C2P: Featuring Large Language Models with Causal Reasoning

Anonymous authors
Paper under double-blind review

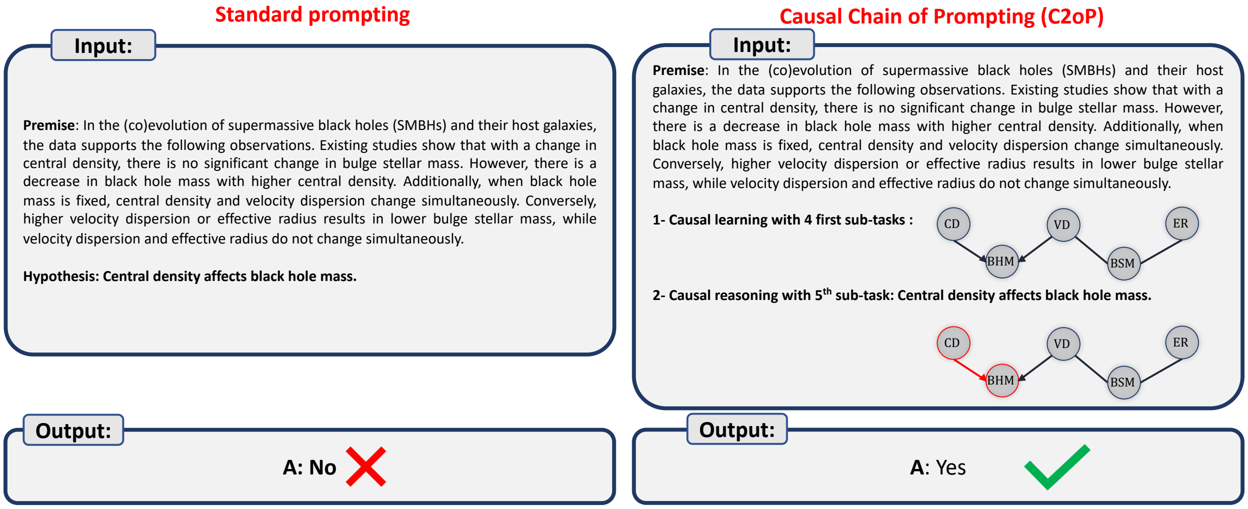
Abstract

Causal reasoning is the primary bottleneck that Large Language Models (LLMs) must overcome to attain human-level intelligence. To address this, we introduce the Causal Chain of Prompting (C2P) as the first reasoning framework that equips current LLMs with causal reasoning capabilities. C2P operates autonomously, avoiding reliance on external tools or modules during both the causal learning and reasoning phases, and can be seamlessly implemented during the training or fine-tuning of LLMs. Experimental results across various benchmark datasets demonstrate a significant improvement in causal learning and subsequent reasoning accuracy of LLMs. We illustrate how C2P enhances LLMs' ability to causally reason in real-world scenarios, addressing complex problems in fields such as healthcare, medicine, economics, education, social sciences, environmental science, and marketing. With few-shot learning, GPT-4 Turbo using C2P with as few as six examples achieves significant performance improvements, boasting over a 33% increase in reasoning accuracy over the most state-of-the-art LLMs, which perform nearly randomly in similar circumstances. This demonstrates the transformative potential of integrating C2P into LLM training or fine-tuning processes, thereby empowering these models with advanced causal reasoning capabilities.

1 Introduction

Recent advancements in Large Language Models (LLMs) have impacted existing AI paradigms and heightened expectations regarding AI's capabilities (Achiam et al., 2023; Brown et al., 2020). Despite significant architectural differences in LLMs, they generally produce outputs based on the most likely results learned from vast amounts of training data (Vaswani et al., 2017). This enables them to acquire extensive knowledge ranging from common sense to specialized domains such as mathematics and science (Jiralerspong et al., 2024). Despite this, the inefficiency of LLMs in addressing causal reasoning questions remains their primary bottleneck and simple tasks can completely break down reasoning in state-of-the-art LLMs (Nezhurina et al., 2024). Additionally, studies such as Kalai & Vempala (2023); Xu et al. (2024) have demonstrated that, despite the training data containing numerous examples of interventions, outcomes, and explanations, as well as similar tasks, hallucinatory responses persist and there is a lack of causal reasoning capability. As a result, while they may talk causality, they are not causal (Zečević et al., 2023). This deficiency represents a fundamental drawback of LLMs as AI systems compared to human intelligence, which goes beyond mere correlations and depends on causal relationships for decision-making (Penn & Povinelli, 2007; Anwar et al., 2024).

Recently, answering cause and effect questions with LLMs has gained extensive interest (Shin et al., 2020; Jin et al., 2023b; Ashwani et al., 2024). To address this issue, it is important to note that LLMs have been utilized in conjunction with external tools to extract causal structures, as demonstrated in (Jiralerspong et al., 2024). However, their architectures lack specialized modules specifically designed to enhance the understanding of cause-and-effect relationships within their outputs (Wang et al., 2023; Imani et al., 2023). Aside from studies that reason causally based on the knowledge that is already in their training data (Petroni et al., 2019; Jiang et al., 2020) and the ones that use LLMs in causality (Kıcıman et al., 2023; Zhang et al., 2023; Feder et al., 2024; Khatibi et al., 2024), chain-of-thought prompting is presented in (Wei et al., 2022) as



Premise: In the (co)evolution of supermassive black holes (SMBHs) and their host galaxies, the data supports the following observations. Existing studies show that with a change in central density, there is no significant change in bulge stellar mass. However, there is a decrease in black hole mass with higher central density. Additionally, when black hole mass is fixed, central density and velocity dispersion change simultaneously. Conversely, higher velocity dispersion or effective radius results in lower bulge stellar mass, while velocity dispersion and effective radius do not change simultaneously.

Hypothesis: Central density affects black hole mass.

A: No ✗

Premise: In the (co)evolution of supermassive black holes (SMBHs) and their host galaxies, the data supports the following observations. Existing studies show that with a change in central density, there is no significant change in bulge stellar mass. However, there is a decrease in black hole mass with higher central density. Additionally, when black hole mass is fixed, central density and velocity dispersion change simultaneously. Conversely, higher velocity dispersion or effective radius results in lower bulge stellar mass, while velocity dispersion and effective radius do not change simultaneously.

1- Causal learning with 4 first sub-tasks :

2- Causal reasoning with 5th sub-task: Central density affects black hole mass.

A: Yes ✓

Figure 1: Example of the standard prompting vs few-shot learned GPT-4 with C2P in open problem in astrophysics (Pasquato et al., 2023)

.

one of the initial attempts in enhancing the reasoning in LLMs that shows improvement based on the data of the given query. However, LLMs still struggle with rigorous numerical and abstract reasoning among many other tasks (Xu et al., 2023). For example, a recent work, Causal Reasoning Assessment Benchmark (CRAB, Romanou et al., 2023), is designed to evaluate the causal understanding of events in real-world narratives. This study demonstrated that most systems perform poorly in identifying cause-and-effect tasks. Similarly, in (Jin et al., 2023b), the CORR2CAUSE dataset is introduced and demonstrated that current models often perform no better than random chance when tasked with causal questions. Following that, in (Jin et al., 2023a), the CLADDER dataset is introduced to asses Average Treatment Effects with LLMs and it is demonstrated that these models struggle with causal tasks. In their study, with the implementation of the proposed framework, CAUSALCoT, progress has been made in evaluating the average treatment effect where LLMs are provided with a collection of causal graphs and various types of queries (associational, interventional, and counterfactual), such as those included in the CLADDER dataset. More recently, in (Ashwani et al., 2024), a novel architecture called the Context-Aware Reasoning Enhancement with Counterfactual Analysis (CARE-CA) framework is presented to enhance causal reasoning and explainability. Their proposed framework incorporates an external explicit causal detection module with ConceptNet (Speer et al., 2017) and counterfactual statements, as well as implicit causal detection through LLMs, showing progress in causal reasoning in short and simple queries. Several other works at the intersection of causal inference and LLMs are discussed in an extensive survey by Liu et al. (2024).

As argued by Pearl (1995), causal Directed Acyclic Graphs (DAGs), along with d-separation, allow for the investigation of cause-and-effect relationships without relying on structural equation models in computational studies. Inspired by Pearl's foundational work, we propose a novel framework, named Causal Chain of Prompting (C2P), to address the inefficiencies of LLMs in handling causal queries (see Figure 1 as an example). We show that by identifying the adjacency matrix, equivalent to and instead of the causal DAG in Pearl's work, the causal relationships among the variables in the premise's cause-and-effect relationships can be effectively reasoned within the context of language models. Contrary to the existing research in addressing weaknesses of LLMs in causal reasoning, the C2P operates autonomously, avoiding reliance on external tools or modules during both the learning and reasoning phases to answer causal questions. Additionally, C2P can easily be implemented in the training or fine-tuning process of LLMs. C2P includes five simple main sub-tasks, as follows: (1) Prompting for extracting the random variables from the provided data. (2) Prompting for extracting all conditional and unconditional relations and cause and effect relations specifically mentioned among the random variables. (3) Prompting to create the initial adjacency matrix with values 1 for all elements except the diagonal elements and the ones corresponding to effect-cause relations (the cause and effects elements are also one). (4) Prompting of the conditional and unconditional independencies and identification of the colliders, step by step, to extract the causal adjacency matrix; and, (5) Prompting for reasoning questions or hypotheses. To evaluate the accuracy and reliability of implementing the C2P on LLMs, we initially assess it using publicly available synthetic datasets such as (Jin et al., 2023b). Subsequently, we evaluate it in more realistic and complex scenarios in real-world problems presented in (Pearl & Mackenzie, 2018) and (Pasquato et al., 2023). Moreover, we present results on few-shot learning with C2P on both synthetics and realistic scenarios.

Contributions. In this work, we present several important contributions to facilitate causal reasoning in language models. Concretely,

1. We introduce the C2P framework as the first reasoning framework to equip LLMs with the ability of causal reasoning within real-world scenarios, without relying on external tools.
2. Through extensive experiments with our framework, we demonstrate a significant improvement of LLMs in causal reasoning in various benchmarks. Additionally, we examine the performance of the C2P framework on more complex and real-world scenarios in various domains.
3. We perform few-shot learning experiments in section 4 with GPT-4 Turbo using our C2P framework and show how integrating C2P during the training or fine-tuning of LLMs can revolutionize existing models. This approach equips them with causal reasoning capabilities akin to the transformative impact of 'Chain of Tought' (Wei et al., 2022) on LLMs, as discussed by Chung et al. (2024).

The codes for sections 3 and 4 are publicly available at `https://anonymous.4open.science/r/C2P-5C2A/`.

2 Preliminaries on Causal Learning and Reasoning

To lay the groundwork for our framework, understanding the fundamental concept of cause and effect is paramount. A foundational and widely acknowledged principle in causality is the one proposed by Reichenbach, which posits the following:

Principle 1 (Common cause principle (Reichenbach, 1991)) *If two random variables X_1 and X_2 are statistically dependent, i.e., $X_1 \not\perp\!\!\!\perp X_2$, then there exists a third variable X_3 that causally influences both. (As a special case, X_3 may coincide with either X_1 or X_2.) Furthermore, this variable X_3 screens X_1 and X_2 from each other in the sense that given X_3, they become independent, $X_1 \perp\!\!\!\perp X_2 | X_3$.*

Markov Property: The Markov property in a DAG G states that each node X_i is conditionally independent of its non-descendants given its parents, namely $X_i \perp\!\!\!\perp \mathrm{NonDe}(X_i)|\mathrm{Pa}(X_i)$, where $\mathrm{NonDe}(X_i)$ denotes the non-descendants of X_i excluding itself, and $\mathrm{Pa}(X_i)$ denotes the parents of X_i. Using the Markov property, we can factorize the joint distribution of all the nodes in the graph into $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|\mathrm{Pa}(X_i))$. To infer the causal graph from probability distributions, a common assumption is faithfulness, namely the validity to infer all the d-separation sets in the graph from the independence relations in the probability distribution. In our work, we also take this broadly taken assumption which holds for most real-world scenarios.

Directed Acyclic graph: A graph G is called a partially directed acyclic graph (PDAG) if there is no directed cycle, that is, if there is no pair (j, k) with directed paths from j to k and from k to j. G is called a directed acyclic graph (DAG) if it is a PDAG and all edges are directed.

Markov Equivalence of Graphs: We denote two DAGs as Markov equivalent if they induce the same joint distribution $P(\mathbf{X})$. The set of DAGs that are Markov equivalent to each other is called a Markov equivalence class (MEC). Causal graphs in the same MEC can be easily identified since they have the same skeleton (i.e., undirected edges) and V-structures (i.e., structures in the form of $X_1 \rightarrow X_2 \leftarrow X_3$ where X_1 and X_3 are not connected).

The be able to perform any level of causal reasoning, the Ladder of Causation is proposed with three main rungs, "seeing", "doing", and "imagining", in (Pearl & Mackenzie, 2018). To fulfill all rungs of the ladder, more information on the causal structure is needed, which can be mostly provided with structural causal models (SCM) (Assuming all the assumptions are satisfied) (Bareinboim et al., 2022). The current causal discovery methods are divided into two groups, constraint-based algorithm such as the PC algorithm, which has a PDAG as an output, that represents the MEC of the true underlying graph and is the best result such methods can extract. Second is score-based methods such as NOTEARS (Zheng et al., 2018), GOLEM (Ng et al., 2020), and DAGMA (Bello et al., 2022), among many others, that extract a DAG that mostly fits the data and demonstrated very high accuracy in extracting Bayesian networks. Score-based methods require solving an optimization problem based on the numerical values of variables, rendering them impractical for implementation in LLM models.

PC algorithm: Following these materials, the PC algorithm is developed based on Reichenbach's common cause principle and Markov property as follows for a causal mechanism with p random variables (Spirtes et al., 2001);

(i) Form a complete undirected graph

(ii) Eliminate edges between variables that are unconditionally independent

(iii) For each pair of variables (X_1, X_2) having an edge between them, and for each variable, X_3 with an edge connected to either of them, eliminate the edge between X_1 and X_2 if $X_1 \perp\!\!\!\perp X_2|X_3$

(iv) For each pair of variables X_1, X_2 having an edge between them, and for each pair of variables $\{X_3, X_4\}$ with edges both connected to X_1 or both connected to B, eliminate the edge between X_1 and X_2 if $X_1 \perp\!\!\!\perp X_2|\{X_3, X_4\}$.

(v) For each triple of variables (X_1, X_2, X_3) such that X_1 and X_2 are adjacent, X_2 and X_3 are adjacent, and A and C are not adjacent, orient the edges $X_1 - X_2 - X_3$ as $X_1 \to X_2 \leftarrow X_3$, if X_2 was not in the set conditioning on which X_1 and X_3 became independent and the edge between them was accordingly eliminated. We call such a triple of variables a V-structure.

(vi) For each triple of variables such that $X_1 \to X_2 - X_3$, and X_1 and X_2 are not adjacent, orient the edge $X_2 - X_3$ as $X_2 \to X_3$. This is called orientation propagation.

d-separation and do calculus (Pearl, 1995): do-calculus, developed by Judea Pearl and others, is a set of rules used to transform and manipulate causal expressions within causal diagrams (or graphical models). It is a generalization of do-calculus, which is itself a formal tool used to reason in causal relationships from a mixture of experimental and observational data. do-calculus consists of three main rules that enable one to rewrite expressions involving interventions (typically represented as $do(x)$, indicating an intervention to set variable X to value x):

1. **Insertion/deletion of observations**: If a variable X_2 is independent of X_1 given a set of other variables X_3, then the conditional distribution of X_2 given $do(X_1)$ and X_3 is the same as the conditional distribution of X_2 given X_3 alone. This rule can be applied when X_1 and X_2 are d-separated by X_3 in the graph modified by removing all arrows pointing into X_1.
2. **Action/observation exchange**: If X_2 is independent of X_1 given X_3 and another set W where X_1 is intervened, then the graph can be modified to reflect this exchange.
3. **Insertion/deletion of actions**: This rule states that if intervening on X_1 does not affect X_2 after accounting for X_3, the intervention can be ignored in the expression.

These rules are crucial for determining the identifiability of causal effects from data, allowing researchers to reason in causal relationships from a combination of experimental and observational data.

d-Separation is a criterion used in the analysis of Bayesian networks to determine whether a set of variables X_1 is independent of another set of variables X_2, given a third set of variables X_3. This concept is foundational in understanding the flow of causal influence in graphical models and helps in deciding whether a path between two variables is "blocked" or not by conditioning on other variables. The basic rule of d-separation is: A path between two variables is blocked if it includes an intermediate variable that is a collider and is not conditioned on or a non-collider that is conditioned on. Here, a **collider** is a variable that has inward arrows from two other nodes (i.e., $X_1 \to X_2 \leftarrow X_3$), whereas a **non-collider** does not meet this criterion.

3 Developing Causal Chain of Prompting

To develop C2P framework, we aim to extract the adjacency matrix of variables in a given premise. The adjacency matrix acts as equivalent to and instead of the causal DAG in Pearl's work leading to answering the causal reasoning question. In developing the Causal Chain of Prompting, we use a few steps similar to the PC algorithm for the given premise.

3.1 Causal Chain of Prompting (C2P)

The C2P frameworks consist of 5 main sub-tasks as follows, for learning and reasoning cause and effect relation for the given premise;

1. Prompting to extract the random variables in the provided data.
2. Prompting to extract all the cause and effect relations along with all conditional and unconditional relations among the random variables specifically mentioned in the given premise.
3. Prompting to create an initial adjacency matrix where all elements are 1, except for the diagonal elements and the elements corresponding to the cause-and-effect relationships specifically mentioned in the given premise (extracted in sub-task 2).
4. Prompting of the conditional and unconditional independency valuation and identifying the colliders, step by step, to extract the causal PDAG.
5. Prompting for cause-and-effect questions or hypotheses.

Each sub-task in C2P includes multiple steps and overall, to execute the frameworks, 9 main prompts have to be completed. The exact prompts for all the steps are provided in the Appendix, A.1.

3.2 Few-shot learning agenda

LLMs have powerful zero-shot capabilities, yet they might struggle with complex tasks with insufficient examples in the training process. In these situations, few-shot learning is a versatile and efficient technique for in-context learning that can be used to quickly adapt LLMs to new tasks and significantly enhance their performance (Min et al., 2022; Touvron et al., 2023). This approach involves providing several examples with desired answers, which condition the LLMs to produce the correct response for new instances with similar patterns. The few-shot learning process of C2P is similar to the previous subsection, with the prompts being based on the more abstract prompts presented there (Due to the token limitation). The prompts and an example of the given story are included in the Appendix, A.2. Depending on the token limitation of employed LLMs, the number of given examples (shots) varies. Comparing the token limitation of GPTs, we use GPT-4 Turbo, which has a 30000 token limit.

Steps from other constraint-based methods can be used instead or along with some steps of the PC algorithm in C2P methods, comprehensively discussed in (Glymour et al., 2019). For instance, steps in the FCI method can be used in cases where the causal sufficiency assumption is violated, i.e., latent variables and selection bias may be present (Spirtes et al., 1995). Additionally, it is important that the classifiers proposed by Ceraolo et al. (2024) that formalize the definition of causal questions and establish a taxonomy for finer-grained classification can be used before our framework to identify the causal question as a prerequisite of our method.

4 Results of the experiment for featuring LLMs with Causal Reasoning

4.1 Experimental setup

The experimental setup is divided into two data groups: synthetic dataset and realistic scenarios.

In a first synthetic dataset, and similar to the experimental setup of (Jin et al., 2023b), we set up the following list of LLMs for the experiments on the CORR2CAUSE dataset. Various types of LLMs are tested in (Jin et al., 2023b) and we only include powerful LLMs in our experiment to avoid reproducing the same results. To test existing LLMs, we first include two BERT-based NLI models in the transformers library (Wolf et al., 2020): BART (Lewis et al., 2019), DistilBART (Shleifer & Rush, 2020). We also evaluate the recent, more efficient models, LLaMa3-8B (Touvron et al., 2023) and LLaMa3-70B (Taori et al., 2023). We assess the GPT-3.5 (i.e., ChatGPT), and the latest GPT-4 (OpenAI, 2023) by April 2023, using the OpenAI API (https://openai.com/api/) with temperature 0. We employed the most recent GPTO as well. Then, we assess C2P on a proposed synthetic dataset, similar to the one in (Jin et al., 2023a), but more complex and with less information provided on the premises. Then we test the reliability of C2P in different numbers of nodes. Due to the limitation of tokens for different versions of GPTs, we employed GP4-Turbo which has a 30000 maximum token limit. We selected 50 samples at random from the corr2cause dataset for our experiment, ensuring a balanced distribution with 50 "Yes" and 50 "No" answers. In the corr2cause dataset, the majority of responses are "No", as a result, they used the F1 score as the primary metric for accuracy. Our experiment design enhances both the realism and comprehensibility of the results. These samples explore reasoning within three causal scenarios: direct cause-and-effect relationships (Fig. 2 i), indirect cause-and-effect relationships (Fig. 2 ii), and the presence of an effect due to two causes (Fig. 2 iii).

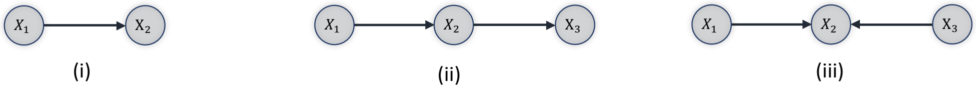


Figure 2: All possible in cause and effect relations. i. X_1 directly causes X_2 (X_2 is directly effect of X_1). ii. X_1 indirectly causes X_3 (X_3 is indirectly effect of X_1). iii. X_1 and X_3 are causes of X_2 (X_2 is common effect of X_1 and X_3)

To evaluate the proposed framework on natural stories, we utilized the state-of-the-art large language model, GPT-4, which excels at story generation. We crafted detailed instructions in the prompts and generated approximately 20 stories for our case study in fields such as healthcare and medicine, economics, education, social sciences, environmental science, and marketing, all emphasizing the importance of causality. This process is similar to the one presented in (Jin et al., 2023b), but our generated stories are more complex. Our approach can be tested using the examples of the Simpson's Paradox discussed in (Pearl & Mackenzie, 2018). However, since these examples are already included in the training data of current LLMs, the models simply repeat the correct answers based on that training data, similar to the parrot study in (Zečević et al., 2023). Consequently, our generated natural stories replicate these examples in a manner that the LLM cannot address within its existing training data.

Additionally, we assess the coevolution of supermassive black holes (SMBHs) and their host galaxies using our proposed framework, replicating the results provided in (Pasquato et al., 2023).

4.2 Evaluation of the C2P on synthetic dataset

Results of the C2P on CORR2CAUSE dataset: In Table 1, we show the performance of the more powerful LLMs in the cause-effect task reported in (Jin et al., 2023b) compared to the results of the C2P framework. According to Table 1, performing pure causal reasoning tasks presents a significant challenge for all existing LLMs, even when it is asked to think step by step similar to (Wei et al., 2022). Notably, several models perform worse than a random guess, indicating a complete failure in this specific task of pure causal reason. Nevertheless, applying our proposed framework to GPT-4 Turbo has boosted at least 33% in the model's ability to respond to causal inquiries, suggesting a substantial improvement over the existing

Table 1: Main results: Comparison of using C2P frameworks compared to the existing LLMs

Models	**F1**	**Precision**	**Recall**	**Accuracy**
Random Baselines				
Random (Proportional)	13.5	12.53	14.62	71.46
Random (Uniform)	20.38	15.11	31.29	62.78
BERT-Based Models				
DistilBART MNLI	26.74	15.92	83.63	30.23
BART MNLI	33.38	31.59	35.38	78.50
LLaMa-Based Models				
LLaMa3-8B	48.97	51.06	47.05	51
LLaMa3-70B	70.8	60.7	56.6	60
GPT-Based Models				
GPT-3.5	48.8	55	44	54
GPT-4	48.4	53.3	48	53
GPT-4O	57.7	59.5	56	59
C2P on GPT-4 Turbo	**92.78**	**95.74**	**90**	**93**
Few-shot learned GPT-4 Turbo	**81.06**	**82.97**	**79.24**	**81**

capabilities of LLMs. It is important to note that the primary factor contributing to the discrepancy between the LLaMa and GPT results and those reported in (Jin et al., 2023b) stems from the distribution of the provided premise. This table indicates that when reasoning questions are posed to LLMs, the responses slightly tend towards the "No" answer. More interestingly, in cases where existing models without C2P responded correctly, their explanations often revealed flawed logic. This highlights the randomness of their responses more clearly.

Robustness analysis: To assess the robustness of the proposed prompts, Table 2 shows the computed accuracy computed in each sub-task of implementing C2P on GPT-4 for different numbers of variables in the given premise.

Table 2: Accuracy by Number of Variables and sub-tasks

Number of variables	**Accuracy**				
	First sub-task	Second sub-task	Third sub-task	Fourth sub-task	Fifth sub-task
3 variables	100%	100%	100%	100%	99%
4 variables	100%	100%	100%	98%	83.2%
5 variables	100%	100%	100%	88.3%	75.2%
6 variables	100%	100%	100%	79.1%	70.1%

4.3 Evaluation of the C2P on natural stories

To evaluate the proposed framework on natural stories, we utilize the same process presented in (Jin et al., 2023b) and generate premises. The result of applying the C2P framework, both step-by-step and few-shot learned, is provided in Table 3 and compared to the result of GPTs. This table clearly demonstrates that the existing method responds randomly to reasoning prompts. It is important to note that even when GPTs provide the correct final answer, the reasoning process leading to that answer is still incorrect. In contrast, GPT-4 Turbo, enhanced with step-by-step prompting of the C2P framework and few-shot learned C2P, boosts the accuracy of these models in similar scenarios involving natural stories.

Table 3: Result of applying the C2P framework and GPTs, TP: True Positives, FP: False Positives, TN: true negatives, FN: False Negatives

Models	**TP**	**FP**	**TN**	**FN**
GPTs				
GPT-4.5	2	2	3	3
GPT-4	2	2	3	3
GPT-4o	3	2	3	3
Featured GPT				
C2P on GPT-4 Turbo	**5**	**0**	**5**	**0**
Few-shot learned GPT-4 Turbo	**4**	**0**	**5**	**1**

Additionally, Table 4 provides an example of such natural stories and demonstrates the result of applying a few-shot learned GP4-Turbo with C2P to it. This showcases how GPT models (including versions 3.5, 4, 4 Turbo, and 4o) fail to respond correctly. It is important to note that the few-shot learning process of GPT-4 Turbo was performed using only six examples from the CORR2CAUSE dataset, without including any natural story examples. Note that there are no natural stories employed in the few-shot learning process.

Table 4: Example of a natural story and the result of few-shot learned GPT-4-Turbo with C2P

Example of true story	**Results of few-shot learned GPT-4 Turbo with C2P**
Premise: eating junk food (A), obesity (C), and watching television (B) have relations with each other. There is a correlation between eating junk food and obesity, and between watching television and obesity. However, eating junk food and watching television are independent from each other. **Hypothesis**: Eating junk food directly affects obesity	Step 1: `Number of random variable:3` `Names of random variable: A, B, C`

	Step 2: `All of Statistical Relations:{`
	`Dependencies: [[A, C], [B, C]],`
	`Unconditional Independencies: [[A, B]],`
	`Conditional Independencies: []}`

	Step 3: `A: {A: 0, B: 1, C: 1},`
	`B: {A: 1, B: 0, C: 1},`
	`C: {A: 1, B: 1, C: 0}`

	Step 4: `A: {A: 0, B: 0, C: 1},`
	`B: {A: 0, B: 0, C: 1},`
	`C: {A: 1, B: 1, C: 0}`

	Step 5: `A: {A: 0, B: 0, C: 1},`
	`B: {A: 0, B: 0, C: 1},`
	`C: {A: 1, B: 1, C: 0}`

	Step 6: `C: [A, B]`
	Step 7: `C: [A, B]`

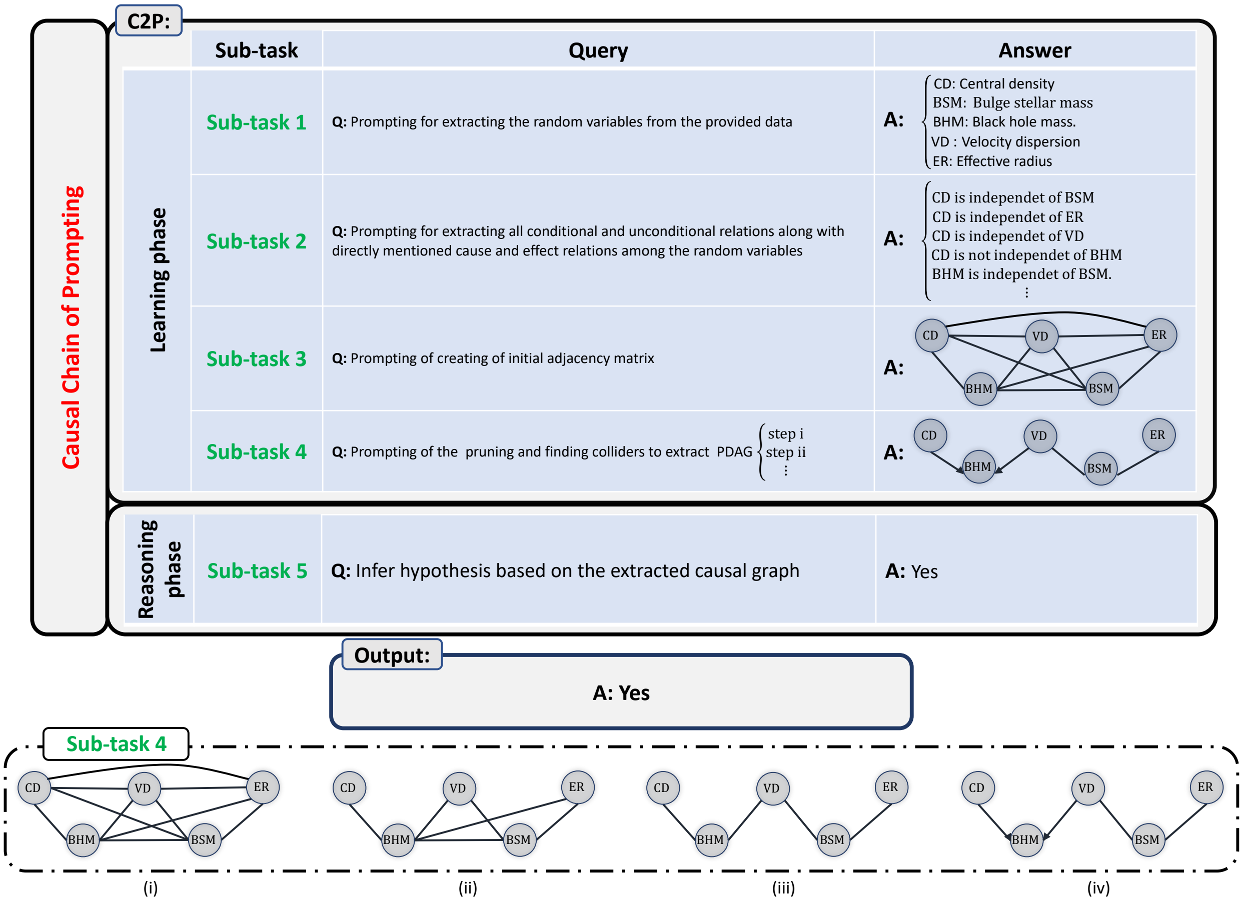
	Step 8: `A: {A: 0, B: 0, C: 1},`
	`B: {A: 0, B: 0, C: 1},`
	`C: {A: 0, B: 0, C: 0}`

	Step 9: `Checking matrix[A][D] = 1 and matrix[D][A] = 0. According to rule 2, this suggests` A is a direct cause of D, or D is a direct effect of A.

4.4 Evaluation of the C2P framework on the (co)evolution of supermassive black holes and their host galaxies

In astrophysics, experiments are impossible. Thus, causal reasoning must rely exclusively on observational data. Similar to the study in (Pasquato et al., 2023), we (co)evolute supermassive black holes (SMBHs) and their host galaxies. We implement C2P on GPT-4 Turbo to a comprehensive catalog of galaxy properties to obtain a PDAG and then reason the causal questions.

Fig. 3 shows the results of each sub-tasks of implementing C2P According to the extracted PDAG, many questions such as the causal given can be answered which is too complicated for the existing LLMs to answer. It is important to note that by implementing C2P, not only many causal questions can be answered with this approach, but it is easy to show which causal questions can not be answered based on the given premise and which more information has to be given to be able to causally reason. More importantly, it is worth noting that this is not the limitation of the C2P framework, any rational agent needs such information to answer such causal questions.



Sub-task	Query	Answer
Sub-task 1	**Q:** Prompting for extracting the random variables from the provided data	**A:** CD: Central density BSM: Bulge stellar mass BHM: Black hole mass. VD : Velocity dispersion ER: Effective radius
Sub-task 2	**Q:** Prompting for extracting all conditional and unconditional relations along with directly mentioned cause and effect relations among the random variables	**A:** CD is independet of BSM CD is independet of ER CD is independet of VD CD is not independet of BHM BHM is independet of BSM. ⋮
Sub-task 3	**Q:** Prompting of creating of initial adjacency matrix	**A:**
Sub-task 4	**Q:** Prompting of the pruning and finding colliders to extract PDAG step i step ii ⋮	**A:**
Sub-task 5	**Q:** Infer hypothesis based on the extracted causal graph	**A:** Yes

Figure 3: Prompts (**Q**) and results (**A**) of sub-tasks application of the C2P framework to real-world complex scenarios and steps of sub-task 3 for the given premise.

5 Discussion, Limitations and Future Work

Causal reasoning and identifiability: Having a PDAG of an SCM of a causal mechanism is the most effective way to achieve this with observational data. Based on this, two main questions arise: Given a PDAG, under what conditions can we make causal reasoning? This involves determining the necessary assumptions and data required to address a causal question—this issue is known as identification. It is generally possible to answer some cause-effect questions based on the PDAG, represented by a causal graph, and lower-level data. These questions are mainly the ones that are related to the part where we have extracted a collider, or if the question is just about the existence of a cause-effect relation; in some cases, these questions can be answered as well. However, it is almost impossible to answer all the causal questions based on the given information, as discussed in (Bareinboim et al., 2022; Pearl & Bareinboim, 2022). As a result, the second question is, what else is needed to answer all the causal questions? In such cases, interventional data is necessary to explicitly extract the direction of causal relations in a mechanism. To do so, while a naive baseline approach would require $O(n^2)$ interventions, various methods have been proposed, such as those in (Kocaoglu et al., 2017; Choo & Shiragur, 2023; von Kügelgen et al., 2024; Squires & Uhler, 2023), for cases in different situations. However, one clear thing is that it is not possible to reason all the causal questions. For instance,

- Directions cannot be discovered: PDAGs typically include both directed and undirected edges. The undirected edges indicate uncertainty about the direction of causality. It is important to note that without further data or assumptions, it is not possible to definitively determine the causal direction for these relationships with any method.
- Full Causal Path Analysis: While PDAGs can indicate possible paths between variables, they may not fully reveal which paths are indeed causal and which are due to confounding or indirect effects. Questions about specific causal pathways can thus be hard to answer definitively.
- Predictions Under Interventions: Questions about the outcomes of hypothetical or actual interventions on one or more variables (do-calculus questions) often require a fully specified causal model. PDAGs, with their partial specification, might not support detailed predictions under interventions without resolving the ambiguities in causal direction.

Causal reasoning is a broad and complex task in AI and LLMs. While current machine learning methods struggle with extracting causal structures and subsequently reasoning causally, this problem is even more intricate in the context of LLMs. The primary reason for this complexity is the inherent structure of LLMs and their reliance on the concept of attention. To enable LLMs to reason causally, the first step is to identify the causal question. Recently, a comprehensive database was compiled, and an AI model was trained to identify causal questions (Ceraolo et al., 2024). This model can be used prior to applying our framework to determine whether a causal reasoning process has to be activated or not. Additionally, more comprehensive examples and scenarios need to be generated to aid in the learning process of an LLM. Due to the vast variety of causal scenarios and the token limitations of current models, our approach must be integrated into the learning or fine-tuning process of an LLM. This ensures that the model can effectively learn and apply causal reasoning within its capabilities.

Scaling Instruction-Finetuned Language Models: LLMs primarily function through pattern recognition without the capability to recognize causal relationships, as indicated by prior research (). This limitation highlights the importance of implementing these models with causal analysis techniques to enhance their causal reasoning ability. Previous attempts to boost LLMs' causal reasoning involved using external tools or manually crafted prompts, which, although insightful, are impractical for easy integration. Our study introduced an algorithm that leverages LLMs' inherent abilities to foster causal reasoning autonomously. Our findings demonstrate a marked increase in the accuracy of LLMs' causal understanding, even in scenarios involving complex causal interactions. Additionally, incorporating a chain of thought into LLMs' training has been proven to enhance performance across various tasks (Chung et al., 2024). Based on the extracted results on the benchmark and provided datasets in our study, and the results of the few-shot learning process, we strongly hypothesize that our approach could similarly benefit the training of language models, particularly in boosting their causal reasoning, even in smaller models that would lead to a revelation in LLMs. This enhancement could be realized using a carefully selected set of examples or synthetic samples created by LLMs, following the guidelines outlined in our algorithm.

Future works As demonstrated in the results section, C2P has significantly enhanced the reasoning capabilities of LLMs. We have shown that extracting the adjacency matrix of the variables in a given premise can function as the DAG in Pearl's works, leading to improved reasoning ability. Our experimental results of few-shot learned LLMs indicate a significant increase in accuracy for tasks where users inquire about direct cause-and-effect relationships or common cause questions. However, the models still struggle with indirect cause-and-effect questions. Pearl demonstrated that the concepts of a DAG and d-separation complement each other in causal reasoning tasks. To address the low accuracy in answering indirect causal reasoning questions, more sub-tasks can be introduced to LLMs to perform d-separation in a more structured manner. As mentioned, due to the token limitation of LLMs, we were able to use only six or even fewer examples in the few-shot process of GPT-4-Turbo. As a result, our proposed framework can be seamlessly integrated into the fine-tuning or learning process of an LLM by introducing more extensive examples with more complex scenarios. Given the accuracy achieved in few-shot learned LLMs, the fine-tuned or trained LLMs with our approach can achieve highly accurate results, effectively addressing the causal reasoning bottleneck of LLMs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Mayank Jindal, Dushyant Singh Sengar, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. Cause and effect: Can large language models truly understand causality? *arXiv preprint arXiv:2402.18139*, 2024.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL `https://doi.org/10.1145/3501714.3501743`.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

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

Roberto Ceraolo, Dmitrii Kharlapenko, Amélie Reymond, Rada Mihalcea, Mrinmaya Sachan, Bernhard Schölkopf, and Zhijing Jin. Causalquest: Collecting natural causal questions for ai agents. *arXiv preprint arXiv:2405.20318*, 2024.

Davin Choo and Kirankumar Shiragur. Subset verification and search algorithms for causal dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 4409–4442. PMLR, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Amir Feder, Yoav Wald, Claudia Shi, Suchi Saria, and David Blei. Causal-structure driven augmentations for text ood generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*, 2023.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

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 preprint arXiv:2306.05836*, 2023b.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*, 2024.

Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*, 2023.

Elahe Khatibi, Mahyar Abbasian, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. Alcm: Autonomous llm-augmented causal discovery framework. *arXiv preprint arXiv:2405.01744*, 2024.

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

Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. *Advances in Neural Information Processing Systems*, 30, 2017.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, et al. Large language models and causal inference in collaboration: A comprehensive survey. *arXiv preprint arXiv:2403.09606*, 2024.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint arXiv:2406.02061*, 2024.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

Mario Pasquato, Zehao Jin, Pablo Lemos, Benjamin L Davis, and Andrea V Macciò. Causa prima: cosmology meets causal discovery for the first time. *arXiv preprint arXiv:2311.15160*, 2023.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl and Elias Bareinboim. *External Validity: From Do-Calculus to Transportability Across Populations*, pp. 451–482. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL `https://doi.org/10.1145/3501714.3501741`.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect.* Basic books, 2018.

Derek C Penn and Daniel J Povinelli. Causal cognition in human and nonhuman animals: A comparative, critical review. *Annu. Rev. Psychol.*, 58:97–118, 2007.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1991.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. Crab: Assessing the strength of causal relationships between real-world events. *arXiv preprint arXiv:2311.04284*, 2023.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

Sam Shleifer and Alexander M Rush. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*, 2020.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 499–506, 1995.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Chandler Squires and Caroline Uhler. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, 23(5):1781–1815, 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*, 2023.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias B Khalil. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *arXiv preprint arXiv:2305.18354*, 2023.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

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

Jiaqi Zhang, Joel Jennings, Cheng Zhang, and Chao Ma. Towards causal foundation model: on duality between causal inference and attention. *arXiv preprint arXiv:2310.00809*, 2023.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

A Appendix

A.1 Prompts for C2P frameworks

The proposed prompts for step-by-step proceeding with C2P are as follows:

1. **Prompting for extracting the random variables from the provided data.**
   - (i) Please give me the number of random variables in the given premise
   - (ii) Write the names of all random variables.
2. **Prompting to extract all the cause and effect relations along with all conditional and unconditional relations among the random variables specifically mentioned in the given premise**
   - If 2 random variables, for instance, R_1 and R_2, are independent, write it in this form: "R_1 is independent of R_2". If there exist 2 random variables, for instance, R_1 and R_2, are conditionally independent given a third random variable, for instance, R_3, write it in this form: "R_1 and R_2 are independent given R_3". If two random variables, for instance, R_1 and R_2, are specially mentioned to have cause and effect relation, write it in this form: "R_1 is the cause of R_2".
3. **Prompting to create an adjacency matrix where all elements are 1, except for the diagonal elements and the elements corresponding to the cause-and-effect relationships specifically mentioned in the given premise.**
   - In this phase, each random variable is treated as a node within a fully connected undirected graph. Then, for each pair, for instance, R_1 and R_2, presented in the form: "R_1 is the cause of R_2", set the element in ["R_2", "R_1"] in the adjacency matrix to 0.
4. **Prompting of the conditional and unconditional independency valuation and identifying the colliders, step by step, to extract the causal PDAG.**
   - (i) Update the adjacency matrix based on the specified unconditional independencies between random variables. Each pair of variables that is declared independent should have their corresponding value set to zero in the adjacency matrix. The initial adjacency matrix and the list of independencies are provided below. Please ensure all independencies are correctly reflected in the updated matrix. Instructions: - For each pair of variables listed as independent, set their corresponding entries in the adjacency matrix to 0.
   - (ii) Update the adjacency matrix based on the specified conditional independencies between random variables. Each pair of variables that is declared independent should have their corresponding value set to zero in the adjacency matrix. The initial adjacency matrix and the list of independencies are provided below. Please ensure all independencies are correctly reflected in the updated matrix. Instructions: - For each pair of variables listed as independent given other variable(s), set their corresponding entries in the adjacency matrix to 0.
   - (iii) Task: Given an initial adjacency matrix, follow these steps: Step 1: Identify all rows (key values) in the matrix where there are two or more than two columns with the value "1" in them. For each identified row, find all pairs of different columns where the values are "1".Ensure to exclude rows that do not contain any pairs from the results. Step 2: Display these pairs, "All Pairs", where each row name is key, and the value is a list of column names that are identified in Step 1.

(iv) Given the "All Pairs" and the list of independencies, follow these instructions step by step: Instruction: For each key in "All Pairs", delete all the pairs that are not mentioned as independent in the "independencies" list and return other with all their values. The "All Pairs" contains pairs of elements associated with each key. The goal is to update this by removing pairs that are not mentioned as independent. The list of independencies provides information about which pairs are independent of each other.

(v) Given the initial adjacency matrix represented and the "All Pairs" list, for each key-value pair ("R") in "All Pairs", modify the initial adjacency matrix as follows: 1- Set the value in the "C_1" row and "R" column to 0: ("C_1", "R") $= 0$. 2- Set the value in the "C_2" row and "R" column to 0: ("C_2", "R") $= 0$.

5. **Prompting for cause-and-effect questions or hypotheses**

   - To extract and understand causal relations in the adjacency matrix: For each specified variable "R" and "C", for instance, that are listed in the adjacency matrix: - If matrix entry at ["R", "C"] $= 1$ and ["C", "R"] $= 1$, then the causal direction between "R" and "C" is undetermined. - If matrix entry at ["R", "C"] $= 1$ and ["C", "R"] $= 0$, then "R" is a direct cause of "C" or "C" is a direct effect of "R". - If matrix entry at ["R", "C"] $= 0$ and ["C", "R"] $= 1$, then "C" is a direct cause of "R" or "R" is a direct effect of "C". If two variables directly affect a third variable, the first two variables are common causes, and the third variable is a collider. Evaluate the hypothesis based on the given partially directed acyclic graph presented as an adjacency matrix with the given Instruction.

A.2 Prompts for few-shot learning of C2P

An example of given prompts for the few-shot learning process of C2P is as follows:

Premise: Suppose there is a closed system of 5 variables, A, B, C, D, and E. All the statistical relations among these 5 variables are as follows: A correlates with C. A correlates with D. A correlates with E. B correlates with D. B correlates with E. C correlates with D. C correlates with E. D correlates with E. However, A is independent of B. A and B are independent given C. B is independent of C. B and C are independent given A. C and E are independent given A, B, and D.

Hypothesis: There exists at least one collider (i.e., common effect) of A and B.

- **"Subtask 1"**- The number of random variables and their names in the given premise in JSON format:

  Output: "number of random variables: 5, "names of random variables": ["A", "B", "C", "D", "E"]

- **"Subtask 2"**- All the dependencies, conditional and unconditional independencies between all random variables extracted in "subtask 1":

  Output: {"All of Statistical Relations": {"Dependencies": [["A", "C"], ["A", "D"], ["A", "E"], ["B", "D"], ["B", "E"], ["C", "D"], ["C", "E"], ["D", "E"]], "Unconditional Independencies": [["A", "B"], ["B", "C"]], "Conditional Independencies": [["A", "B"], ["B", "C"], ["C", "E"]]}}

- **"Subtask 3"**- The adjacency matrix of all random variables extracted in "subtask 1" where each random variable is treated as a node within a fully connected undirected graph:

  Output: {"A": {"A": 0, "B": 1, "C": 1, "D": 1, "E": 1}, "B": {"A": 1, "B": 0, "C": 1, "D": 1, "E": 1}, "C": {"A": 1, "B": 1, "C": 0, "D": 1, "E": 1}, "D": {"A": 1, "B": 1, "C": 1, "D": 0, "E": 1}, "E": {"A": 1, "B": 1, "C": 1, "D": 1, "E": 0}}

- **"Subtask 4"**- Update the adjacency matrix extracted in the output of "subtask 3" based on the specified unconditional independencies between random variables. Each pair of variables that are declared independent should have their corresponding value set to zero in the adjacency matrix. - For each pair of variables listed as unconditional independent in "subtask 2", we set their corresponding entries in the adjacency matrix to 0. - We do not change any other entries except those specified by the independence.

Output: {"A": {"A": 0, "B": 0, "C": 1, "D": 1, "E": 1}, "B": {"A": 0, "B": 0, "C": 0, "D": 1, "E": 1}, "C": {"A": 1, "B": 0, "C": 0, "D": 1, "E": 1}, "D": {"A": 1, "B": 1, "C": 1, "D": 0, "E": 1}, "E": {"A": 1, "B": 1, "C": 1, "D": 1, "E": 0}}

- **"Subtask 5"**- Update the adjacency matrix in the output of "Subtask 4" based on the specified conditional independencies between random variables extracted in "subtask 2". Each pair of variables that are declared conditional independent should have their corresponding value set to zero in the adjacency matrix. - For each pair of variables listed as conditionally independent given other variable(s), we set their corresponding entries in the adjacency matrix to 0. - We do not change any other entries except those specified by the conditional independence.

  Output: {"A": {"A": 0, "B": 0, "C": 1, "D": 1, "E": 1}, "B": {"A": 0, "B": 0, "C": 0, "D": 1, "E": 1}, "C": {"A": 1, "B": 0, "C": 0, "D": 1, "E": 0}, "D": {"A": 1, "B": 1, "C": 1, "D": 0, "E": 1}, "E": {"A": 1, "B": 1, "C": 0, "D": 1, "E": 0}}

- **"Subtask 6"**: In this subtask, for each key value in the "Adjacency Matrix" mentioned in the output of "Subtask 5", the task is to extract only pairs of columns with values "1" step by step according to the provided steps and put it in "Candidates" dictionary: follow these steps: Step 1: Identify all rows (key values) in the matrix where there are at least two or more columns with the value "1". For each identified row, find all pairs of different columns where the values are "1". Ensure to exclude rows that do not contain any pairs from the results. Make sure to check each row individually and include all valid pairs for every row. Step 2: Display these pairs in a simplified JSON format, where each row name is a key, and the pair values are a list of column names that are identified in step 1. If there are no such rows, show an empty JSON object. Step 3: Do not include any rows with one or zero columns with "1" values in the output. Example output format: Candidates: { "row1 name": [["1value column1 name", "1value column2 name"], ["1value column1 name", "1value column3 name"], ["1value column2 name", "1value column3 name"]], "row2 name": [["1value column5 name", "1value column6 name"]], ... } The values in each row have to be pairs and the output cannot be as follows: { "row1 name": [["1value column1 name"], ["1value column1 name"], ["1value column2 name", "1value column3 name"]], "row2 name": [["1value column5 name"]], ... } Please provide only the desired output formatted exactly as shown in the example without any further explanation.

  Output: {"A": [["C", "D"], ["C", "E"], ["D", "E"]], "B": [["D", "E"]], "C": [["A", "D"]], "D": [["A", "B"], ["A", "C"], ["A", "E"], ["B", "C"], ["B", "E"], ["C", "E"]], "E": [["A", "B"], ["A", "D"], ["B", "D"]]}

- **"Subtask 7"**- In this subtask, given the output of "Subtask 6", "Candidates", and "Unconditional Independencies" in "Subtask 2", the task is to identify and extract all the pairs in the "Candidates" that are also present in the "Unconditional Independencies" step by step according to the provided steps. follow these steps: 1- For each pair in the "Candidates" list, check if it is present in the "Unconditional Independencies" list. 3-Only keep all the pairs from "Candidates" that are also present in "Unconditional Independencies". If a pair in "Candidates" is found in "Conditional Independencies", keep it. 4-Remove any pairs in "Candidates" that are not found in "Conditional Independencies". If a pair in "Candidates" is not found in "Unconditional Independencies", remove it. 5-Output the result as the modified "Candidates" dictionary without any additional text or explanation. Only the updated "Candidates" dictionary and nothing else.

  Output: {"D": [["A", "B"], ["B", "C"]], "E": [["A", "B"]]}

- **"Subtask 8"**- Given the adjacency matrix in the output of "Subtask 5" and the "All Pairs" list in the output of "Subtask 7", for each key-value pair ("R") in "All Pairs", we modify the adjacency matrix as follows: -For each key "R" and pair ("C1", "C2") in the candidates, change the values in key "R" and pairs ("C1", "C2") to zero. -Ensure that only the specified modifications are made, and all other entries in the adjacency matrix remain unchanged.

  Output: {"A": {"A": 0, "B": 0, "C": 1, "D": 1, "E": 1}, "B": {"A": 0, "B": 0, "C": 0, "D": 1, "E": 1}, "C": {"A": 1, "B": 0, "C": 0, "D": 1, "E": 0}, "D": {"A": 0, "B": 0, "C": 0, "D": 0, "E": 1}, "E": {"A": 0, "B": 0, "C": 0, "D": 1, "E": 0}}

- **"Subtask 9"**- In this subtask, given the "Adjacency Matrix" mentioned in the output of "Subtask 8", evaluate the correctness of the hypothesis step by step according to the provided rules. Rules: 1- If matrix["R"]["C"] = 1 and matrix["C"]["R"] = 1, then the causal direction between "R" and "C" is undetermined. 2- If matrix["R"]["C"] = 1 and matrix["C"]["R"] = 0, then "R" is a direct cause of "C", or "C" is a direct effect of "R". 3- If matrix["R"]["C"] = 0 and matrix["C"]["R"] = 1, then "C" is a direct cause of "R", or "R" is a direct effect of "C". 4- Test each variable in the matrix, for instance, "R", "R" is a collider if the matrix entries in ["R", "C1"] = 0, ["C1", "R"] = 1, ["R", "C2"] = 0, and ["C2", "R"] = 1 Perform it step by step and provide the final "Yes" or "No" answer.

  Output: Step 1: According to the rules, a node is a collider (common effect) if the matrix entries in ["R", "C1"] = 0, ["R","C2"] = 0, ["C1", "R"] = 1 and ["C2", "R"] = 1.

  Step 2: According to rule 4, "R" is a collider if the matrix entries in ["R", "D"] = 0, ["R", "E"] = 0, ["D", "R"] = 1 and ["E", "R"] = 1.

  Step 3: Checking each variable in the matrix

  Node C: ["C", "A"] = 1 and ["A", "C"] = 1, so C is not a common effect of A and B.

  Node D: ["D", "A"] = 0 and ["A", "D"] = 1, and ["D", "B"] = 0 and ["B", "D"] = 1, so D is a common effect of A and B.

  Node E: ["E","A"] = 0 and ["A","E"] = 1, and ["E","B"] = 0 and ["B", "E"] = 1, so E is a common effect of A and B.

  Step 4: So, the hypothesis that there exists at least one collider (i.e., common effect) of A and B is correct. The answer is "Yes".