From Causal Parrots to Causal Prophets? Towards Sound Causal Reasoning with Large Language Models

Rahul B. Shrestha[*] and **Simon Malberg**[*] and **Georg Groh**
School of Computation, Information and Technology
Technical University of Munich, Germany
{rahul.shrestha, simon.malberg}@tum.de, grohg@cit.tum.de
[*]These authors contributed equally to this work

Abstract

Causal reasoning is a fundamental property of human and machine intelligence. While *large language models* (LLMs) excel in many natural language tasks, their ability to infer causal relationships beyond memorized associations is debated. This study systematically evaluates recent LLMs' causal reasoning across three levels of Pearl's *Ladder of Causation*—associational, interventional, and counterfactual—as well as commonsensical, anti-commonsensical, and nonsensical causal structures using the CLADDER dataset. We further explore the effectiveness of prompting techniques, including *chain of thought* (CoT), *self-consistency* (SC), and *causal chain of thought* (CAUSALCoT), in enhancing causal reasoning, and propose two new techniques *causal tree of thoughts* (CAUSALToT) and *causal program of thoughts* (CAUSALPoT). While larger models tend to outperform smaller ones and are generally more robust against perturbations, our results indicate that all tested LLMs still have difficulties, especially with counterfactual reasoning. However, our CAUSALToT and CAUSALPoT significantly improve performance over existing prompting techniques, suggesting that hybrid approaches combining LLMs with formal reasoning frameworks can mitigate these limitations. Our findings contribute to understanding LLMs' reasoning capacities and outline promising strategies for improving their ability to reason causally as humans would. We release our code and data[1].

1 Introduction

Causal reasoning, the ability to infer cause-and-effect relationships, is a fundamental property of intelligence (Jin et al., 2023) in humans and machines alike. While LLMs have achieved significant progress in natural language processing (Radford et al., 2019; Zhao et al., 2024), their ability to perform genuine causal reasoning is debated. Existing studies indicate that models perform poorly when facing complex causal structures (Romanou et al., 2023), engaging in counterfactual reasoning (Wu et al., 2024b), or applying formal causal reasoning when commonsense rules do not apply (Jin et al., 2023). Some findings suggest that LLMs act more like "causal parrots", simply reciting causal knowledge from their training data rather than engaging in true causal inference (Zečević et al., 2023). Understanding and improving LLMs' causal reasoning capabilities remains critical for ensuring reliable LLM-supported decision-making, particularly in high-stakes domains such as healthcare, economics, or public policy.

This study aims to bridge the gap by systematically evaluating LLMs on causal reasoning tasks, addressing the following research questions:

1. **How well do LLMs perform in different disciplines of causal reasoning?** We evaluate a diverse set of models on causal reasoning tasks from the CLADDER dataset (Jin et al., 2023) spanning Pearl and Mackenzie's (2018) *Ladder of Causation*, including associational, interventional, and counterfactual reasoning. The latter two in particular constitute essential capabilities of humans and machines when planning and interacting with their environment.

2. **How well do LLMs generalize to causal reasoning tasks where they cannot rely on learned commonsense knowledge?** We systematically modify causal problems with anti-commonsensical and nonsensical perturbations and test LLMs' performance. This exposes how much LLMs rely on learned world knowledge when facing unknown causal reasoning challenges.

3. **Can prompting techniques and external tools enhance LLMs' causal reasoning?**

[1] Our code and data can be found here: https://github.com/rahulbshrestha/causal-reasoning

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 319–333
May 3-4, 2025 ©2025 Association for Computational Linguistics

> We evaluate *zero-shot chain of thought* (Kojima et al., 2022), *causal chain of thought* (Jin et al., 2023), and *chain of thought with self-consistency* (Wang et al., 2023). Finally, we introduce new causal variants of *tree of thoughts* (Yao et al., 2023) and *program of thoughts* (Chen et al., 2023) with an integration of the *DoWhy* causal inference library (Sharma and Kiciman, 2020) and demonstrate how these can elevate causal reasoning performance.

Our main contributions include (1) a comprehensive evaluation of recent LLMs' causal reasoning abilities, updating previous work, (2) an assessment of causal reasoning improvements coming through prompting techniques, and (3) two new prompting techniques CAUSALToT and CAUSALPoT to enhance LLMs' causal reasoning over previous baselines, borrowing ideas from formal causal reasoning techniques accessible to humans.

2 Background and Related Work

2.1 Ladder of Causation

The *Ladder of Causation*, introduced by Pearl and Mackenzie's (2018), structures causality into three levels, often referred to as a ladder with three rungs:

Rung 1 (Association): The lowest rung represents statistical associations and seeks to answer the question "What?" This level involves identifying patterns or correlations in the data without implying causation. For example, "What is the probability of lung cancer among smokers?"

Rung 2 (Intervention): This rung focuses on the effects of interventions, addressing the question "What if?" It examines the impact of actively altering a variable and observing its influence on other variables. For example, "If I stop smoking, will my risk of lung cancer decrease?"

Rung 3 (Counterfactual): This highest rung involves counterfactual reasoning, which answers the questions "Why?" or "What if I had acted differently?" This level entails imagining hypothetical scenarios based on observed data. For instance, "Given that I have lung cancer, if I had never smoked, would I still have developed the disease?"

2.2 Causal Reasoning with LLMs

LLMs have been proposed for use in several causal *natural language processing* (NLP) tasks, such as causal discovery (e.g., Kıcıman et al., 2024; Long et al., 2024), causal effect estimation (e.g., Jin et al., 2023), and counterfactual reasoning (e.g., Lewis and Mitchell, 2024). Liu et al. (2025) survey existing work on the interplay between LLMs and causal inference, separating approaches that use causal inference frameworks for LLMs and approaches that use LLMs for causal tasks. Similarly, Yu et al. (2025) provide a comprehensive overview of previous work using LLMs for causal reasoning, dividing into methods that use LLMs as the main reasoning engine and methods that use LLMs only as a helper to traditional methods.

While these works often find that LLMs outperform existing algorithms in these tasks, LLMs still seem to have difficulties with some more challenging tasks. Counterfactual reasoning on hypothetical and unusual causal structures in particular presents a challenge to LLMs, showing a degradation of reasoning performance compared to non-counterfactual settings (Lewis and Mitchell, 2024; Li et al., 2022; Wu et al., 2024c). Li et al. (2022) find that counterfactual reasoning of smaller language models seems to be largely driven by simple lexical triggers. They observe that only their largest model tested, GPT-3, was able to not only override real-world knowledge in counterfactual scenarios but also show somewhat greater sensitivity to more detailed linguistic cues.

Zečević et al. (2023) argue that LLMs merely behave like "causal parrots" simply reciting causal knowledge from their training data. This indicates that LLMs reason in ways different from what trained humans would do. Chi et al. (2024) discuss how autoregressive transformer-based LLMs are not inherently causal. LLMs are able to imitate causal reasoning only as long as similar causal knowledge is available in their training data (Zhang et al., 2023) or relevant domain-specific context and causal knowledge is provided (Cai et al., 2024).

2.3 Causal Reasoning Benchmarks

Multiple LLM-specific causal reasoning benchmarks and evaluation frameworks have emerged. Some notable benchmarks include CLADDER (Jin et al., 2023), CORR2CAUSE (Jin et al., 2024), CAUSALBENCH (Zhou et al., 2024; Wang, 2024), CRAB (Romanou et al., 2023), IfQA (Yu et al., 2023), and CRASS (Frohberg and Binder, 2022). For a comprehensive list of additional benchmarks, readers may refer to Liu et al. (2025).

2.4 Methodological Advances

While many works focus on measuring the causal reasoning abilities of LLMs, some proposals were made for how to turn LLMs into better causal reasoners. Wu et al. (2024a) explore how causality can improve LLMs at all stages of their lifecycle, looking at token embeddings, training, alignment, inference, and evaluation. Just like with other NLP tasks, fine-tuning the LLMs may improve their accuracy also on causal tasks, as shown by Cai et al. (2024) for causal discovery. Liu et al. (2023) even found that Code-LLMs seem to acquire better causal reasoning abilities than text-only LLMs and tend to be robust against format perturbations.

More advanced prompting techniques such as *chain of thought* (Wei et al., 2022b) were shown to improve reasoning performance of LLMs, although not with all LLMs and on all reasoning tasks (Wang and Shen, 2024; Yu et al., 2025). Jin et al. (2023) introduce a new causal variant of *chain of thought* called CAUSALCOT. On CLADDER, they demonstrate how a GPT-4 LLM achieves 62.03% accuracy without CAUSALCOT and 70.40% accuracy with CAUSALCOT.

Gendron et al. (2024) propose a new counterfactual causal inference framework (Counterfactual-CI) for causal discovery reaching an accuracy of 60.53% on CLADDER with a GPT-4o LLM. CARE-CA (Ashwani et al., 2024) attempts to improve LLM causal reasoning by enriching prompts with relevant causal concepts from a knowledge graph and counterfactual insights. The authors demonstrate CARE-CA's abilities on CLADDER, reporting a 63.0% accuracy with a T5 LLM, versus a 60.0% accuracy with T5 alone. Similar to CARE-CA, the G²-Reasoner (Chi et al., 2024) retrieves related general knowledge from a vector database and incorporates it in a goal-oriented prompt to guide the LLM in the reasoning process. While the authors do not evaluate the G²-Reasoner on CLADDER, they report performance improvements similar to CARE-CA on other datasets.

Yu et al. (2025) use Python scripts to solve 100 causal questions from CLADDER, achieving an accuracy of 76%. However, their method does not leverage the full potential of external causal inference tools, merely leveraging Python as a calculator for relatively simple computations. Their approach led to only a marginal improvement compared to the 75% accuracy achieved with CAUSALCOT. In contrast, our work integrates the external causal inference library *DoWhy* (Sharma and Kiciman, 2020) and evaluates performance on a larger, balanced dataset from CLADDER.

3 Methods

3.1 Dataset

Similar to several previous works, our experiments are based on CLADDER (Jin et al., 2023), a dataset that tests formal causal reasoning capabilities. The causal questions in the dataset are represented in natural language, yet the questions are grounded in symbolic logic and ground truth answers derived using an oracle causal inference engine (Pearl and Mackenzie, 2018).

Choice of CLADDER Arguably, formal causal reasoning, and CLADDER in particular, make an ideal test bench for LLMs' causal reasoning abilities. The necessity to formalize multi-step thought processes makes transparent whether the LLM identifies true causation rather than just correlations. Further, the symbolic grounding offers much potential to comprehensively evaluate the integration of external tools and reasoning frameworks. A review of causal reasoning benchmarks by Yang et al. (2024) referred to CLADDER as "the most advanced causal benchmark available currently, as it holistically tests the LLM's ability to synthesize several different components into a complex causal model, and then interprets the effects of interventions or changes within that model". CLADDER addresses key design issues identified in other benchmarks by (1) covering all three rungs of the *Ladder of Causation*, including interventional and counterfactual questions, (2) requiring multi-step causal reasoning rather than simple one-step answers, and (3) testing for reasoning rather than retrieval by including perturbed versions of queries.

Dataset Structure CLADDER questions test the ability to correctly plan and execute the estimation of a causal effect. Each question has a binary answer: *yes* or *no*. Questions cover all three rungs of the *Ladder of Causation*, span across nine distinct query types (e.g., marginal probability or average treatment effect), and represent one of three degrees of alignment with commonsense knowledge, namely commonsensical, anticommonsensical, and nonsensical.

Sampling and Perturbations For our experiments, we sampled 1,000 commonsensical questions from CLADDER, maintaining a distribution of question types similar to the original dataset (see Appendix A for details on the distribution). We excluded questions with the “backdoor adjustment” query type, as they do not require formal calculations and multi-step reasoning. Unfortunately, the publicly available CLadder dataset contains the anti-commonsensical and nonsensical counterparts to only some but not all of the commonsensical questions in our sample. Therefore, we created new anti-commonsensical and nonsensical perturbations of the 1,000 sampled commonsensical questions using GPT-4o:

- **Anti-commonsensical perturbations**: Given a causal relationship $X \rightarrow Y$ (e.g., *smoking → lung cancer*), we replaced Y with a randomly selected noun unrelated to X (e.g., *smoking → ice cream sales*).

- **Nonsensical perturbations**: Given a causal relationship $X \rightarrow Y$ (e.g., *smoking → lung cancer*), both X and Y were replaced with randomly-generated four-letter words (e.g., *xacx → msad*).

The CLadder paper applied similar perturbations, including anti-commonsensical and nonsensical variants, but used a fixed set of words for substitutions. In contrast, we let GPT-4o generate random words, introducing greater variability in the perturbations. To ensure grammatical and logical soundness, we manually verified all generated perturbations.

Details about the exact prompt used for the perturbations can be found in Appendix B. An example image illustrating the two perturbations can be found in Figure 6 in the Appendix.

3.2 Models

We list the models used for our experiments in Table 1. Our selection includes a diverse range of open and closed-weight models with different parameter counts. All models were tested with a temperature of 1.0 to create sufficient variance in the answers, especially for generating diverse alternative thoughts with some of the tested prompting methods[2].

Model	Version
Mistral 7B	2024-06-01
WizardLM 2 8x22B	2024-04-16
Llama 3.1 8BB	2024-07-23
Llama 3.1 70B	2024-07-23
Llama 3.1 Nemotron 70B	2024-10-16
Claude 3.5 Haiku	2024-10-22
Claude 3.5 Sonnet	2024-10-22
GPT-3.5-Turbo	2023-11-06
GPT-4o mini	2024-07-18
GPT-4o	2024-08-06
o3-mini	2025-01-31
DeepSeek V3	2025-01-03
DeepSeek R1	2025-01-22

Table 1: LLMs evaluated in this study.

We performed a memorization test to check if the dataset was part of the models’ training data, similar to the one performed by Kıcıman et al. (2024). We found no evidence of the LLMs having memorized CLadder questions. The prompts used for this test can be found in Appendix C.

3.3 Prompting Techniques

We test various prompting techniques to see if they improve the causal reasoning abilities of LLMs.

Input-Output Prompting In this simple baseline approach, the LLM is prompted with a question and an instruction to answer with a ’yes’ or ’no’ in the end.

Zero-shot Chain of Thought In this approach (CoT), the prompt “Let’s think step by step” (Kojima et al., 2022) is appended to each question.

Causal Chain of Thought We use the *causal chain of thought* (CausalCoT) prompt from Jin et al. (2023). CausalCoT is a six-step instruction prompt for solving formal causal inference problems. The exact prompt can be found in Appendix B.

Causal Chain of Thought with Self-Consistency We implement self-consistency (SC) decoding (Wang et al., 2023) with the CausalCoT prompt. With SC, multiple CausalCoT reasoning chains are sampled from the LLM and their majority answer is selected as the final answer. We evaluate SC with 3, 5, and 10 parallel reasoning chains (SC-{3,5,10}).

[2] We ran tests with GPT-3.5-Turbo and observed only minor accuracy differences when changing the temperature (average overall accuracy was 56.6% with temperature 0.0 vs. 57.5% with temperature 1.0). The CLadder dataset reports an accuracy of 52.18% for GPT-3.5-Turbo.

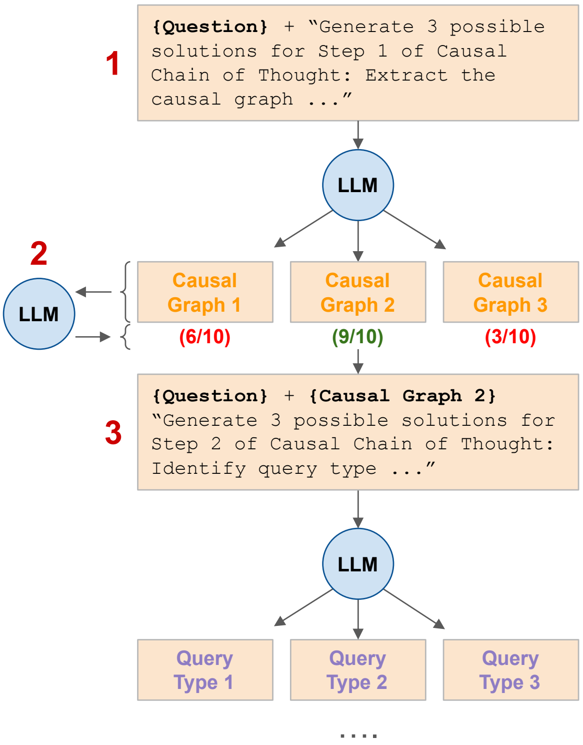

1
{Question} + "Generate 3 possible solutions for Step 1 of Causal Chain of Thought: Extract the causal graph ..."
LLM
2
LLM
Causal Graph 1
(6/10)
Causal Graph 2
(9/10)
Causal Graph 3
(3/10)
3
{Question} + {Causal Graph 2} "Generate 3 possible solutions for Step 2 of Causal Chain of Thought: Identify query type ..."
LLM
Query Type 1
Query Type 2
Query Type 3
....

Figure 1: *Causal tree of thoughts* (CAUSALTOT): (1) The LLM generates three possible causal graphs as the first step of CAUSALCOT. (2) Each graph is evaluated with a score between 1 and 10 by the LLM. (3) The highest-scoring solution, along with the question, is then used to generate three query types. This iterative process continues for the six steps of CAUSALCOT.

Causal Tree of Thoughts We propose a new causal adaptation of the *tree of thoughts* (Yao et al., 2023) prompting technique. This CAUSALTOT technique follows the six distinct steps of CAUSALCOT, but is able to consider multiple alternative thought candidates for each step $i = 1, \cdots, 6$. Unlike SC, CAUSALTOT self-evaluates thought candidates after each step and selects the best one to proceed with. Figure 1 provides a high-level overview of how CAUSALTOT operates, illustrating its iterative process of generating, evaluating, and selecting causal thoughts for a question. Throughout this process, CAUSALTOT maintains a memory of the reasoning state $s = (x, z_{1\dots i})$ consisting of the causal question x and all causal thoughts $z_{1\dots i}$ so far.

For **thought generation**, CAUSALTOT queries an LLM p_θ with a GEN_i prompt (see Appendix B for the exact prompts used) to generate k_i alternative thought candidates following the thoughts from previous steps. Hereby, k_i and GEN_i are different for each of the six steps and are curated to cater for the unique requirements of each step:

$$\{z_{i+1}^{(1)}, \cdots, z_{i+1}^{(k_i+1)}\} \sim p_\theta^{GEN_{i+1}}(z_{i+1}|s) \qquad (1)$$

For **thought evaluation**, CAUSALTOT self-selects the best thought by assigning a score between 1 and 10 to each thought and continuing with the highest-scoring thought z_i^*:

$$z_i^* \sim p_\theta^{EVAL_i}(z_i^*|\{z_i^{(1)}, \cdots, z_i^{(k_i)}\}) \qquad (2)$$

where $EVAL_i$ is the prompt for voting and selecting the best thought. Once the best thought has been chosen, the process is repeated for the following steps, starting with the GEN_{i+1} prompt again. This way, CAUSALTOT greedily decodes a *causal chain of thought* towards the final answer.

In their error analysis of CAUSALCOT, Jin et al. (2023) argue that steps 2, 3, and 5 pose the greatest challenges to the LLM. Further, causal graphs extracted by the LLM in step 1 sometimes differ from the ground truth causal graphs. Hence, we decide to set $k_i = 3$ for $i \in \{1, 2, 3, 5\}$ to explore alternative thoughts for each of these critical and error-prone steps. For the two other steps, we forego any branching (i.e., $k_i = 1, i \in \{4, 6\}$), as these steps tend to be handled rather reliably by the LLM.

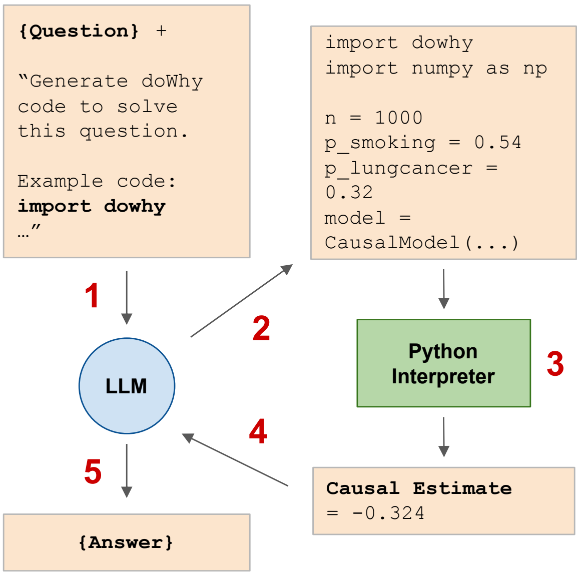
Causal Program of Thoughts We also introduce a causal version of the *program of thoughts* (Chen et al., 2023) prompting technique, which uses *DoWhy* (Sharma and Kiciman, 2020), a Python library for causal inference that supports explicit modeling and testing of causal assumptions. CAUSALPOT uses an LLM to generate *DoWhy* code c to calculate a causal estimate for a question x:

$$c \sim p_\theta^{CODE}(c|x) \qquad (3)$$

The *DoWhy* code is executed by a Python interpreter f using *REPL* (LangChain Contributors, 2024). A causal estimate $\hat{e}$ is then computed and, along with the question, provided to the LLM to generate a final answer y:

$$\hat{e} = f(c) \qquad (4)$$

$$y \sim p_\theta^{ANSWER}(y|x, \hat{e}) \qquad (5)$$



{Question} +

“Generate doWhy code to solve this question.

Example code:
import dowhy
…”

```
import dowhy
import numpy as np

n = 1000
p_smoking = 0.54
p_lungcancer =
0.32
model =
CausalModel(...)
```

```
Causal Estimate
= -0.324
```

{Answer}

Figure 2: *Causal program of thoughts* (CAUSALPOT): (1) The question and example code are input into the LLM. (2) The LLM generates *DoWhy* code. (3) The generated code is executed by the Python interpreter to compute a causal estimate. (4) This estimate is returned to the LLM. (5) The LLM decides the final answer.

Figure 2 provides a high-level overview of the CAUSALPOT methodology. The exact $CODE$ and $ANSWER$ prompts used can be found in Appendix B. We include three code examples, one from each rung, in the $CODE$ prompt. These examples are not part of the sampled dataset and are intended to guide the LLM in (1) using the correct libraries from the *DoWhy* library to solve the problem, (2) generating artificial data based on the information in the causal question, which is used to calculate the causal estimate, and (3) ensuring that the generated code is in the correct format for execution by the Python interpreter.

Additionally, in the input prompt, we added an instruction for the LLM to not mistake $p(X \mid Y)$ with $p(X \cap Y)$, which we frequently noticed in our experiments. The LLM would evaluate a statement like “the probability of smoking and lung cancer” as $p(Smoking \mid LungCancer)$ rather than $p(Smoking \cap LungCancer)$.

4 Experiments

4.1 LLMs’ Causal Reasoning Performance

RQ1: *How well do LLMs perform in different disciplines of causal reasoning?*

To establish each model’s baseline causal reasoning performance, we let the models predict the correct answers using input-output prompting on reasoning problems from the associational, interventional, and counterfactual rungs within the commonsensical subset of our CLADDER sample. Results can be seen in the left half of Table 2 (under RQ1).

The results indicate significant discrepancies between the LLMs. The weakest model, Mistral 7B, performs only somewhat better than random guessing while the strongest model, DeepSeek R1, achieves an average accuracy of 89.1% on the commonsensical questions. The largest and most recent LLMs seem to outperform the smaller and older LLMs.

The highest accuracy using input-output prompting reported in the original CLADDER paper (Jin et al., 2023) on commonsensical questions was 62.27% with GPT-4. In our test, this accuracy is beaten by 11 out of the 13 evaluated models. While there may be minor differences in our sample and testing procedure vs. Jin et al.’s (2023), we hypothesize that the most likely explanation is a strong general improvement in causal reasoning performance in newer generations of LLMs.

When comparing results across the three rungs, it seems that a majority of the evaluated LLMs generally perform best on associational questions, followed by interventional questions, and lastly counterfactual questions, as the *Ladder of Causation* suggests. This is unsurprising, as counterfactual questions are inherently more complex, requiring a deeper understanding of advanced causal inference concepts. We provide example questions from each rung of the dataset in Appendix D.

The three lowest-performing LLMs, Mistral 7B, Llama 3.1 8B, and GPT-3.5-Turbo surprisingly perform significantly better on interventional problems than on associational problems. For all but two LLMs, questions from the counterfactual rung are the most difficult, showing mostly sharp accuracy drops compared to the other two rungs. This matches observations in related works that LLMs have difficulties with counterfactual reasoning (Lewis and Mitchell, 2024; Li et al., 2022; Wu et al., 2024c).

To understand why and how models fail to reach correct answers, we manually assessed the model outputs for GPT-4o and GPT-4o mini, representing two high-performing models of different sizes. We classify a random sample of 100 incorrectly answered reasoning questions into four error types:

- **Type 1: Misinterprets the question.** The

		RQ1				*RQ2*							
		Commonsensical				**Anti-commonsensical**				**Nonsensical**			
Model	**Avg.**	**Avg.**	**R1**	**R2**	**R3**	**Avg.**	**R1**	**R2**	**R3**	**Avg.**	**R1**	**R2**	**R3**
Mistral 7B	53.1	55.3	50.1	63.5	56.5	52.9	50.9	61.4	50.8	51.0	48.9	61.9	47.7
WizardLM 2 8x22B	77.4	79.2	88.4	82.7	68.1	77.4	86.7	86.3	63.6	75.6	83.0	85.8	63.1
Llama 3.1 8B	59.1	64.9	68.1	79.7	54.3	58.6	61.7	69.0	50.3	53.8	57.3	68.0	43.2
Llama 3.1 70B	78.6	79.7	86.2	87.8	69.1	79.0	83.7	89.8	68.8	77.1	81.0	86.8	68.3
Llama 3.1 Nemo. 70B	81.2	82.8	91.4	**90.9**	70.1	80.7	83.7	88.8	73.6	80.0	85.2	86.8	71.4
Claude 3.5 Haiku	78.0	78.6	87.9	87.3	64.8	79.9	90.4	92.4	63.1	75.5	85.4	87.8	59.3
Claude 3.5 Sonnet	84.1	85.3	94.1	85.8	76.1	84.2	90.4	90.9	74.6	82.8	90.6	87.8	72.4
GPT-3.5-Turbo	57.5	58.2	54.3	72.6	55.0	56.5	52.1	72.1	53.3	57.7	51.9	67.5	58.8
GPT-4o mini	78.4	79.8	92.1	85.3	64.6	78.1	86.7	83.2	66.8	77.2	91.4	77.7	62.6
GPT-4o	82.0	84.3	95.1	**90.9**	70.1	82.4	90.1	**93.9**	68.8	79.3	89.9	88.3	64.1
o3-mini	86.8	86.3	96.8	88.3	74.6	86.9	92.3	88.8	80.4	**87.1**	92.6	**89.8**	**80.2**
DeepSeek V3	80.9	81.2	95.8	87.3	63.3	81.5	92.8	88.3	66.6	80.1	92.6	86.3	64.3
DeepSeek R1	**88.1**	**89.1**	**97.3**	86.8	**81.9**	**88.2**	**94.3**	89.8	**81.2**	86.9	**94.3**	88.8	78.4
Average	75.8	77.3	84.4	83.8	66.8	75.9	81.2	84.2	66.3	74.2	80.3	81.8	64.1

Table 2: The table shows the causal reasoning accuracy of the evaluated models on the three dataset parts commonsensical (for RQ1), as well as anti-commonsensical and nonsensical (for RQ2). For each model and dataset part, the average accuracy per rung (R1: associational, R2: interventional, R3: counterfactual) and the average across the three rungs are reported. The leftmost column contains the overall average accuracy across all reasoning questions.

Error Type	GPT-4o	GPT-4o mini
Type 1	42	33
Type 2	23	40
Type 3	8	13
Type 4	27	14

Table 3: For a sample of 100 incorrect answers, we identify the primary reasoning error that caused the wrong answer and classify it into one of four types: misinterprets the question (Type 1), relies on intuition over computation (Type 2), incorrect data extraction (Type 3), applies incorrect formula (Type 4).

model misunderstands the causal relationships in the question.

- **Type 2: Relies on intuition over computation.** Instead of performing probability calculations based on the given data, the model just provides an intuitive answer.

- **Type 3: Incorrect data extraction.** The model extracts incorrect probability data from the natural language question.

- **Type 4: Applies incorrect formula.** The model understands the question but uses the wrong formula to compute the causal effect.

Table 3 reports the errors observed. Both, GPT-4o and GPT-4o mini seem to interpret the available data mostly correctly but fail to determine the right approach to solve the problem (incl. misinterpreting the question and relying on intuition rather than calculations), or carry out calculations with an incorrect formula. The smaller GPT-4o mini seems to rely on intuition more often than its larger sibling GPT-4o, leading to relatively fewer calculation-related errors. Since GPT-4o attempts actual calculations more often, its most common errors affect the correct execution of these mathematical calculations.

4.2 Reliance on Learned Knowledge

RQ2: *How well do LLMs generalize to causal reasoning tasks where they cannot rely on learned commonsense knowledge?*

If LLMs do not perform genuine causal reasoning but rely on commonsense knowledge acquired during training, one would expect that performance drops sharply when models must reason about unfamiliar structures. To test this, we repeat the previous experiment on the anti-commonsensical and nonsensical parts of the dataset. The anti-commonsensical problems contain entities likely familiar to the LLM, but with uncommon causal relationships. The nonsensical problems contain random four-letter words with unfamiliar causal relationships.

The right half of Table 2 (under RQ2) shows the results of these experiments. While the two smallest 7B and 8B LLMs show the largest relative performance drop, the remaining medium-sized and large models seem to reason almost as well about the anti-commonsensical and nonsensical problems as about the commonsensical problems, sometimes even showing a slight accuracy increase. This could indicate that genuine causal reasoning about unfamiliar structures is an ability emerging in LLMs with scale (Wei et al., 2022a). Reasoning performance seems to be slightly higher on the anti-commonsensical problems than on the nonsensical problems, suggesting that models reason better when at least the entities are familiar, even though causal relationships between these entities are not. This may represent a confirmation of the results by Li et al. (2022) suggesting that reasoning abilities of recent LLMs still somewhat depend on simple lexical cues, which are present in the anti-commonsensical problems but not in the nonsensical problems.

4.3 Improvements through Prompting Techniques and Tool Usage

RQ3: *Can prompting techniques and external tools enhance LLMs' causal reasoning?*

Prompting techniques and usage of external tools have been shown to improve LLMs' reasoning performance, often substantially (Wei et al., 2022b; Wang et al., 2023; Yao et al., 2023; Xu et al., 2023). Jin et al. (2023) have demonstrated how the *causal chain of thought* (CAUSALCOT) prompting technique can improve GPT-4's causal reasoning accuracy on CLADDER from 62.03% to 70.40%, on average. In Section 3, we introduced two new causal prompting techniques CAUSALTOT and CAUSALPOT.

Table 4 shows the accuracy of different prompting techniques on CLADDER using a GPT-4o LLM. We chose GPT-4o as a base model for this experiment as it strikes a reasonable balance between competitive reasoning accuracy, low cost, and short runtime. Interestingly, we observe CAUSALCOT to perform 2.4%-points worse than input-output prompting when used with GPT-4o. It is worth noting though that GPT-4o with input-output prompting already achieves an overall average accuracy of 82.0%, which is substantially higher than the accuracy Jin et al. (2023) measured for GPT-4. This may indicate that recent advancements in model architectures and training procedures made sophisticated prompting techniques dispensable on the CLADDER problems. Noticeably, simple zero-shot COT improved causal reasoning accuracy by 1.2%-points, on average, versus input-output prompting. *Self-consistency* (SC) performed similar to input-output prompting, independent of the number of parallel reasoning chains.

Our new prompting techniques CAUSALTOT and CAUSALPOT outperform input-output prompting by an average of 4.4%-points and 8.8%-points, respectively. With CAUSALTOT, GPT-4o even reaches close to the performance of o3-mini and DeepSeek R1, the strongest reasoning models we evaluated. With CAUSALPOT, GPT-4o surpasses o3-mini by 4.0%-points and DeepSeek R1 by 2.7%-points, on average. This shows that in the domain of formal causal reasoning, the domain-specialized prompting techniques applied in CAUSALTOT and CAUSALPOT can match or even outperform the extensive but non-specialized test-time thinking done by o3-mini and DeepSeek R1.

Remarkably, CAUSALTOT outperforms all other tested prompting techniques on questions from the associational rung, but loses accuracy on the other two rungs, especially on counterfactual questions. On the other hand, CAUSALPOT achieves slightly lower performance on associational questions than many of the other prompting techniques but maintains a fairly consistent accuracy throughout all rungs. With that, CAUSALPOT seems to be the first prompting technique that performs similarly well on counterfactual questions as on associational or interventional questions.

An error analysis for each of our methods can be found in Appendix E.

5 Conclusion

Connecting to previous work on causal reasoning in LLMs, we have presented a systematic evaluation of causal reasoning abilities of the most recent LLMs. Our findings indicate that the latest models perform substantially better than older LLMs evaluated in previous works. These state-of-the-art LLMs seem to reason well, even on challenging causal reasoning tasks and unfamiliar causal structures. One exception is counterfactual reasoning, which still poses significant challenges to state-of-the-art LLMs. While we found that previous prompting techniques designed to improve LLMs' reasoning performance no longer show the desired improvements on recent LLMs, we proposed two new causal prompting techniques. As demonstrated, CAUSALTOT and CAUSALPOT can significantly elevate reasoning performance, even

		RQ3											
		Commonsensical				**Anti-commonsensical**				**Nonsensical**			
Model	**Avg.**	**Avg.**	**R1**	**R2**	**R3**	**Avg.**	**R1**	**R2**	**R3**	**Avg.**	**R1**	**R2**	**R3**
GPT-4o	82.0	84.3	95.1	90.9	70.1	82.4	90.1	**93.9**	68.8	79.3	89.9	88.3	64.1
GPT-4o + CoT	83.2	85.3	95.6	92.9	71.1	83.7	92.6	**93.9**	69.6	80.7	91.6	90.9	64.6
GPT-4o + CausalCoT	79.6	81.1	93.1	88.3	65.3	80.2	91.6	90.9	63.3	77.4	89.1	87.8	60.3
GPT-4o + SC-3	82.2	83.0	93.8	89.8	68.6	82.3	90.6	89.3	70.4	81.3	90.9	89.3	67.6
GPT-4o + SC-5	81.6	82.7	94.1	87.8	68.6	81.2	84.9	84.3	75.9	80.8	88.6	85.3	70.6
GPT-4o + SC-10	82.0	83.5	93.8	90.4	69.6	80.3	89.9	88.8	66.3	82.3	91.1	90.4	69.3
GPT-4o + CausalToT	86.4	87.5	**96.3**	91.9	76.4	86.5	**93.8**	**93.9**	75.4	85.3	**92.6**	90.9	75.1
GPT-4o + CausalPoT	**90.8**	**92.5**	91.6	**94.4**	**92.5**	**90.3**	89.1	92.9	**90.2**	**89.6**	90.6	**91.9**	**87.4**

Table 4: The table shows the causal reasoning accuracy with different prompting techniques using GPT-4o.

of recent LLMs. CausalPoT appears to be the only causal prompting technique that substantially improves performance on counterfactual reasoning problems.

Limitations and Future Work

In this paper, we focus exclusively on formal causal reasoning and do not evaluate LLMs' capabilities on informal reasoning tasks. This is because we believe that several works already discuss informal causal reasoning with LLMs and, while their results are insightful and relevant, we see formal causal reasoning problems as more suitable to assess whether LLMs can genuinely reason. Nonetheless, our proposed methods CausalToT and CausalPoT were specifically designed for formal causal reasoning and problem formulations similar to those included in CLadder. We leave it to future work to ideate similar methods that generalize beyond the scope of formal causal reasoning.

Some readers may criticize the limited breadth of evidence put forward in our analysis, where we evaluate all methods on CLadder only, with CLadder being a synthetic dataset and our experimental sample being limited to 1,000 examples. We certainly encourage future work to continue to evaluate LLMs on several datasets. However, we also note that CLadder alone is perhaps one of the most comprehensive evaluation tasks for formal causal reasoning (Yang et al., 2024), covering all rungs of the *Ladder of Causation*, as well as commonsensical, anti-commonsensical, and nonsensical problem formulations, and nine different query types. In addition, the authors conduct a broad range of quality checks including grammaticality, human readability, and naturalness/perplexity (Jin et al., 2023). For these reasons, we argue that CLadder served as the ideal evaluation bench to rigorously evaluate our methods within the constraints of our resources.

Our anti-commonsensical and nonsensical perturbations were generated using GPT-4o, raising concerns that this may have made it easier for GPT-4o to recognize its own perturbations, potentially leading to an artificial inflation of its performance. However, a similar decline of performance from commonsensical to anticommonsensical to nonsensical seen in GPT-4o is evident in other LLMs. We also see a larger performance decline for GPT-4o than what was reported in the CLadder paper (Jin et al., 2023) for GPT-4, suggesting that GPT-4o scores are not substantially inflated.

For future work, we still recommend evaluating CausalPoT and CausalToT on other formal causal reasoning datasets, such as Corr2Cause (Jin et al., 2024), to assess their effectiveness and generalizability. We believe that leveraging external libraries could enhance the performance of LLMs in these tasks.

Ethical Considerations

While we have shed light on the causal reasoning abilities of current LLMs, no general evaluation can replace a detailed assessment of a specific LLM in the context of its final use case. Using LLMs for causal reasoning comes with risks and our results should not be seen as a free pass for using LLMs for purely machine-based decision-making. An oversimplification of complex causal phenomena may lead to high-stakes errors, particularly in domains such as healthcare or public policy. Open dissemination of powerful LLM-based causal methods risks malicious applications, including generating deceptive causal claims. Mitigation strategies may include careful curation of training data, the integration of formal causal inference tools, transparent reporting of model capabilities and limitations, and stricter governance of high-stakes use cases.

References

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Mayank Jindal, Dushyant Singh Sengar, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: Can large language models truly understand causality?

Hengrui Cai, Shengjie Liu, and Rui Song. 2024. Is knowledge all large language models needed for causal reasoning?

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024. Unveiling causal reasoning in large language models: Reality or mirage? In *Advances in Neural Information Processing Systems*, volume 37, pages 96640–96670. Curran Associates, Inc.

Jörg Frohberg and Frank Binder. 2022. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models.

Gaël Gendron, Jože M. Rožanec, Michael Witbrock, and Gillian Dobbie. 2024. Counterfactual causal inference in natural language with large language models.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. Cladder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 31038–31065. Curran Associates, Inc.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation?

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality.

LangChain Contributors. 2024. *LangChain Python Integration Documentation*. Accessed: 2024-12-01.

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models.

Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2022. Counterfactual reasoning: Do language models need world knowledge for causal understanding?

Xiao Liu, Da Yin, Chen Zhang, Yansong Feng, and Dongyan Zhao. 2023. The magic of if: Investigating causal reasoning abilities in large language models of code.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. 2025. Large language models and causal inference in collaboration: A comprehensive survey.

Stephanie Long, Tibor Schuster, and Alexandre Piché. 2024. Can large language models build causal graphs?

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*, 1st edition. Basic Books, Inc., USA.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. 2023. Crab: Assessing the strength of causal relationships between real-world events.

Amit Sharma and Emre Kiciman. 2020. Dowhy: An end-to-end library for causal inference.

Lei Wang and Yiqing Shen. 2024. Evaluating causal reasoning capabilities of large language models: A systematic analysis across three scenarios. *Electronics*, 13(23).

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Zeyu Wang. 2024. CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024a. Causality for large language models.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024b. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024c. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks.

Binfeng Xu, Zhiyuan Peng, Bowen Lei, Subhabrata Mukherjee, Yuchen Liu, and Dongkuan Xu. 2023. Rewoo: Decoupling reasoning from observations for efficient augmented language models.

Linying Yang, Vik Shirvaikar, Oscar Clivio, and Fabian Falck. 2024. A critical review of causal reasoning benchmarks for large language models.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Qingzhen Liu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. 2025. Causaleval: Towards better causal reasoning in language models.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. 2023. Understanding causality with large language models: Feasibility and opportunities.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models.

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. 2024. Causalbench: A comprehensive benchmark for causal learning capability of llms.

A Comparison of Sampled and Original Dataset Distributions

Table 5 presents the number of causal questions across different properties in the sampled and original datasets. Figures 3, 4, and 5 illustrate the data distribution for each metric in both datasets, showing that the distribution of the sampled dataset (1,000 causal questions) closely matches that of the original commonsensical dataset (8,690 causal questions).

Metric	Sampled	Original
Answer		
No	504	4345
Yes	496	4345
Query Type		
Marginal Prob.	209	1702
ATE	174	1518
Conditional Prob.	174	1518
ATT	137	1288
Counterfactual	95	874
NIE	92	870
NDE	73	552
Collider Bias	23	184
Explaining Away	22	184
Rung		
Rung 1	405	3584
Rung 2	398	3404
Rung 3	197	1702

Table 5: Number of causal questions per metric for sampled and original dataset

B Prompts

The code and prompts used for all experiments can be found in `https://github.com/rahulbshrestha/causal-reasoning`

Specifically, the prompts used for the memorization test, *causal chain of thought* and *program of thoughts* can be found in `https://github.com/rahulbshrestha/causal-reasoning/blob/main/src/prompts.py`

C Memorization Test

To verify that the dataset was not included in the training data for each model, we conducted a memorization test as outlined in Kıcıman et al.'s (2024).

For the basic test, we asked the LLMs whether they were familiar with the CLADDER dataset using the following prompt:

"Do you know about the dataset CLadder*: Assessing Causal Reasoning in Language Models? If yes, please provide the names of the authors, the number of questions in the dataset, and an example row from the dataset."*

We observed that all LLMs either fabricated the information for all three values or stated that they did not recognize the dataset.

For a more rigorous evaluation, we employed a memorization test prompt inspired by (Kıcıman et al., 2024). Details on the exact prompt used are provided in Appendix B. In this test, the LLM was tasked with recalling three partial questions from the dataset. To enhance the likelihood of successful reconstruction, the LLM was first provided with additional contextual information, including the dataset's name, URL, a description extracted from the README file, and two few-shot examples from the dataset.

The three partial questions are presented below. The italicized portions were deliberately omitted from the prompt, and the LLM was expected to reconstruct them.

Q1: The overall probability of manager signing the termination letter is 39%. For managers who don't sign termination letters, *the probability of employee being fired is 22%. For managers who sign termination letters, the probability of employee being fired is 60%. Is employee being fired less likely than employee not being fired overall?*

Q2: For unvaccinated individuals, the probability of smallpox survival is 35%. For vaccinated individuals, *the probability of smallpox survival is 40%. Does vaccination status positively affect smallpox survival through getting smallpox and vaccination reaction?*

Q3: For infants with nonsmoking mothers, the probability of high infant mortality is 88%. For infants with smoking mothers, *the probability of high infant mortality is 64%. For infants with smoking mothers, would it be less likely to see high infant mortality if the infant had a nonsmoking mother? Let's think step by step. Answer with 'yes' or 'no' at the end.*

We observed that the LLMs failed to reconstruct the questions accurately, instead generating random data that did not match the original dataset.

D Sample Questions from CLadder

In this section, we present sample data points from the CLadder dataset. The "Info" and "Question" together form the causal question. "Answer" represents the ground truth. "Query Type" indicates one of nine distinct query categories. "Rung" specifies the causal hierarchy level of the answer (1 = Association, 2 = Intervention, 3 = Counterfactual). "Formal Form" provides the mathematical representation of the query type, and "Reasoning" outlines the step-by-step approach to solving the question.

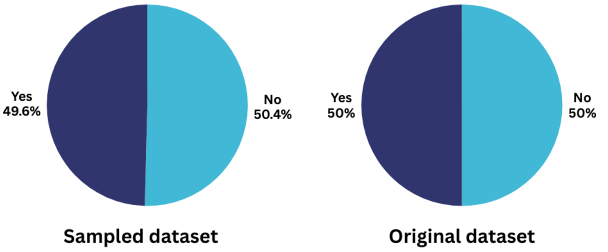


Figure 3: Comparison of Answer Distributions (Yes/No) in Original vs. Sampled Datasets

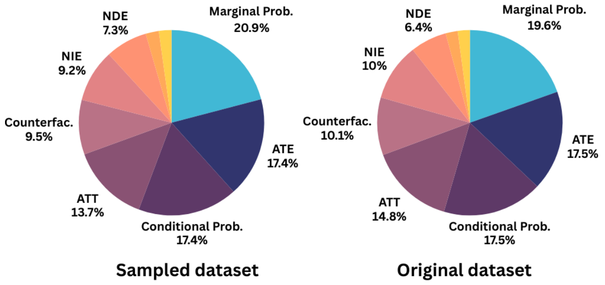


Figure 4: Comparison of Query Types Distributions in Original vs. Sampled Datasets

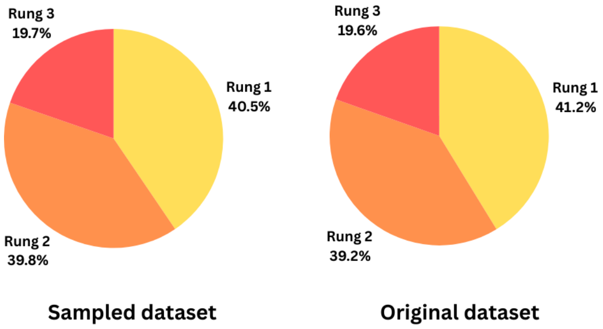


Figure 5: Comparison of Rung Type Distributions in Original vs. Sampled Datasets

D.1 Rung 1 Question

Info: The overall probability of alarm set by husband is 73%. The probability of alarm not set by husband and ringing alarm is 10%. The probability of alarm set by husband and ringing alarm is 40%.

Question: Is the chance of ringing alarm larger when observing alarm set by husband?

Answer: yes

Query type: correlation
Rung: 1
Formal form: $P(Y|X)$
Reasoning:

1. Let X = husband; V_2 = wife; Y = alarm clock.
2. $X \to V_2, X \to Y, V_2 \to Y$
3. $P(Y|X)$
4. $P(X = 1, Y = 1)/P(X = 1) - P(X = 0, Y = 1)/P(X = 0)$
5. $P(X = 1) = 0.73$
   $P(Y = 1, X = 0) = 0.10$
   $P(Y = 1, X = 1) = 0.40$
6. $0.40/0.73 - 0.10/0.27 = 0.18$
7. $0.18 > 0$

D.2 Rung 2 Question

Info: For CEOs who fire employees and managers who don't sign termination letters, the probability of employee being fired is 43%. For CEOs who fire employees and managers who sign termination letters, the probability of employee being fired is 81%. For CEOs who don't fire employees and managers who don't sign termination letters, the probability of employee being fired is 63%. For CEOs who don't fire employees and managers who sign termination letters, the probability of employee being fired is 98%. The overall probability of CEO's decision to fire the employee is 26%.

Question: Will manager signing the termination letter decrease the chance of employee being fired?
Answer: no
Query type: ATE
Rung: 2
Formal form:
$E[Y|do(X = 1)] - E[Y|do(X = 0)]$
Reasoning:

1. Let V_1 = CEO; V_3 = director; X = manager; Y = employee.
2. $V_1 \to V_3, V_1 \to X, X \to Y, V_3 \to Y$
3. $E[Y|do(X = 1)] - E[Y|do(X = 0)]$
4. $\sum_{V_1=v} P(V_1 = v) * [P(Y = 1|V_1 = v, X = 1) - P(Y = 1|V_1 = v, X = 0)]$
5. $P(Y = 1|V_1 = 0, X = 0) = 0.43$
   $P(Y = 1|V_1 = 0, X = 1) = 0.81$
   $P(Y = 1|V_1 = 1, X = 0) = 0.63$
   $P(Y = 1|V_1 = 1, X = 1) = 0.98$
   $P(V_1 = 1) = 0.26$
6. $0.74 * (0.81 - 0.43) + 0.26 * (0.98 - 0.63) = 0.38$
7. $0.38 > 0$

D.3 Rung 3 Question

Info: For those who choose to take the stairs and penguins who are sad, the probability of penguin death is 28%. For those who choose to take the stairs and penguins who are happy, the probability of penguin death is 60%. For those who choose to take the elevator and penguins who are sad, the probability of penguin death is 35%. For those who choose to take the elevator and penguins who are happy, the probability of penguin death is 74%. For those who choose to take the stairs, the probability of penguin happiness is 57%. For those who choose to take the elevator, the probability of penguin happiness is 22%.

Question: Does my decision negatively affect penguin survival through penguin mood?
Answer: yes
Query type: NIE
Rung: 3
Formal form: $E[Y_{X=0,V_2=1} - Y_{X=0,V_2=0}]$
Reasoning:

1. Let X = my decision; V_2 = penguin mood; Y = penguin survival.
2. $X \to V_2, X \to Y, V_2 \to Y$
3. $E[Y_{X=0,V_2=1} - Y_{X=0,V_2=0}]$
4. $\sum_{V_2=v} P(Y = 1|X = 0, V_2 = v) * [P(V_2 = v|X = 1) - P(V_2 = v|X = 0)]$
5. $P(Y = 1|X = 0, V_2 = 0) = 0.28$
   $P(Y = 1|X = 0, V_2 = 1) = 0.60$
   $P(Y = 1|X = 1, V_2 = 0) = 0.35$
   $P(Y = 1|X = 1, V_2 = 1) = 0.74$
   $P(V_2 = 1|X = 0) = 0.57$
   $P(V_2 = 1|X = 1) = 0.22$
6. $0.22 * (0.60 - 0.28) + (1 - 0.22) * (0.74 - 0.35) - (0.57 * (0.60 - 0.28) + (1 - 0.57) * (0.74 - 0.35)) = 0.0704 + 0.2964 - (0.1824 + 0.1671) = 0.3668 - 0.3495 = 0.0173$
7. $0.0173 > 0$

Step	Error Count
Step 1	3
Step 2	18
Step 3	17
Step 4	12

Table 6: For a sample of errors in CAUSALCOT, we identify the step at which GPT-4o made a mistake.

E Error Analysis

E.1 Causal Chain of Thought

For CAUSALCOT, we conducted an error analysis on 50 samples, categorizing failures based on the specific step in the prompt where the model made an error. Table 6 presents our findings. If a model fails at an earlier step, we do not assess its performance on subsequent steps. Our results align with those reported by (Jin et al., 2023), who found that LLMs most frequently fail at Steps 2, 3, and 5 of CAUSALCOT. Similarly, we observed errors in Steps 2 and 3, but we did not encounter any cases where the model successfully completed the first four steps and then failed at Step 5.

E.2 Chain of Thought with Self-Consistency

To verify that COT-SC produces different answers for different reasoning chains—i.e., that there is variation—we analyzed the distribution of ‘yes’ and ‘no’ responses.

As shown in Tables 7 and 8, the model generated varying distributions of answers. In Table 5.1, the highest frequency of responses falls into the (10 yes, 0 no) or (0 yes, 10 no) categories, with frequencies decreasing from there. This suggests that sampling different reasoning chains (for nearly half the dataset) does not significantly impact most questions. The same pattern holds for SC-5 (5 chains), which may explain why increasing to SC-10 (10 chains) does not improve accuracy.

Distribution	# Answers
(10 yes + 0 no) or (0 yes + 10 no)	403
(9 yes + 1 no) or (1 yes + 9 no)	169
(8 yes + 2 no) or (2 yes + 8 no)	127
(7 yes + 3 no) or (3 yes + 7 no)	128
(6 yes + 4 no) or (4 yes + 6 no)	109
(5 yes + 5 no)	59

Table 7: Distribution of answers in SC-10 (10 chains)

Distribution	# Answers
(5 yes + 0 no) or (0 yes + 5 no)	535
(4 yes + 1 no) or (1 yes + 4 no)	246
(3 yes + 2 no) or (2 yes + 3 no)	218

Table 8: Distribution of answers in SC-5 (5 chains)

Step	Error Count
Step 1	4
Step 2	23
Step 3	12
Step 4	11

Table 9: For a sample of errors in CAUSALTOT, we identify the step at which GPT-4o made a mistake.

E.3 Causal Tree of Thoughts

For CAUSALCOT, we conducted an error analysis on 50 samples, categorizing failures based on the specific step in the prompt where the model made an error. Table 9 presents our findings. We categorize mistakes the LLM made in generating or evaluating thoughts under the same step. If a model fails at an earlier step, we do not assess its performance on subsequent steps.

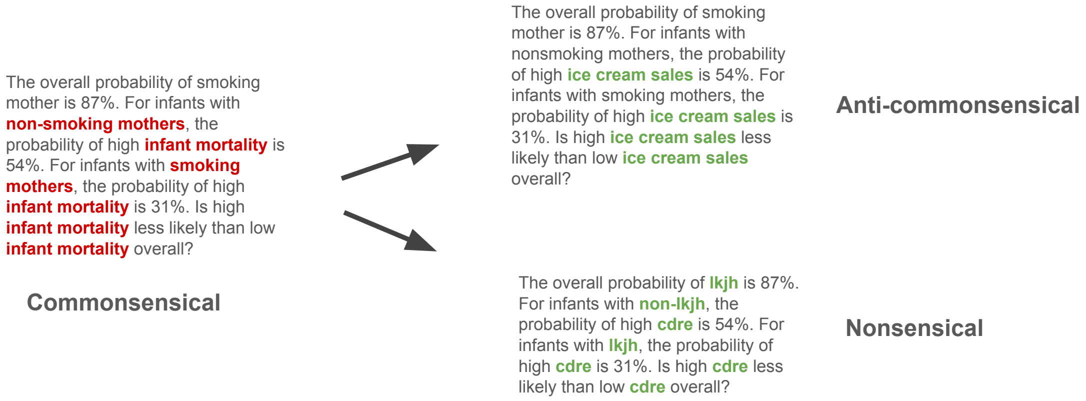
E.4 Causal Program of Thoughts

For CAUSALPOT, we conducted an error analysis on 50 samples, categorizing failures based on 3 different error types. Results are shown in Table 10.

- **Type 1: Incorrect causal graph extracted.** The model extracts an incorrect causal graph based on the question.
- **Type 2: Incorrect library function call.** The model uses the wrong library function when estimating causal effects. This often results from misidentifying the required rung type for solving the causal question.
- **Type 3: Incorrect code produced (code execution failure).** The model generates incorrect code due to formatting errors, incorrect library names, or other issues, causing execution failures or runtime errors.

Error Type	**Error Count**
Type 1	10
Type 2	37
Type 3	13

Table 10: For a sample of errors in CAUSALPOT, we classify the primary reasoning mistake into three types.



The overall probability of smoking mother is 87%. For infants with **non-smoking mothers**, the probability of high **infant mortality** is 54%. For infants with **smoking mothers**, the probability of high **infant mortality** is 31%. Is high **infant mortality** less likely than low **infant mortality** overall?

The overall probability of smoking mother is 87%. For infants with nonsmoking mothers, the probability of high **ice cream sales** is 54%. For infants with smoking mothers, the probability of high **ice cream sales** is 31%. Is high **ice cream sales** less likely than low **ice cream sales** overall?

The overall probability of **lkjh** is 87%. For infants with **non-lkjh**, the probability of high **cdre** is 54%. For infants with **lkjh**, the probability of high **cdre** is 31%. Is high **cdre** less likely than low **cdre** overall?

Figure 6: Generating the anti-commonsensical and nonsensical perturbed datasets