Unveiling and Causalizing CoT: A Causal Pespective

Jiarun Fu[1] , **Lizhong Ding**[1] , **Hao Li**[1] , **Pengqi Li**[1] , **Qiuning Wei**[1] , **Xu Chen** [2]

[1]Beijing Institute of Technology
[2]Renmin University of China
3120235197@bit.edu.cn

Abstract

Although Chain-of-Thought (CoT) has achieved remarkable success in enhancing the reasoning ability of large language models (LLMs), the mechanism of CoT remains a "black box". Even if the correct answers can frequently be obtained, existing CoTs struggle to make the reasoning understandable to human. In this paper, we unveil and causalize CoT from a causal perspective to ensure both correctness and understandability of all reasoning steps (to the best of our knowledge, the first such). We model causality of CoT via structural causal models (SCM) to unveil the reasoning mechanism of CoT. To measure the causality of CoT, we define the CoT Average Causal Effect (CACE) to test the causal relations between steps. For those steps without causality (wrong or unintelligible steps), we design a role-playing causal query algorithm to causalize these steps, resulting a causalized CoT with all steps correct and understandable. Experimental results on both open-source and closed-source LLMs demonstrate that the causal errors commonly in steps are effectively corrected and the reasoning ability of LLMs is significantly improved.

1 Introduction

> "We do not have knowledge of a thing until we grasped its cause."
>
> — Aristotle

Causality [Pearl, 2009; Yao *et al.*, 2021] provides a unique perspective for exploring the mechanisms of machine learning algorithms, making significant contributions in areas such as trustworthy learning [Li *et al.*, 2024; Xie *et al.*, 2024] and stable learning [Liu and Kuang, 2023; Wu *et al.*, 2023a; Zhang *et al.*, 2024b], etc. Large language models (LLMs) have made remarkable progresses in recent years [Wu *et al.*, 2023b; Yang *et al.*, 2024; Liu *et al.*, 2021], and it has been confirmed that LLMs can perform step-by-step reasoning through a Chain-of-Thought (CoT), which breaks down complex problems into a sequence of step-by-step thinking processes [Kojima *et al.*, 2022; Hu *et al.*, 2023; Chen *et al.*, 2024]. Although CoT has led to remarkable achievements, it does not always provide positive outcomes [Sprague *et al.*,

2024] and sometimes hinders reasoning performance [Lanham *et al.*, 2023]. So it is crucial to explore the mechanism of CoT and make it understandable. Therefore, analyzing CoT or LLMs' reasoning through causality offers a natural solution. [Chen *et al.*, 2022; Kıcıman *et al.*, 2023; Bhattacharjee *et al.*, 2024; Jin *et al.*, 2023b].

However current research suggests LLMs behave like "causal parrots", merely reciting causal knowledge without truly understanding causality[Zečević *et al.*, 2023; Wu *et al.*, 2024], which limits the applicability of existing causal-based CoT methods, only in knowledge based tasks [Wu *et al.*, 2024] and causal inference tasks [Jin *et al.*, 2023a; Zhang *et al.*, 2024a]. Beyond the community of causality, researches on the unveiling mechanism of CoT focus on the upper bound of reasoning ability [Feng *et al.*, 2024; Chen *et al.*, 2024] and contextual demonstration ability [Madaan *et al.*, 2023], even so CoT remains a black box. We summarize the shortcomings of current researches as follows:

1. The lack of models for unveiling the mechanism of CoT to make the reasoning of LLMs interpretable.

2. The lack of algorithms that causalize all steps of CoT to make them correct and understandable.

In this paper, we unveil the reasoning mechanism of CoT from a causal perspective and causalize CoT to make LLMs reasoning both correct and understandable. Since causality is the most understandable and learnable logical relationship for humans [Rubin, 1980; Pearl, 2009; Kaddour *et al.*, 2022], we assume that CoT's effectiveness in reasoning derives from its reflection of real-world causal relationships involved in problem-solving. The example in Figure 1 serves as a demonstration of this assumption. We construct a constructing structural causal model (SCM) of CoT to model the causal relations between reasoning steps, and further establish the SCM for unveiling mechanism of reasoning. In order to quantify the causality of CoT, we propose the CoT Average Causal Effect (CACE) to measure the causal relationship between steps of CoT from both answer and logic aspects. Based on empirical evidence, the first step of CoT largely determines the logic, we introduce the First-Step Causal Effect (FSCE) to quantify the casual logic of CoT reasoning. To test whether CoT has been causalized, we complete the causal inference of steps by CACE and FSCE to test the causality of each steps. For those CoTs have not been causalized, we apply a role-playing causal query algorithm including the causalizing process and the refine process to establish
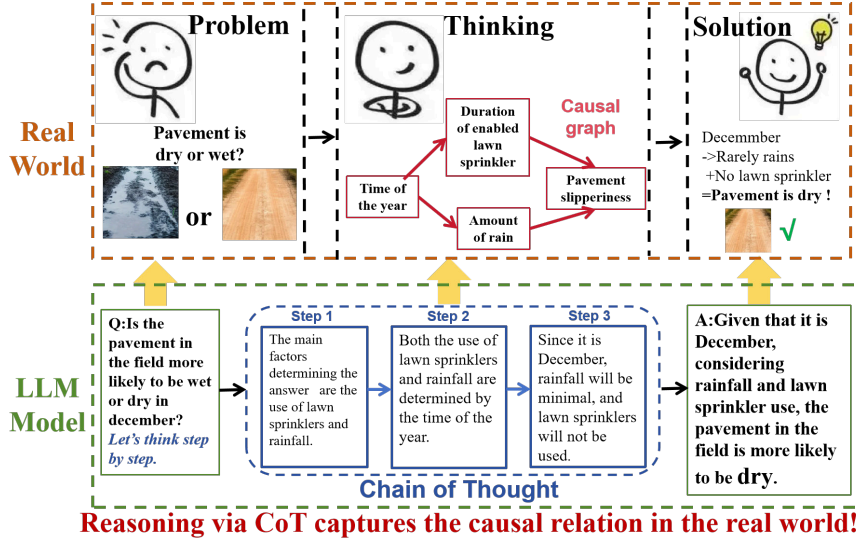
Figure 1: We assume the ability of CoT to reason correctly stems from its reflection of real-world causal relationships. As shown in the figure above, in the real world, based on our life experience, we infer the causal graph between variables such as month, rainfall, sprinkler use, and pavement slipperiness, then use this graph to deduce the answer. Similarly, LLMs employ Chain-of-Thought (CoT) to perform a reasoning process that aligns with these real-world causal relationships, ultimately arriving at the correct answer.

the causal logic for all steps. Through extensive experiments on both open-source and closed-source LLMs, we correct multiple causal error types between steps and improve significantly in capabilities of LLMs reasoning.

Our main contributions are as follows:

1. To the best of our knowledge, we are the first to unveils the mechanism of CoT through causality. We apply SCM for modeling the causality of CoT to make the reasoning of LLMs interpretable.

2. In order to measure the causality of CoT, we propose Average Causal Effect (CACE) and First-Step Causal Effect (FSCE) to infer the causal logic between each reasoning step. Through causal inference, we summarize common causality errors in CoT.

3. We propose a role-playing causal query algorithm to causalize reasoning steps that lack causal logic, ensuring the correctness and understandability of CoT.

2 Background

Since SCM have been proven effective in exploring potential mechanisms, we introduce them to model the causality of CoT.

2.1 Structural causal model (SCM)

A structural causal model [Pearl, 2009; Yao *et al.*, 2021; Kaddour *et al.*, 2022] $\mathcal{M}$ is a 3-tuple $\langle V, U, \mathcal{F} \rangle$, where:

1. $U = \{u_1, ..., u_m\}$ is the set of exogenous variables, also called background variables, which are determined by factors outside the model;

2. $V = \{v_1, ..., v_n\}$ is the set of endogenous variables, which are determined by the variables in the model;

3. $\mathcal{F} = \{f_{v_1}, ..., f_{v_n}\}$ is the set of structural functions determining V, where $f_{v_i} : (pa(v_i), U) \mapsto v_i$ and $pa(v_i) \subseteq (V \setminus \{v_i\})$.

SCM are often presented for the case of intervention, which is denoted using *do-operator* $do(\cdot)$ to implement

$do(T = t)$ [Singh *et al.*, 2020; Kaddour *et al.*, 2022], where T represens treatment and t represents actual treatment content. Treatment effect is used to measure the impact of an interventional treatment, that is, the causal relationship between t and the outcome of interest Y. The most widely used average treatment effect (ATE) is defined as $\gamma(t) := \mathbb{E}[Y \mid do(T = t)]$ and Conditional Average Treatment Effect (CATE) is defined as $\gamma(t, u_i) := \mathbb{E}[Y \mid do(T = t), u_i]$, where $u_i \in U$ are exogenous variables. It is important to note that conditioning on $T = t$ means that we are looking at the outcomes for the subset of the population who have received the treatment t in the observation data, so $\mathbb{E}[Y \mid do(T = t)] \neq \mathbb{E}[Y \mid T = t]$.

2.2 Chain-of-Thought reasoning

Referring to previous CoT work [Qiao *et al.*, 2022; Chu *et al.*, 2024; Xiang *et al.*, 2025], we define the notations as follows: the question Q, the instruction IS, LLM p_{LM} and the answer $A = [a_1, a_2, \ldots, a_n]$. The model takes the Q and IS as inputs to produce the answer A as its output:

$$p(A \mid IS, Q) = \prod_{i=1}^{n} p_{\text{LM}}(IS, Q, a_{<i}). \tag{1}$$

LLM generates its reasoning paths $C = [c_1, c_2, ..., c_n]$ by explaining them step by step before producing the final answer A:

$$p(A, C \mid IS, Q) = p(A \mid C, IS, Q) \cdot p(C \mid IS, Q),$$

$$p(C \mid IS, Q) = \prod_{i=1}^{n} p_{\text{LM}}(IS, Q, c_{<i}),$$

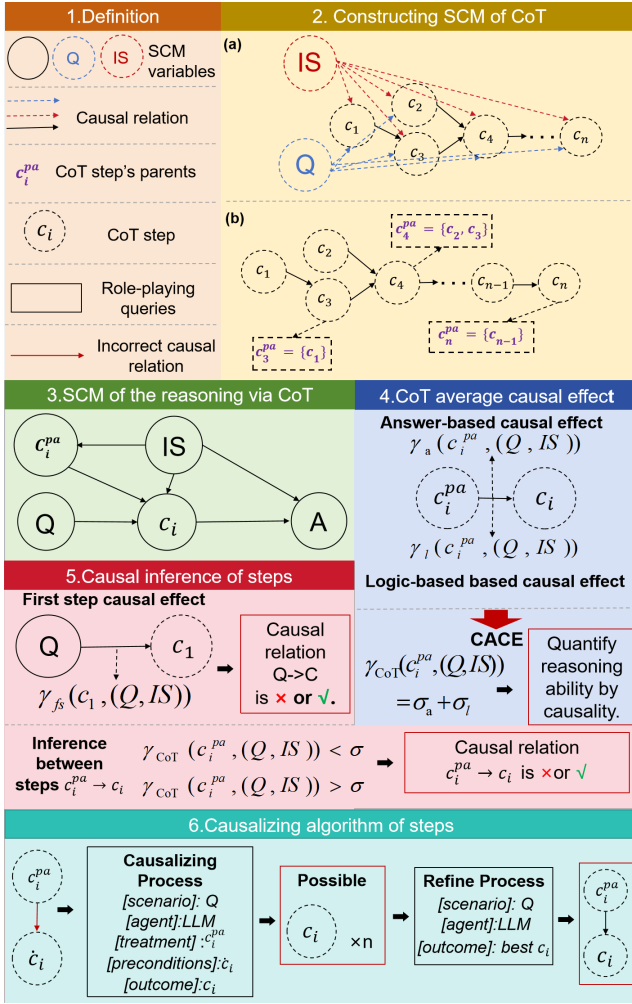$$p(A \mid IS, Q, C) = \prod_{j=1}^{n} p_{\text{LM}}(IS, Q, C, a_{<j}).$$

Figure 2: From modeling CoT to causalizing CoT.

3 Opening the black box of CoT: By SCM modeling

Since causality is one of the most intuitive forms of logic for humans, we aim to unveil the mechanism of CoT from a causal pespective. Building on this idea, we assume that the reasoning ability of CoT arises from real-world causality, where causal relationships exist between each step, thus reflecting the correct causal graph inherent in the problem's solution. We model causality of CoT via SCM to make LLMs' reasoning understandable. With reasonable assumptions on the causal structure, we define the CACE to measure the causality of CoT.

3.1 Constructing SCM of CoT

We treat IS and Q as exogenous variables in SCM since they do not change during reasoning. Then we define $C = [c_1, c_2, ..., c_n]$ as endogenous variables, $p_{LM}(\cdot)$ as structural functions. As we assume the CoT reflects the causal graph, c_i should be influenced by its parent nodes c_i^{pa}, Q, and IS, where c_i^{pa} is a subset of the power set $2^{[c_1, c_2, ..., c_{i-1}]}$. Then we can get the constructing SCM of CoT as:

$$c_i = p_{LM}\left(IS, Q, c_i^{pa}\right). \tag{2}$$

As shown in Figure 2, we explain the detials of constructing SCM of CoT and the definition of $C^{pa} = [c_1^{pa}, c_2^{pa}, ..., c_n^{pa}]$.

3.2 SCM of the reasoning via CoT

Based on the constructing SCM of CoT, we redefine the reasoning via CoT fomular (2.2) to causal version to model CoT reasoning mechanism. We define LLMs answer the Q with a_i based on IS and step-by-step reasoning path c_i , which is generated based on c_i^{pa}. We format SCM of the reasoning via CoT as:

$$p(a_i, c_i \mid IS, Q, c_i^{pa}) = $$
$$p(a_i \mid c_i, IS, Q, c_i^{pa}) \cdot p(c_i \mid IS, Q, c_i^{pa}).$$

Following the constructing SCM of CoT, we can define:

$$p(C \mid IS, Q, C^{pa}) = \prod_{i=1}^{n} p_{LM}(IS, Q, c_{<i}^{pa}),$$

$$p(A \mid IS, Q, C, C^{pa}) = \prod_{j=1}^{n} p_{LM}(IS, Q, C, C^{pa}, a_{<j}).$$

We illustrate SCM of the reasoning via CoT in Figure 2. One important point to note is that, according to the Markov rule of SCM, c_i has a direct causal relationship only with its parent nodes c_i^{pa}, not with the previous steps $c < i$ from a traditional perspective. Therefore, the causal path $Q \rightarrow c_i \rightarrow A$ shows that c_i can serve as a mediator between the question Q and the answer A. Similarly, the causal relationship between a_{i-1} and a_i should only be reflected by c_i, otherwise, a collider bias will occur between A and the parent nodes of C, C^{pa}.

3.3 CoT average causal effect

Since we have successfully established a SCM for CoT, the subsequent question is how to quantify the causality within CoT. Since the causality of CoT should possess both logical coherence and correctness, we quantify the causal relationship $c_i^{pa} \rightarrow c_i$ based on the logic of reasoning and the correctness of answers corresponding to a_i. Using the $do(\cdot)$ operator, we define the logic-based average causal effect γ_l and the answer-based average causal effect γ_a as follows:

$$
\begin{aligned}
&\gamma_a(c_i^{pa}, (Q, IS)) \\
=&\mathbb{E}\left[(a_i \mid IS, Q, do(c_i^{pa})) - (a_i \mid IS, Q, c_i^{pa})\right], \\
&\gamma_l(c_i^{pa}, (Q, IS)) \\
=&\mathbb{E}[(c_i \mid IS, Q, do(c_i^{pa})) - (c_i \mid IS, Q, c_i^{pa})].
\end{aligned} \tag{3}
$$

Based on γ_l and γ_a, we propose CoT average causal effect of CoT (CACE) γ_{CoT} to quantify causal relationships of $c_i^{pa} \rightarrow c_i$:

$$
\begin{aligned}
&\gamma_{CoT}(c_i^{pa}, (Q, IS)) \\
=&\gamma_a(c_i^{pa}, (Q, IS)) + \gamma_l(c_i^{pa}, (Q, IS)).
\end{aligned} \tag{4}
$$

Assumptions for SCM of CoT

To eliminate the well-known selection bias, we introduce three standard assumptions that are commonly relied upon in existing SCM [Rubin, 1980; Pearl, 2009; Yao et al., 2021; Li et al., 2023; Zeng et al., 2024; Jin et al., 2023a; Zhang et al., 2024a; Wu et al., 2024].

Assumption 1 (Stable Unit Treatment Value (SUTVA)). *The distribution of the a_i and c_i with $p_{LM}(\cdot)$ are assumed to be independent of the interventions $do(c_i^{pa})$.*

Since $do(c_i^{pa})$ intervenes in the step of reasoning and modifies $p_{LM}(\cdot)$, the distributions of A and C remain unaffected by this intervention. SUTVA ensures that equation (4) accurately reflects the causal relationships $c_i^{pa} \rightarrow c_i$ [Qi *et al.*, 2023; Wu *et al.*, 2023a; Zhang *et al.*, 2024b].

Assumption 2 (Unconfoundedness). *The distribution of $do(c_i^{pa})$ is independent of a_i and c_i, given exogenous variables IS, Q.*

Our second assumption extends the unconfoundedness assumption to equation (4), assuming that there are no unobserved variables that have a causal relationships with the intervention. The unconfoundedness assumption ensures that the treatment of c_i^{pa} accurately reflects the causal relationship $c_i^{pa} \rightarrow c_i$ [Xie *et al.*, 2020; Qi *et al.*, 2023; Zuo *et al.*, 2024].

Assumption 3 (Overlap). *Every unit should have a nonzero probability to receive either treatment status. Formally, $0 < p(do(c_i^{pa}) \mid Q, IS) < 1$.*

The overlap assumption is fundamental in the field of causal inference, which ensures the validity and feasibility of c_i^{pa} in (4) [Li *et al.*, 2023; Zeng *et al.*, 2024; Jin *et al.*, 2023a].

By using γ_{CoT}, we can successfully quantify the causal relationships between steps in CoT, enabling us to test whether reasoning steps have causal logic.

4 CauCoT: Causalized Chain-of-Thought

We have successfully constructed the SCM for CoT to demonstrate the causality within its reasoning mechanism. However, to make LLMs reasoning both correct and understandable, we still need a method to causalize CoT in order to ensure steps of CoT both correct and casually logical. In this section, we propose our method Causalized CoT (CauCoT), including causal inference of steps and causalizing algorithm of steps.

4.1 Causal inference of steps

Since steps of causalized CoT are both correct and understandable, to test whether there is correct causal relation between steps, we need to quantify the causality inherent in CoT from both the correctness of answer and the logic of reasoning aspects. Based on (4), we can effectively measure the causal relationships between the steps of CoT. However, we still lack a method to measure the causal logic between CoT and the question Q (since IS is assumed to have an equal causal effect on all SCM components except for Q, it does not require separate analysis). Based on empircial evidence, if the first step, c_1, does not have causality to answer Q, subsequent steps will be unable to has causal logic since they are built on incorrect basic (this will be further discussed in the Appendix A.3). Therefore, we propose the First-Step Causal Effect (FSCE), denoted as γ_{fs}, to measure the causality between CoT and the question. We define γ_{fs} as:

$$
\begin{aligned}
&\gamma_{fs}\left(c_1,(Q, IS)\right) \\
&= \mathbb{E}[(a_1 \mid IS, Q, do(c_1)) - (a_1 \mid IS, Q, c_1)].
\end{aligned} \tag{5}
$$

To ensure uniformity of CauCoT, we apply LLMs to finish $\mathbb{E}(\cdot)$ by scoreing, prompt examples are as follow:

> **The prompt of γ_{fs}**
>
> Now that we try to answer the question Q step by step and c_1 is the first step. Now you need to carefully evaluate the impact of c_1 to answer Q correctly. Please the impact and the full score is 100.

> **The prompt of γ_a**
>
> Now that we try to answer the question Q step by step through reasoning path c_i. Now you need to carefully evaluate the impact of c_i^{pa} to answer Q based on c_i is final step. Please the impact and the full score is 100.

> **The prompt of γ_l**
>
> Now that we try to answer the question Q step by step through reasoning path c_i. Now you need to carefully evaluate the impact of c_i^{pa} to generate c_i. If the full score is 100, you need to score the size of the impact.

To complete the causal inference of steps , we still need a criterion for determining whether the causality exists. Therefore, we define σ as the causalized confidence degree and provide the following definition:

Definition 1. *We say there exists a causal relation between c_i and c_i^{pa} if*

$$
\gamma_{CoT}(c_i^{pa}, (Q, IS)) < \sigma. \tag{6}
$$

To ensure the flexibility of causal inference of steps we redefine (4) as :

$$
\begin{aligned}
&\gamma_{CoT}(c_i^{pa}, (Q, IS)) \\
&= \alpha \gamma_a(c_i^{pa}, (Q, IS)) + \beta \gamma_l(c_i^{pa}, (Q, IS)),
\end{aligned} \tag{7}
$$

where α and β are hyperparameters (the setting of α and β will be further discussed in the experiment.), with $\alpha + \beta = 1$. These parameters can be used to adjust the emphasis on different aspects of the causal relationship, such as prioritizing the reasoning logic or the correct answer. It is important to note that σ is not a strictly fixed value; it can be adjusted based on the specific scenario. For example, in mathematical reasoning, where strict logic between steps is emphasized, a higher value for σ should be used. In contrast, for knowledge reasoning, which may rely more on continuous causality between steps, a relatively lower value for σ may be appropriate. In summary, σ can be adjusted according to the type of task to align with the requirements of different reasoning tasks. The settings for σ will be discussed further in Appendix A.3, A.4.

4.2 Causalizing algorithm of steps

In the steps of causal inference, we have completed the causal inference in CoT and obtained a set of CoT including steps without the causality, wthih defined as $\sigma = [\dot{C}_1, \dot{C}_2, \ldots, \dot{C}_n]$. Then we need to causalize those CoT to

make every steps has causal relation. We define causal relation as follow:

Definition 2. *We say C has correct causality to answer Q when any $c_i \in C$ has causal relation with c_i^{pa}, and c_i is faithful to answer the qusetion Q.*

Inspired by [Han, 2024], where role-playing queries generate more targeted and unique responses, we apply a two-step algorithm consisting of role-playing causal queries, which includes the causalizing process and the refinement process, to establish the causal relationships in all steps.

Causalizing process

The definition of the roles in the causalizing Process is as follows: LLMs act as the [agent], attempting to find the correct reasoning steps. The agent depends on the [scenario] Q, such as an economist in an economic Q or a medical professional in a medical Q. c_i^{pa} represents the [treatment], $[c_i]$ is the [outcome], and the current incorrect steps $\dot{c}_i \in \dot{C}_i$ is [preconditions]. Therefore, our role-playing causal queries ask the LLM to assume the role of the [agent]and predict the [outcome] after the [treatment], given the [preconditions] in a hypothetical [scenario]. The [treatment] and [preconditions] provided assist the LLM in better capturing the correct causal relationships, as we are presenting incorrect answers based on the [preconditions]. An example of the prompt used in the causalizing process is as follows:

> **Causalizing Process**
>
> You are [agent] who is answering question Q and trying to think about the problem step by step, where c_i^{pa} and $\dot{c}_i$ should have strong causal relation to answer the question correctly. Since $\dot{c}_i$ is wrong step, now generate c_i that can meet the strong causal relationship with the previous step c_i^{pa}, make $c_i^{pa} \to c_i$ is correct reasoning path to correctly answer Q. Please list $[c_i]$ chains with strong credibility and explain your result.

Refine process

Applying refining, the large language model conducts self-reviews during the process of generating answers to verify the correctness and consistency of its responses. In the our refine process, we prompt the LLM to rethink its previous steps and select the most accurate and faithful response to answer the question Q,

> **Refine process**
>
> You are [agent] who is answering question Q, as previously described. Among the $[c_i]$ listed above, choose chain c_i that are most likely to have strong reasoning to answering question Q. For the chosen chain, explain the reasoning.

CauCoT algorithm: Loop until Causalized

Although we introduce a refine process to enhance the quality and efficiency of causalizing algorithm of steps, making all steps of CoT causalized is still a challenging task. Therefore, we further propose the CauCoT algorithm to ensure the reliability of the causalizing of steps.

Algorithm 1 CauCoT Algorithm

Input: CoT C, prompt IS, Question Q, causalized confidence degree σ, and large language model p_{LM}.
Output: Causalized C

1: Initialize large language model p_{LM} with prompt IS.
2: **for** each c_i in C **do**
3:      **if** $i = 1$ **then**
4:          Finish γ_{fs} inference by (5)
5:          **while** $\gamma_{fs}(c_1, (Q, IS)) < \epsilon$ **do**
6:              Do causalizing algorithm of steps with $\dot{c}_1$
7:          **end while**
8:          **continue**
9:      **end if**
10:      Finish $\gamma_{\text{CoT}}(c_i^{pa}, (Q, IS))$ by (4)
11:      **while** $\gamma_{\text{CoT}}(c_i^{pa}, (Q, IS)) < \sigma$ **do**
12:          Do causalizing algorithm of steps with $\dot{c}_i$
13:      **end while**
14: **end for**

---

5 Experimental results and analyses

We evaluate the causalization performance of CauCoT on the Polluted CoT dataset using both open-source and closed-source LLMs. First, we analyze four common causal errors in CoT that CauCoT addresses. Subsequently, through experiments about evaluation of reasoning and causality, we prove that CauCoT demonstrate that the causal errors commonly in steps are effectively corrected and the reasoning ability of LLMs is significantly improved.

5.1 Datasets

We select the PROCESSBENCH (PB) dataset [Zheng *et al.*, 2024]. It consists of 3,400 tests cases, primarily focused on competition and Olympiad-level math problems. Each test case contains a step-by-step solution with error location annotated by human experts, and half of the data is causalized, while the other half contains causal errors.

Specifically, it contains sample questions from the following four datasets: **GSM8K** [Cobbe *et al.*, 2021],**Math** [Hendrycks *et al.*, 2021],**OlympiadBench** [He *et al.*, 2024] and **Omni-MATH** [Gao *et al.*, 2024] (Details will be show in Appendix A.1).

5.2 Common causal errors in CoT

Through experiments, we verified that CauCoT can effectively correct causal errors in Polluted CoT. We summarize the potential causal errors in CoT as follows (The detailed analysis and example wii be showed in Appendix A.5):

Causality measure error [Chwialkowski *et al.*, 2014; Scheines and Ramsey, 2017].

Causality measure error refers to the inaccuracy in causal inference caused by the incorrect measurement or transmission of causal variables within the CoT.

Collider error [Schneider, 2020; Holmberg and Andersen, 2022].

Collider error is an error caused by the inability to accurately measure the impact of two variables CoT on answering the question (the next step is the collision point).

Causal sensitivity error [Cinelli *et al.*, 2019].

Causal sensitivity error refers to the failure to accurately measure the impact of two causal variables in CoT on an-

Table 1: EM results of reasoning QA. The first row of the table indicates the content of each column. The first column lists the datasets, the second column shows the LLMs used for testing (including different versions of Qwen, Deepseek, and Llama). Columns three to five display the EM results of the baseline models we compared, while the sixth column presents the results of our method, CauCoT. We highlight the top three results in red (first), blue (second), and purple (third).

Dataset	Model	0-shot	CoT	PB	CauCoT
GSM8K	Qwen2.5-3b	0.485	0.791	0.415	0.848
	Qwen2.5-7b	0.517	0.854	0.425	0.871
	Llama3-8b	0.499	0.845	0.433	0.864
	Qwen2.5-72b	0.789	0.915	0.500	0.943
	Deepseek-v3-37B	0.850	0.915	0.500	0.950
Math	Qwen2.5-3b	0.426	0.472	0.376	0.638
	Qwen2.5-7b	0.498	0.576	0.448	0.677
	Llama3-8b	0.443	0.519	0.462	0.653
	Qwen2.5-72b	0.570	0.820	0.499	0.882
	Deepseek-v3-37B	0.616	0.902	0.500	0.935
Olympiadbench	Qwen2.5-3b	0.062	0.106	0.316	0.388
	Qwen2.5-7b	0.076	0.153	0.390	0.512
	Llama3-8b	0.069	0.133	0.356	0.442
	Qwen2.5-72b	0.115	0.197	0.499	0.634
	Deepseek-v3-37B	0.179	0.254	0.500	0.665
Omnimath	Qwen2.5-3b	0.142	0.165	0.350	0.452
	Qwen2.5-7b	0.181	0.242	0.388	0.556
	Llama3-8b	0.161	0.251	0.352	0.441
	Qwen2.5-72b	0.217	0.362	0.498	0.675
	Deepseek-v3-37B	0.309	0.381	0.500	0.722

swering the question. It occurs when the next step fails to account for the sensitivity of the variables' relationship, leading to an incorrect conclusion.

Mediation error [Pearl, 2014].

Mediation error occurs when CoT fails to correctly identify the mediator variable, which is essential in explaining the causal relationship between other variables. This error arises when CoT ignores the role of the mediator, leading to a misunderstanding of establishing a direct causal relationship between two variables, resulting in an incorrect conclusion.

5.3 Main results

In this section, we evaluate the performance of CauCoT in reasoning questions. We analyze the experiments from two perspectives: the effectiveness of regular answers and the causality of the CoT. This dual analysis allows us to assess both the accuracy of the answers and the degree to which CauCoT ensures correct causal relationships in the reasoning process.

Evaluations of reasoning ability.

For baselines, we compare our method with the conventional CoT prompting method [Wei *et al.*, 2022] and PROCESS-BENCH (PB) [Zheng *et al.*, 2024], which help demonstrate the superiority of CauCoT and its effectiveness in explaining CoT's reasoning ability from a causal perspective. Additionally, a 0-shot QA method is used as a control group. We employ Exact Match (EM) as our evaluation metric, and all tasks are evaluated in a 0-shot setting.

Comparison with Baselines.

The reasoning QA results are shown in Table (1). As expected, CauCoT outperforms all other methods across all datasets and open-source large models. Notably, on more complex logical problem datasets, such as Olympiad-Bench and Omni-MATH, the improvement with CauCoT is more pronounced compared to relatively simpler datasets like GSM8K and Math. Another point to highlight is that, in datasets like GSM8K and Math, CoT, or even 0-shot, can outperform PB. However, when faced with more complex datasets like OlympiadBench and Omni-MATH, PB actually yields better performance. This is because PB contains a subset of data that has already been causalized, which aids LLMs in providing correct answers. In contrast, conventional CoT methods do not ensure causality, which can result in less than half of the answers being correct. CauCoT, however, shows significant improvement, confirming that causality is key to ensuring CoT can perform reasoning correctly.

Evaluations of causality

For the causality of CoT, we compare CauCoT with CoT and PB by analyzing the changes in causal effects between each step. Across four datasets, we report the heterogeneous effect (HE) and the factual average treatment effect (ATE). The heterogeneous effect is defined as: ($HE = \sqrt{n^{-1} \sum_{i=1}^{n} \left((\hat{a}_i - a_i \mid do(c_i)) \right)^2}$). Additionally, we report the factual average treatment effect ($\hat{ATE}$) to evaluate the overall impact of the intervention on CoT reasoning.

Comparison with Baselines.

As shown in Figure 3, we analyze the causality between steps. CauCoT achieves the strongest causality across all datasets, and its inference EM is also the highest, confirming that CauCoT successfully causalizes CoT. When the causality between CoT steps is stronger than that in PB, its inference performance surpasses PB. Conversely, when the

causality is weaker than that in PB, its inference performance lags behind CoT. It further reinforces the fact that causality is a critical determinant of CoT's reasoning ability.
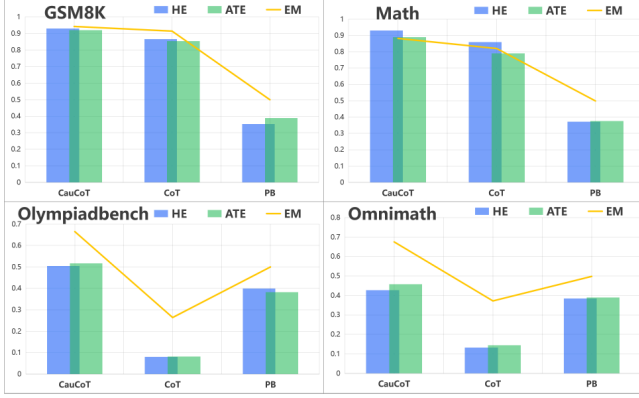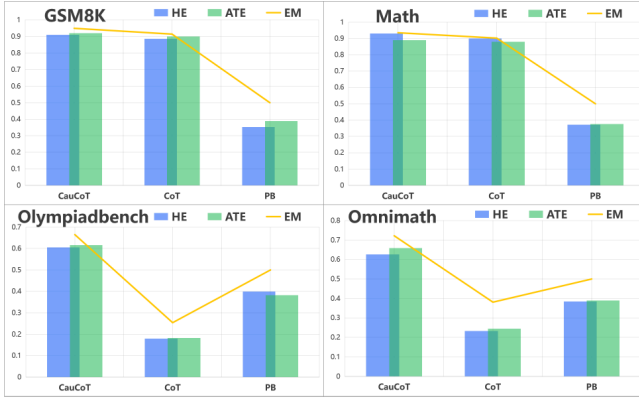

Figure 3: Causalized Evaluation on Qwen2.5-72B

Figure 4: Causalized Evaluation on Deepseek-v3-37B

Hyperparamter experiments

Here, we discuss the setting of hyperparameters α and β based on Qwen2.5-72b. As shown in the Figure 5, when both factors are balanced, CauCoT achieves the strongest causality. It is noteworthy that when $\alpha = 0, \beta = 1$, the performance drop of CauCoT is much larger than when $\alpha = 1, \beta = 0$. This indicates that in order to causalize CoT, causality should not only be considered from a logical perspective (γ_l), but the causal logic with answers (γ_a). This demonstrates that, to unveil the mechanisms of CoT and further enhance the performance of large models' reasoning, **correctness and understandability of steps are equally important.**

5.4 Discussions of experimental results

We make the following discussion based on experimental results:

1.Based on the evaluations of reasoning ability and causality, the stronger the causality between reasoning steps, the stronger the reasoning ability of CoT is. This proves that **our proposed SCM successfully unveils the mechanisms of CoT reasoning from a causal perspective, making the reasoning of LLMs interpretable.**

2.CauCoT successfully to correct the causal errors in the polluted dataset, and CauCoT achieves the strongest reasoning ability and causality across all datasets and LLMs. Thus, **we successfully achieves an algorithms that causalize all steps of CoT to make them correct and understandable.**

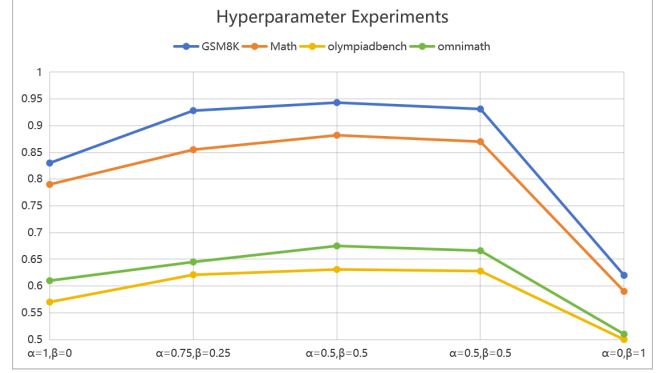We will also provide further analysis of the experiment results in the Appendix.

Figure 5: Hyperparamter experiments

6 Conclusion

We unveil and causalize CoT from a causal perspective, ensuring LLMs' reasoning both correct and understandable. We model the causality of CoT via SCM, illustrating that steps' reasoning ability stems from causality. To test whether CoT has been causalized, we proposed causal inference of steps from both logic and answer aspects. we design a role-playing causal query algorithm to causalize steps without causality, resulting on our method CauCoT. Experimental results on both open-source and closed-source LLMs demonstrate that CauCoT effectively correct the causal error in steps and improve the reasoning ability of LLMs. In summary, our proposed SCM provides a new paradigm for modeling LLMs' reasoning from a causal perspective. CauCoT achieves an unprecedented causalizing algorithms of CoT, offering a causal solution for enabling correct and understandable LLMs' reasoning. Our work also presents the causality community with an innovative methodology for causally analyzing the reasoning mechanisms of CoT or LLMs.

References

[Bhattacharjee *et al.*, 2024] Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. Towards llm-guided causal explainability for black-box text classifiers. In *AAAI*, 2024.

[Burns and Wieth, 2004] Bruce D Burns and Mareike Wieth. The collider principle in causal reasoning: why the monty hall dilemma is so hard. *Journal of Experimental Psychology: General*, 133(3):434, 2004.

[Chen *et al.*, 2022] Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. Disco: Distilling counterfactuals with large language models. *arXiv:2212.10534*, 2022.

[Chen *et al.*, 2024] Qiguang Chen, Libo Qin, Jiaqi Wang, Jinxuan Zhou, and Wanxiang Che. Unlocking the ca-

pabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. In *NeurIPS*, 2024.

[Chu *et al.*, 2024] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

[Chwialkowski *et al.*, 2014] Kacper P Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. *Advances in neural information processing systems*, 27, 2014.

[Cinelli *et al.*, 2019] Carlos Cinelli, Daniel Kumor, Bryant Chen, Judea Pearl, and Elias Bareinboim. Sensitivity analysis of linear structural causal models. In *ICML*, 2019.

[Cobbe *et al.*, 2021] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

[Díaz and van der Laan, 2013] Iván Díaz and Mark J van der Laan. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The international journal of biostatistics*, 9(2):149–160, 2013.

[Feng *et al.*, 2024] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. In *NeurIPS*, 2024.

[Gao *et al.*, 2024] Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv:2410.07985*, 2024.

[Han, 2024] Sukjin Han. Mining causality: Ai-assisted search for instrumental variables. *arXiv:2409.14202*, 2024.

[He *et al.*, 2024] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv:2402.14008*, 2024.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv:2103.03874*, 2021.

[Holmberg and Andersen, 2022] Mathias J Holmberg and Lars W Andersen. Collider bias. *Jama*, 327(13):1282–1283, 2022.

[Hu *et al.*, 2023] Mengkang Hu, Yao Mu, Xinmiao Yu, Mingyu Ding, Shiguang Wu, Wenqi Shao, Qiguang Chen, Bin Wang, Yu Qiao, and Ping Luo. Tree-planner: Efficient close-loop task planning with large language models. *arXiv:2310.08582*, 2023.

[Imai *et al.*, 2010] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010.

[Jin *et al.*, 2023a] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: assessing causal reasoning in language models. In *NeurIPS*, 2023.

[Jin *et al.*, 2023b] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? *arXiv:2306.05836*, 2023.

[Kaddour *et al.*, 2022] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv:2206.15475*, 2022.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv:2205.11916*, 2022.

[Kıcıman *et al.*, 2023] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv:2305.00050*, 2023.

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

[Li *et al.*, 2023] Baohong Li, Anpeng Wu, Ruoxuan Xiong, and Kun Kuang. Two-stage shadow inclusion estimation: An iv approach for causal inference under latent confounding and collider bias. In *ICML*, 2023.

[Li *et al.*, 2024] Haoxuan Li, Chunyuan Zheng, Yanghao Xiao, Peng Wu, Zhi Geng, Xu Chen, and Peng Cui. Debiased collaborative filtering with kernel-based causal balancing. *arXiv:2404.19596*, 2024.

[Liu and Kuang, 2023] Chenxi Liu and Kun Kuang. Causal structure learning for latent intervened non-stationary data. In *ICML*, 2023.

[Liu *et al.*, 2021] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv:2103.10385*, 2021.

[Madaan *et al.*, 2023] Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. What makes chain-of-thought prompting effective? a counterfactual study. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[Papana *et al.*, 2011] A. Papana, D. Kugiumtzis, and P. G. Larsson. Reducing the bias of causality measures. *Physical Review E*, 83(3), 2011.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Pearl, 2012] Judea Pearl. On measurement bias in causal inference. *arXiv:1203.3504*, 2012.

[Pearl, 2014] Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4), 2014.

[Qi *et al.*, 2023] Zhengling Qi, Rui Miao, and Xiaoke Zhang. Proximal learning for individualized treatment regimes under unmeasured confounding. *Journal of the American Statistical Association*, 2023.

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

[Rubin, 1980] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 1980.

[Scheines and Ramsey, 2017] Richard Scheines and Joseph Ramsey. Measurement error and causal discovery. In *CEUR workshop proceedings*, 2017.

[Schneider, 2020] Eric B Schneider. Collider bias in economic history research. *Explorations in Economic History*, 78, 2020.

[Singh *et al.*, 2020] Rahul Singh, Liyuan Xu, and Arthur Gretton. Kernel methods for causal functions: Dose, heterogeneous, and incremental response curves. *arXiv:2010.04855*, 2020.

[Sprague *et al.*, 2024] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv:2409.12183*, 2024.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, H Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *In NeurIPS*, 2022.

[Wu *et al.*, 2023a] Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. Stable estimation of heterogeneous treatment effects. In *ICML*, 2023.

[Wu *et al.*, 2023b] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023.

[Wu *et al.*, 2024] Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. Causality for large language models. *arXiv:2410.15319*, 2024.

[Xiang *et al.*, 2025] Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, et al. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-though. *arXiv:2501.04682*, 2025.

[Xie *et al.*, 2020] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *arXiv:2010.04917*, 2020.

[Xie *et al.*, 2024] Feng Xie, Biwei Huang, Zhengming Chen, Ruichu Cai, Clark Glymour, Zhi Geng, and Kun Zhang. Generalized independent noise condition for estimating causal structure with latent variables. *Journal of Machine Learning Research*, 25:1–61, 2024.

[Yang *et al.*, 2024] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv:2412.15115*, 2024.

[Yao *et al.*, 2021] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *In TKDD*, 2021.

[Zečević *et al.*, 2023] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *arXiv:2308.13067*, 2023.

[Zeng *et al.*, 2024] Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Zhang *et al.*, 2024a] Congzhi Zhang, Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. Causal prompting: Debiasing large language model prompting based on front-door adjustment. *arXiv:2403.02738*, 2024.

[Zhang *et al.*, 2024b] Shengyu Zhang, Ziqi Jiang, Jiangchao Yao, Fuli Feng, Kun Kuang, Zhou Zhao, Shuo Li, Hongxia Yang, Tat-Seng Chua, and Fei Wu. Causal distillation for alleviating performance heterogeneity in recommender systems, 2024.

[Zheng *et al.*, 2024] Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning. *arXiv:2412.06559*, 2024.

[Zuo *et al.*, 2024] Aoqi Zuo, Yiqing Li, Susan Wei, and Mingming Gong. Interventional fairness on partially known causal graphs: A constrained optimization approach. *arXiv:2401.10632*, 2024.

A   Appendix

A.1   Datasets

We select the PROCESSBENCH (PB) dataset [Zheng *et al.*, 2024]. It consists of 3,400 tests cases, primarily focused on competition and Olympiad-level math problems. Each test case contains a step-by-step solution with error location annotated by human experts, and half of the data is causalized, while the other half contains causal errors.

Specifically, it contains sample questions from the following four datasets:

GSM8K [Cobbe *et al.*, 2021] contains high quality linguistically diverse grade school math word problems.

Math [Hendrycks *et al.*, 2021] is a challenging competition math problems dataset. Each problem requires a complete step-by-step solution to arrive at the correct answer.

OlympiadBench [He *et al.*, 2024] is an Olympiad-level bilingual multimodal science benchmark that contains Olympiad-level math and physics competition problems, including the Chinese college entrance examination. Each problem requires expert-level annotations to complete step-by-step reasoning. We focus OlympiadBench's physics part in our experiment.

Omni-MATH [Gao *et al.*, 2024] is a mathematics-focused, comprehensive and challenging benchmark specifically designed to assess LLMs' mathematical reasoning ability at the Olympiad level. It is rigorously manually annotated. The questions are carefully divided into more than 33 sub-areas covering more than 10 different difficulty levels.

A.2   LLMs CoT reasoning: Causal view analysis

Combining the results in Table 1, Figure 3 and Figure 4, we make the following analyses:

1.More complex problems require causal reasoning: As previously mentioned, the causal relationship between PB and correct answers is stronger than that of CoT in relatively complex datasets. A similar phenomenon is observed in Table 1. It indicates that problems requiring more complex reasoning demand reasoning paths with correct causal relationships for accurate answers. The greater improvements achieved by CauCoT over CoT on complex datasets compared to simpler datasets further confirm this conclusion.

2.Stronger language models have a greater ability to capture causal information: On the PB dataset, Qwen2.5-72b, and Deepseek-v3-37B achieves nearly 0.5 accuracy, indicating that the model can fully capture and understand half of the correct causal information in the dataset. This demonstrates from the perspective of LLMs capability that causality is a necessary factor in determining whether the reasoning is correct.

A.3 σ evaluation

we analyze the numerical settings of the σ. For the implementation of the causal base model, we consistently choose Qianwen2.5-72b.

Table 2: σ Evaluation. The first column lists the datasets used for evaluation. The second column shows the values of σ set in the experiments, and the last column represents the proportion of CoT that successfully causalized.

Dataset	σ values	Percents of successfully causalizing
GSM8K	50	100%
	75	100%
	100	100%
Math	50	100%
	75	100%
	100	96%
Olympiad	50	100%
	75	65%
	100	54%
Omnimath	50	100%
	75	76 %
	100	69 %

As shown in the Table 2, when the dataset is relatively complex, a higher σ value makes updates more difficult to complete. This also highlights the necessity of setting σ appropriately, allowing CauCoT to adjust the settings according to different scenarios to ensure feasibility.

A.4   Ablation experiment

We design baselines as :

CauCoT-WOutRD: Removing γ_{td}, skip the step of validating the causal relationship between Q and C.

CauCoT-WOutRF: Removing Refine process, do not apply exhaustive thinking.

CauCoT-WutLoop: Removing loop, trust the result of a single causalize attempt.

Due to the minimal performance difference of CauCoT on GSM8K, we do not consider its performance on this dataset in the ablation experiments.
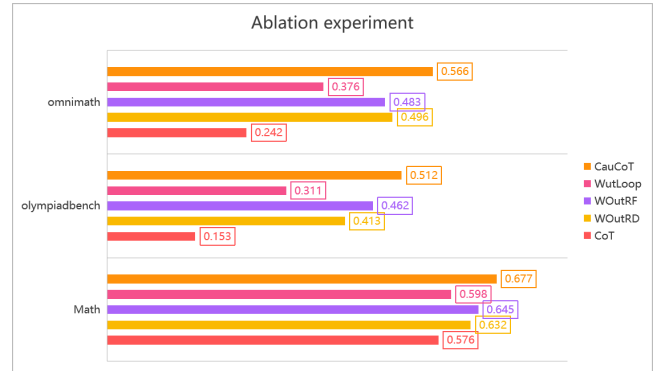

Figure 6: Ablation experiment

As shown in the Figure 6, the performance of the three baselines we proposed is inferior to that of CauCoT. The accuracy difference between WOutRD and CauCoT is equal to the percentage of errors in the first step of PB. This demonstrates that when there is an incorrect causal relationship be-

tween Q and C, subsequent causalize steps are also difficult to complete. This confirms the necessity of setting the equation 5. The performance of WOutRF proves the importance of slow thinking, which becomes even more apparent when dealing with relatively complex datasets (Omnimath and Olypiadbench). The performance of CauCoT-WutLoop demonstrates that existing LLMs still have an insufficient understanding of causality.

A.5   Examples of causal errors in CoT

Causality measure error [Papana *et al.*, 2011; Pearl, 2012; Chwialkowski *et al.*, 2014; Scheines and Ramsey, 2017].

Causality measure error refers to the inaccuracy in causal inference caused by the incorrect measurement or transmission of causal variables within the CoT. For instance, as shown in the first example in Table 3, this is a typical mathematical calculation problem. When an error occurs in calculation steps, incorrect causal variable data will be passed on to the next step, ultimately leading to an incorrect answer to the question.

Collider error. [Burns and Wieth, 2004; Schneider, 2020; Holmberg and Andersen, 2022].

Collider error is an error caused by the inability to accurately measure the impact of two variables CoT on answering the question (the next step is the collision point). Like the second example in Table 3, it is a problem that combines calculation and logical deduction. When both causal variables—division calculation and logical deduction based on the remainder—appear simultaneously, the next CoT step fails to accurately capture information from both, leading to an error in the calculation.

Causal sensitivity error [Imai *et al.*, 2010; Díaz and van der Laan, 2013; Cinelli *et al.*, 2019].

Causal sensitivity error refers to the failure to accurately measure the impact of two causal variables in CoT on answering the question. It occurs when the next step fails to account for the sensitivity of the variables' relationship, leading to an incorrect conclusion. For example, in the second case in Table 3, it is a problem that involves both calculation and logical deduction. When both causal variables—such as division calculation and logical deduction based on the remainder—appear simultaneously, the next CoT step fails to accurately capture the influence of both, resulting in an error in the calculation.

Mediation error [Pearl, 2014].

Mediation error occurs when CoT fails to correctly identify the mediator variable, which is essential in explaining the causal relationship between other variables. This error arises when CoT ignores the role of the mediator, leading to a misunderstanding of the causal pathway. For example, in the fourth case in Table 3, CoT overlooks the iron pickaxe as a mediating variable and mistakenly establishes a direct causal relationship between two variables, resulting in an incorrect conclusion.

Moreover, we have emphasized the necessity of verifying the causal relationship between the first step and the Q, as discussed in formula (5). For example, in question 3 in Table 3, if the causal relationship between the first step and Q is incorrect, it negatively impacts subsequent steps, leading to a failure in the reasoning process. This issue will also be further explored in Appendix 5.3.

B   Technical report

B.1   Details of LLMs

We updated the reasoning chains in the ProcessBench dataset (https://huggingface.co/datasets/Qwen/ProcessBench) using the Qwen2.5-72B model (https://huggingface.co/Qwen/Qwen2.5-72B) via the Transformer Library (https://huggingface.co/docs/transformers/en/index). The hyperparameters for generation remained consistent across all prompt methods. During the update process, we utilized a pipeline as a high-level helper, and set the task type to "text-generation" while keeping other hyperparameters at their default values.

Additionally, we compared the question-answering performance before and after updates using large models: Qwen2.5-3B, Qwen2.5-7B, Llama3-8B, Qwen2.5-72B, and Deepseek-v3-37B. For the Qwen2.5-3B, Qwen2.5-7B, Llama3-8B, and Qwen2.5-72B models, we set the voting count to 8, based on prior work [Zheng *et al.*, 2024]. The final output was determined by majority voting.

Link of LLMs intructions:
Qwen2.5-3B:
(https://huggingface.co/Qwen/Qwen2.5-3B).
Qwen2.5-7B:
(https://huggingface.co/Qwen/Qwen2.5-7B).
Llama3-8B:
(https://huggingface.co/meta-llama/Meta-Llama-3-8B).
Qwen2.5-72B:
(https://huggingface.co/Qwen/Qwen2.5-72B).
Deepseek-v3-37B:
(https://github.com/deepseek-ai/DeepSeek-V3).

More specifically, the generation hyperparameters remained consistent across all prompt methods, as shown in the table 4:

Due to the large size of the Deepseek-v3-37B model, we use the OpenAI Python API library and an API key obtained from Deepseek's official request to call it. The generation hyperparameters are shown in the Table 5:

Table 5: Hyperparameters of Deepseek-v3-37B in experiments

Hyperparameter	Value
stream	False
model	deepseek-chat

Table 3: The table below provides examples of common causal errors in CoT that were corrected by CauCoT. The first column indicates the dataset to which the example belongs. The second column contains the content of the corresponding question. The third column specifies the step where the causal error occurred. The fourth column details the specific content of the error in that step, with the exact errors highlighted in red. The fifth column presents the corrected content by CauCoT, with the corrected parts marked in green.

Dataset	Question	Steps	Setps without causality	Causalized CoT
GSM8K	A company sold 4000 gallons of milk in jars to Mr. Marcellus' store at the cost of $3.5 per gallon. However, Mr. Marcellus later realized 2/5 of the amount of milk he purchased had passed the expiry date and could not be sold. He returned the sour milk to the company and ordered a refund. Calculate how much he got in refunds.	Step 3	"...Total gallons purchased = 2/5 * 4000 = 800 gallons.", "Now, we need to calculate the cost of the expired milk that needs to be refunded: Cost of expired milk = Cost per gallon * Amount of expired milk = $3.5 ∗ 800 = $2,800.", "Therefore, Mr. Marcellus got a refund of $2,800 for the unsellable milk. [The answer is $2800."✗]	"...Total gallons purchased by the fraction of expired milk: 4000 gallons ∗ (2/5) = 1600 gallons.", "Now, we need to calculate the cost of the expired milk that needs to be refunded: Cost of expired milk = Cost per gallon * Amount of expired milk = $3.5 ∗ 1600 = $5,600.", "Therefore, Mr. Marcellus got a refund of $5,600 for the unsellable milk. [The answer is $5,600."✓]
Math	Carlos Montado was born on Saturday, November 9, 2002. On what day of the week will Carlos be 706 days old? (Spell out the entire day of the week.	Step 2	"... Divide 706 by 7 : $\frac{706}{7} = 101$ with a remainder of 1. This means that 706 days from his birth date is equivalent to adding one day to his birth day.", "Since Carlos was born on a Saturday, we add 1 day to Saturday: Saturday + 1 day = Sunday.", "Therefore, the day of the week Carlos will be 706 days old is: [Sunday."✗]	"...Since the result is 100 and reminder is 6, reduce 1 day to Saturday, the day Carlos was born. This will give us the day of the week he will be 706 days old.", "As Carlos was born on a Saturday, we reduce 1 day to Saturday: Saturday - 1 day = Firday.", "Therefore, the day of the week Carlos will be 706 days old is: [Firday."✓]
Olympiad	Bobby, Peter, Greg, Cindy, Jan, and Marcia line up for ice cream. In an acceptable lineup, Greg is ahead of Peter, Peter is ahead of Bobby, Marcia is ahead of Jan, and Jan is ahead of Cindy. For example, the lineup with Greg in front, followed by Peter, Marcia, Jan, Cindy, and Bobby, in that order, is an acceptable lineup. Compute the number of acceptable lineups.",	Step 1	"To determine the number of acceptable lineups for Bobby, Peter, Greg, Cindy, Jan, and Marcia, we need to consider the constraints given: First, Greg is ahead of Peter. Second, Peter is ahead of Bobby. Third, Marcia is ahead of Jan. Fourth, Jan is ahead of Cindy. These constraints imply that the relative positions of Greg, Peter, Bobby, Marcia, Jan, and Cindy must be in the following order: Greg, Peter, Bobby, Marcia, Jan, Cindy."..."Thus, the number of acceptable lineups is [35"✗])	"To determine the number of acceptable lineups for Bobby, Peter, Greg, Cindy, Jan, and Marcia, we need to consider the constraints given: First, Greg is ahead of Peter. Second, Peter is ahead of Bobby. Third, Marcia is ahead of Jan. Fourth, Jan is ahead of Cindy. These constraints imply that the relative positions of Greg, Peter, Bobby, Marcia, Jan, and Cindy must be in the following two order: Greg, Peter, Bobby and Marcia, Jan, Cindy."..." The number of ways to select 3 positions from 6 is given by the combination formula acceptable lineups is [20."✓]
GSM8K	One dwarf can mine 12 pounds of ore per day with his bare hands. He can mine twice as much with an iron pickaxe and 50% more with a steel pickaxe than with an iron pickaxe. How many pounds of ore can 40 dwarves with steel pickaxes mine in a month with 30 days?	Step 2	"... First, we calculate the amount of ore one dwarf can mine with a steel pickaxe. We know that one dwarf can mine 50% more with a steel pickaxe than with an iron pickaxe" ..." So, the total amount of ore mined by 40 dwarves with steel pickaxes in a month is:[21600."✗]	"...First, we calculate the amount of ore one dwarf can mine with a iron pickaxe. We know that one dwarf can mine twice as much with an with an iron pickaxe..."So, the total amount of ore mined by 40 dwarves with steel pickaxes in a month is:[43,200. "✓]"

Table 4: Hyperparameter Settings for Different Models

Hyperparameter	Qwen2.5-3B	Qwen2.5-7B	Llama3-8B	Qwen2.5-72B
temperature	1	1	1	1
top p	0.9	0.9	0.9	0.9
trust remote code	True	True	True	True
times of votes	8	8	8	8
max tokens	32768	32768	32768	32768
tensor parallel size	4	4	4	8