper-parameter settings and case studies.

Summary Of Weaknesses:

1. It seems that the key of DeCoT resampling method is reweighing the possible CoTs, which results in two questions. The first one is the performance is directly limited by the quality and diversity of generated CoTs. Then does DeCoT perform relative better in larger LLMs than smaller ones? It should be checked and explained in detail.
2. The second concern is that statistically reweighing needs statistically analyzing, which means there should be more quantitative exploration on how the final result after reweighing beats other ones. Without doing so, we will not be able to find a reliable statistical explanation for the performance improvement.
3. There should be more powerful approaches as the compared baselines, such as GPT-4 which is the most knowledgeable LLM, to evaluate the upper limit and to quantize the improvement of the proposed DeCoT sampling method.

Comments, Suggestions And Typos:

1. There is a lack of intuition of meaning of ACEs and the causal intervention in the front-door adjustment via ACE scores. What is the role of ACE, i.e., the decreased confidence of the answer with counterfactual context as the evidence, in the reweighing process of CoTs?
2. It would be better to supplement the details about the expenses, such as used the tokens and inference time in their experiment results.

Soundness: 3.5

Overall Assessment: 3.5

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Best Paper: No

Ethical Concerns:

N/A

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: DEvz

Add **Author-Editors Confidential Comment**

[–] ## Response to Reviewer

ACL ARR 2024 February Paper2816 Authors Junda Wu (/profile?id=~Junda_Wu1) (privately revealed to you)

30 Mar 2024, 02:38 ACL ARR 2024 February Paper2816 Official Comment Readers: Program Chairs, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers Submitted, Paper2816 Authors Show Revisions (/revisions?id=G7yPfvRvLo0)

Comment:

We sincerely appreciate your time in reviewing our paper. Your comments and suggestions are very insightful and helpful to our work.

Responses to Weaknesses 1:

According to the results in Table 1, we observe that DeCoT performs relatively better in larger LLMs than in smaller ones, which suggests that the performance of our method is related to the quality and diversity of generated CoTs. As shown in Table 1, the improvements of GPT3.5 on CoT are 7.67 (EM) and 8.84 (F1) are larger than those of LlaMA-2 (6.47 and 8.2) and Flan-T5 (5.95 and 6.6). Such observation is also consistent with the evaluation results related to the diversity in [1]. We will add a detailed discussion in an updated version.

Response to Weaknesses 2:

In addition to the results that our approach can outperform the baselines validated in Table 1, we further report the standard errors from our methods for every backbone model in each dataset (F1 and EM metrics).

Stand. Error	HotpotQA-EM	HotpotQA-F1	MuSiQue-EM	MuSiQue-F1	SciQ-EM	SciQ-F1	WikiHop-EM	WikiHop-F1
Flan-T5	1.23	1.52	2.04	2.05	1.66	1.31	1.46	1.49

Stand. Error	HotpotQA-EM	HotpotQA-F1	MuSiQue-EM	MuSiQue-F1	SciQ-EM	SciQ-F1	WikiHop-EM	WikiHop-F1
LLaMA-2	1.63	2.08	2.26	2.19	1.68	1.51	1.62	1.61
GPT-3.5	1.15	1.48	2.29	2.21	1.81	1.37	1.72	1.64

We observe relatively low randomness in our sampling method, which showcases the robustness of DeCoT.

Responses to Weaknesses 3:

For a fair comparison between our approach and the baseline methods [2, 3, 4], we consider GPT-3.5 as the comparable backbone model used in [2, 3, 4]. We will consider adding more powerful and recent LLMs as backbone models in our updated version of experiments.

Response to Comment 1:

With a decreased confidence of the answer with counterfactual context as the evidence, the selected CoT can be evaluated as more sensitive to external counterfactual knowledge, which suggests the CoT can better use the external knowledge and generate a higher quality response, as illustrated in Figure 2. We have provided more case studies in Table 3 of Section 6.7 that provide more insights into specific use cases.

Response to Comment 2:

We have analyzed the impact and expenses of selected counterfactual entity tokens in Section 6.5 and 6.6. We have also discussed comparative inference time and API call numbers in Section 6.5 and Section 2. We will add more details about the general inference cost as suggested.

References: (Most of the references listed below are already mentioned in the original paper. We reindexed the references for easy presentation in the rebuttal.)

[1] Chen, Yihan, et al. "Benchmarking large language models on controllable generation under diversified instructions." arXiv preprint arXiv:2401.00690 (2024).

[2] Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." arXiv preprint arXiv:2210.03629 (2022).

[3] Shi, Weijia, et al. "Trusting your evidence: Hallucinate less with context-aware decoding." arXiv preprint arXiv:2305.14739 (2023).

[4] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in neural information processing systems 35 (2022): 24824-24837.

Add **Author-Editors Confidential Comment**

[–] Response to Authors

ACL ARR 2024 February Paper2816 Reviewer DEvz

02 Apr 2024, 01:57 ACL ARR 2024 February Paper2816 Official Comment Readers: Program Chairs, Paper2816 Senior Area Chairs, Paper2816 Area Chairs, Paper2816 Reviewers Submitted, Paper2816 Authors Show Revisions (/revisions?id=tKxpcpAh22m)

Comment:

Thank you for your clarification~

Add **Author-Editors Confidential Comment**

[–] Supplementary Materials by Program Chairs

ACL ARR 2024 February Program Chairs

16 Feb 2024, 13:34 ACL ARR 2024 February Paper2816 Supplementary Materials Readers: Program Chairs, Paper2816 Reviewers, Paper2816 Authors,

Paper2816 Area Chairs, Paper2816 Senior Area Chairs Show Revisions (/revisions?id=4b8_GF-JzrH)

Previous URL: /forum?id=JGVgR5L0iDq (/forum?id=JGVgR5L0iDq)

Previous PDF: pdf (/attachment?id=4b8_GF-JzrH&name=previous_PDF)

Response PDF: pdf (/attachment?id=4b8_GF-JzrH&name=response_PDF)

Reassignment Request Action Editor: No, I want the same action editor from our previous submission and understand that a new action editor may be assigned if the previous one is unavailable

Reassignment Request Reviewers: No, I want the same set of reviewers from our previous submission and understand that new reviewers may be assigned if any of the previous ones are unavailable

A1: yes

A1 Elaboration For Yes Or No: Section 8

A2: n/a

A3: yes

A3 Elaboration For Yes Or No: Abstract and Section 1

B: no

B1: n/a

B2: n/a

B3: n/a

B4: n/a

B5: n/a

B6: yes

B6 Elaboration For Yes Or No: Section 6

C: yes

C1: yes

C1 Elaboration For Yes Or No: Section 6

C2: yes

C2 Elaboration For Yes Or No: Section 6

C3: yes

C3 Elaboration For Yes Or No: Section 6

C4: yes

C4 Elaboration For Yes Or No: Section 6

D: no

D1: n/a

D2: n/a

D3: n/a

D4: n/a

D5: n/a

E: no

E1: n/a

Note From EiCs: These are the confidential supplementary materials of the submission. If you see no entries in this comment, this means there haven't been submitted any.

Add **Author-Editors Confidential Comment**