Leveraging LLMs for Causal Inference and Discovery

Zhuofan Sun
School of Economics and Management
Tsinghua University
sunzf23@mails.tsinghua.edu.cn

Qingyi Li
Shenzhen International Graduate School
Tsinghua University
li-qy23@mails.tsinghua.edu.cn

Abstract

Causal inference is widely applied in the social sciences to analyze the effects of a specific treatment. Causal inference tools rely on uncovering the underlying causal graph in advance, a process known as causal discovery. Traditionally, constructing causal graphs has depended on expert domain knowledge; however, the rich knowledge embedded in large language models (LLMs) offers a promising alternative. Nevertheless, LLMs alone perform poorly in inferring complete causal graphs, primarily because they fail to account for the directed acyclic nature of causal graphs. To address this limitation, we propose a novel approach that combines LLMs with statistical causal discovery algorithms to better leverage the expert-like capabilities of LLMs. Experimental results demonstrate that the proposed method significantly improves the accuracy of causal ordering and effectively reduces errors in downstream causal effect estimation tasks.

1 Introduction

Estimating the causal effects of variables on an outcome is a fundamental challenge in diverse scientific disciplines, including epidemiology, economics, and atmospheric sciences. Inferring causal effects from observational data, however, remains a difficult task due to the dependence of effect estimates on the assumed causal graph. Although significant progress has been made in graph discovery algorithms, particularly in specific parametric contexts (Shimizu et al. (2006)Hoyer et al. (2008)Hyvärinen et al. (2010)Rolland et al. (2022)), studies on real-world datasets, such as those in atmospheric science and healthcare (Huang et al. (2021)Tu et al. (2019)), underscore the persistent challenges of reliably inferring causal graphs from data (Reisach et al. (2021)). Consequently, causal effect inference studies often depend on human experts to provide the causal graph.

In this paper, we propose a novel approach that leverages Large Language Models (LLMs) as virtual domain experts to automate the extraction of causal order, an essential component for causal effect inference. Building on the observation that the topological causal order of variables is sufficient for effect inference (Proposition 4.1), we argue that eliciting causal order is a more tractable and appropriate question for experts, whether human or machine. Unlike the identification of direct graph edges, which depends on the inclusion of other variables to account for direct and indirect effects, causal order depends solely on the variables in question.

For instance, consider the data-generating process: lung cancerß doctor visit ß positive X-ray. An expert might confirm a direct causal edge from lung cancer to positive X-ray when considered in isolation. However, when the observed variable set includes doctor visit, the correct graph excludes a direct edge between lung cancer and positive X-ray, instead reflecting mediation via doctor visit. Importantly, the causal order lung cancer $<$ positive X-ray remains consistent across both scenarios.

This distinction has critical implications for using LLMs to infer graph structures. Existing methods typically prompt LLMs to evaluate causal edges pairwise (Kıcıman et al. (2023)Long et al. (2022)). However, as demonstrated above, the accuracy of pairwise prompts is unreliable, as the response

Preprint. Under review.

depends on the broader variable set under consideration. Including all other variables in a prompt is computationally impractical for large graphs. Consequently, we argue that instead of attempting to infer the full graph structure, a more feasible and effective approach is to focus on inferring the causal order—a simpler and locally determined property that is unaffected by the availability of other variables. Moreover, causal order is sufficient for downstream tasks such as effect inference, obviating the need for the full graph structure.

To address these challenges, we introduce a triplet-based prompting strategy that outperforms pairwise prompting methods (Kıcıman et al. (2023)Willig et al. (2022)Long et al. (2022)) in determining causal order. The triplet prompt also reduces the likelihood of introducing cycles into the inferred graph.

Recognizing potential failure modes of LLMs, we propose two algorithms to integrate LLM-inferred causal order with existing graph discovery techniques. The first algorithm uses LLM-derived causal order to guide a constraint-based algorithm (e.g., PC) in orienting undirected edges, while the second incorporates LLM causal order as a prior in a score-based algorithm like CaMML. Our empirical evaluation on six benchmark datasets demonstrates that these LLM-enhanced algorithms significantly outperform baseline causal discovery methods in inferring causal order.

Our contributions are summarized as follows:

- We propose a novel triplet-based prompting strategy for estimating causal order using LLMs, offering a more reliable alternative to inferring graph structures directly.
- We introduce algorithms that combine LLM-derived causal order with existing discovery methods, enhancing their performance in causal inference tasks.
- Through comprehensive experiments, we show that our approach achieves superior results in causal order inference compared to existing methods.

2 Related Work

Historically, the fields of causal discovery (Glymour et al. (2019)Rolland et al. (2022)Teyssier and Koller (2005)Zheng et al. (2018)Lachapelle et al. (2020)) and causal effect inference (Pearl (2009)) have been studied independently. While conventional approaches rely on learning a full causal graph for effect inference (Hoyer et al. (2008)Mooij et al. (2016)Maathuis et al. (2010)Gupta et al. (2022)), our work demonstrates that this step can be simplified: causal order alone is sufficient for effect inference, eliminating the need to infer the entire graph structure.

Our study is closely related to recent efforts in Large Language Model (LLM)-based, knowledge-driven causal discovery (Kıcıman et al. (2023)Ban et al. (2023)Long et al. (2022)Willig et al. (2022)). Unlike traditional causal discovery algorithms that rely on statistical patterns in data, LLM-based methods leverage metadata, such as variable names, to predict causal structures. These approaches typically employ edge-wise prompts for each pair of variables, aggregating the results to infer causal graphs (Kıcıman et al. (2023)Long et al. (2023)Willig et al. (2022)). However, we argue that this approach has a fundamental limitation: the existence of an edge may depend on the presence or absence of other variables in the dataset. As such, causal order—being independent of other variable sets—is a more robust and suitable output to elicit from LLMs. To improve causal order inference, we propose a novel triplet-based prompting strategy, which represents an advance over pairwise prompting methods and may also be of broader interest for prompting LLMs in causal reasoning tasks.

Recognizing the potential for errors in LLM outputs, we propose a more principled approach that integrates LLMs with established causal discovery algorithms. While Long Long et al. (2023) enhance constraint-based algorithms for full graph discovery using LLM outputs, and Ban Ban et al. (2023) utilize LLMs as priors for score-based causal discovery methods, we extend this paradigm specifically to causal order estimation. To this end, we present LLM-adapted versions of both constraint-based and score-based algorithms, thereby combining the strengths of LLM-driven insights and data-driven discovery techniques.

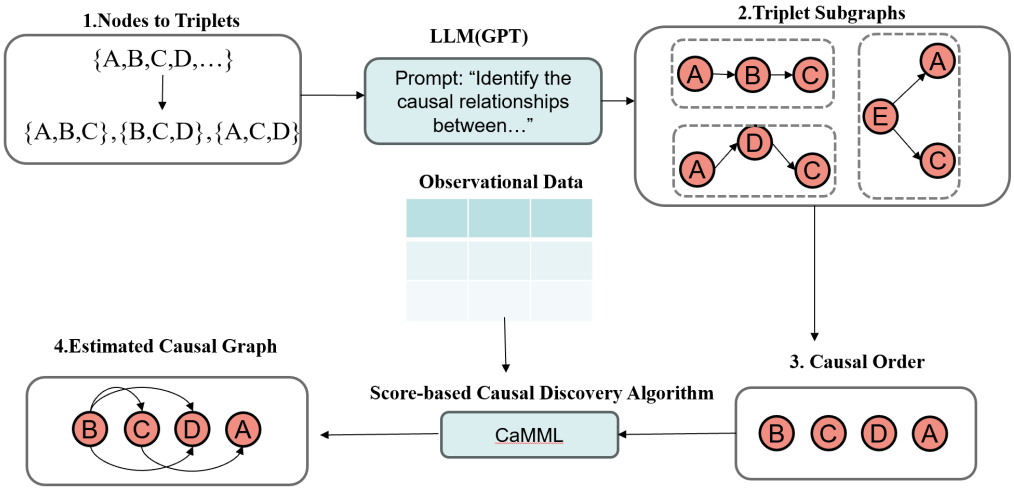

Figure 1: the method workflow

3 Problem Formulation

Let $G(X, E)$ represent a causal directed acyclic graph (DAG) consisting of a set of variables $X = \{X_1, \ldots, X_n\}$ and a set of directed edges E among the variables in X. A directed edge $X_i \to X_j \in E$ signifies the direct causal influence of the variable X_i on the variable X_j. Define $\text{pa}(X_i) = \{X_k \mid X_k \to X_i\}$ as the set of parents of X_i, and $\text{de}(X_i) = \{X_k \mid X_k \leftarrow \cdots \leftarrow X_i\}$ as the set of its descendants. A sequence π of variables in X is called a topological order if and only if for every edge $X_i \to X_j \in E$, $\pi_i < \pi_j$. This work focuses on a downstream application of causal graph discovery: causal effect inference. The *average causal effect* (ACE) of a variable X_i on another variable X_j is defined as$\text{ACE}_{X_j|X_i} = E[X_j \mid \text{do}(X_i = x_i)] - E[X_j \mid \text{do}(X_i = x_i^*)]$, where X_i is referred to as the treatment variable, and X_j is the target variable. The operator $\text{do}(X_i = x_i)$ denotes an external intervention that sets X_i to the value x_i. The interventional expectation $E[X_j \mid \text{do}(X_i = x_i)]$ differs from the conditional expectation $E[X_j \mid X_i = x_i]$ as the former reflects the effects of actively setting X_i, rather than conditioning on its observed value. To estimate $E[X_j \mid \text{do}(X_i = x_i)]$ from observational data, the *back-door adjustment* formula is employed. In a given DAG G, a set of variables Z satisfies the *back-door criterion* relative to the treatment and target pair (X_i, X_j) if:

(i) No variable in Z is a descendant of X_i, and

(ii) Z blocks every path between X_i and X_j that contains an arrow pointing into X_i.

Here, a path in a causal DAG is defined as a sequence of unique vertices $X_i, X_{i+1}, \ldots, X_j$ where each consecutive pair (X_k, X_{k+1}) is connected by a directed edge ($X_k \to X_{k+1}$ or $X_{k+1} \to X_k$). If a set of variables Z satisfies the back-door criterion relative to (X_i, X_j), the interventional expectation $E[X_j \mid \text{do}(X_i = x_i)]$ can be computed using the formula as established in Theorem 3.3.2 of Pearl (2009) $E[X_j \mid \text{do}(X_i = x_i)] = E_{z \sim Z}\big[E[X_j \mid X_i = x_i, Z = z]\big]$. To ensure the identifiability of causal effects, we assume the absence of latent confounding variables.

4 LLM-based Causal Discovery Algorithms

4.1 Method Overview

The process proceeds as follows:

- **Nodes to Triplets.** The initial set of variables (e.g., $\{A, B, C, D, \ldots\}$) is first decomposed into all possible three-variable combinations or triplets, such as (A, B, C), (B, C, D), and so forth. For each triplet, a language model (e.g., GPT-3.5-Turbo) generates a local causal

subgraph by identifying the pairwise and conditional causal relationships among the three variables.

- **Triplet Subgraphs.** The resulting causal subgraphs for all triplets are collected, each specifying edges (e.g., $A \rightarrow B$, $B \rightarrow C$) derived from the triplet structure. These subgraphs serve as building blocks for global causal inference.

- **Majority Voting for Edge Orientation.** To integrate the triplet-level causal relationships into a global structure, a majority voting mechanism is employed to resolve the directionality of edges between pairs of nodes. For instance, if $A \rightarrow B$ is inferred four times, $A \leftarrow B$ twice, and "no connection" once, the edge $A \rightarrow B$ is selected. In cases where ties occur (e.g., equal votes for $C \rightarrow D$ and $C \leftarrow D$), GPT-4 is used as an expert system to break the tie and determine the final edge orientation.

- **Getting Final Causal Order.** Once edge directions are resolved through majority voting and expert intervention, a global causal order of variables (e.g., A, B, C, D) is determined. This order corresponds to a fully oriented DAG, where edges indicate causal relationships between variables.

- **Using Causal Order as Prior for Discovery Algorithms.** The inferred causal order serves as a prior for classical causal discovery algorithms, such as the PC algorithm or CaMML, which further refine the causal structure using observational data. Specifically, the causal order helps these algorithms constrain the search space and identify the most plausible causal graph.

- **Using Final Graph for Downstream Causal Effect Inference.** The resulting causal graph is then utilized for downstream causal inference tasks, such as estimating causal effects of treatments on outcomes. Using back-door adjustment sets, valid estimations of causal effects are ensured, enabling reliable causal reasoning.

4.2 Prompt Technique Based on Triplets

Some existing literature has explored the use of LLMs as virtual experts to perform pair-wise causal discovery tasks. However, this approach cannot be directly applied to causal discovery. The reason is that the pair-wise method may lead to cycles in the resulting causal graph. In this paper, we propose a triplet-based prompt method to reduce the likelihood of cycle generation. By introducing the triplet structure, we not only clarify the causal relationships between variables but also effectively limit the potential circular dependencies that may arise during the inference process, thereby improving the accuracy and validity of the causal graph.

After splitting the nodes into triplets, we use a large model to simultaneously determine the relationships among the three nodes within each triplet. After obtaining the results for each triplet, we apply the majority principle for each edge, selecting the most frequent result (A influences B, B influences A, or no causal relationship between A and B). Finally, we determine the causal order based on the causal graph we obtained. During this process, the triplet structure effectively reduces the generation of cyclic dependencies and ensures the correctness of causal inference. Through this approach, we are able to obtain a more consistent and reliable causal graph, and further establish the causal order of the variables, providing a solid foundation for subsequent causal reasoning tasks. We adopt the ICL prompting strategy, and the template for this strategy is shown in the figure.

Task: Given three variable names and their descriptions, determine the causal relationships among them. Specifically, identify whether one variable influences another, or if there is a reciprocal or no influence.

You should output the relationship in the form of a causal chain (e.g., A $\rightarrow$ B $\rightarrow$ C) or indicate if there is no relationship between them.

Example:

Given Variables:

- **Variable A**: Temperature
  Description: The measure of the warmth or coldness of an environment. Typically measured in degrees Celsius or Fahrenheit.
- **Variable B**: Ice Melting
  Description: The process of ice turning into water when heated.
- **Variable C**: Water Evaporation
  Description: The process where water molecules transition from liquid to gas form, typically due to heat.

Causal Relationship:

Based on the descriptions, the relationships are as follows:

$$\text{Temperature} \rightarrow \text{Ice Melting} \rightarrow \text{Water Evaporation}$$

Given Variables:

- **Variable A**: [Name of Variable A]
  Description: [Brief description of what Variable A represents, including any relevant context]
- **Variable B**: [Name of Variable B]
  Description: [Brief description of what Variable B represents, including any relevant context]
- **Variable C**: [Name of Variable C]
  Description: [Brief description of what Variable C represents, including any relevant context]

Causal Relationship:

Figure 2: Prompt Template

4.3 Causal discovery with given causal order

Although LLMs ensure a certain level of predictive accuracy and the use of the triplet method effectively avoids the occurrence of cycles, this approach does not distinguish between direct and indirect effects. In contrast, some traditional causal discovery algorithms based on observation data can infer the causal relationships between variables from the statistical distributions of the data. We adopt causal order as an intermediate bridge, and by introducing causal order, we can better differentiate between direct and indirect effects, providing a clear path for further causal inference. This method combines the powerful language understanding capabilities of LLMs with the data-driven reasoning advantages of traditional algorithms, thus enabling more precise identification of causal relationships.

In this paper, we adopt the CAM algorithm. The underlying assumption of the CAM algorithm is that the link functions f_i have an additive structure. Following previous work, we perform sparse regression on each component and use hypothesis testing for additive models to decide on the existence of edges. This specific pruning process has already been mentioned in earlier algorithms.

In other words, the CAM algorithm is a pruning method that gradually eliminates impossible causal edges in a graph based on observation data. In the original CAM algorithm, the pruning starts from a fully connected graph, where all nodes are connected. However, given the causal order, we can initialize the pruning process with a semi-connected graph according to the causal order.

4.4 The Algorithm

Algorithm 1 Combining CAM and LLM to get $\hat{\pi}$ for a given set of variables.

Require: observation data $\mathcal{D}$, variables $\{X_1, \ldots, X_n\}$, Expert $\mathcal{E}$, CAM parameter θ
Ensure: Estimated topological order $\hat{\pi}$ of $\{X_1, \ldots, X_n\}$.

1: Step (I) $\hat{G} \leftarrow \mathcal{E}(X_1, \ldots, X_n)$
2: Step (II) estimated causal order$\hat{\pi} \leftarrow \hat{G}$
3: Step (III) $\hat{G} \leftarrow CAM(\hat{\pi}|\theta)$
4: Step (IV) $\hat{\pi} \leftarrow$ topological ordering of $\hat{G}$
5: **return** $\hat{\pi}$

5 Experiments and Results

5.1 Metric

We use the topological divergence between an estimated topological order $\hat{\pi}$ and the real graph adjacency matrix A as the metric. It's defined as $D_{\text{top}}(\hat{\pi}, A) = \sum_{i=1}^{n} \sum_{j:\hat{\pi}_i > \hat{\pi}_j} A_{ij}$;

Below, we show that D_{top} is a valid metric to check the correctness of the estimated causal effects. That is $D_{\text{top}}(\hat{\pi}, A) = 0$ is equivalent to obtaining the correct back-door adjustment set from $\hat{\pi}$.We start with the fact that the causal order is sufficient to find a valid back-door set.

Proposition 5.1 *For an estimated topological order $\hat{\pi}$ and a true topological order π of a causal DAG G with the corresponding adjacency matrix A, $D_{top}(\hat{\pi}, A) = 0$ iff $Z = \{X_k | \hat{\pi}_k < \hat{\pi}_i\}$ is a valid adjustment set relative to (X_i, X_j), $\forall \pi_i < \pi_j$.*

Proof. The statement of proposition is of the form $A \Leftrightarrow B$ with A being "$D_{\text{top}}(\hat{\pi}, A) = 0$" and B being "$Z = \{X_k | \hat{\pi}_k < \hat{\pi}_i\}$ is a valid adjustment set relative to (X_i, X_j), $\forall i, j$".We prove $A \Leftrightarrow B$ by proving (i) $A \Rightarrow B$ and (ii) $B \Rightarrow A$.

(i) Proof of $A \Rightarrow B$: If $D_{\text{top}}(\hat{\pi}, A) = 0$, for all pairs of nodes (X_i, X_j), we have $\hat{\pi}_i < \hat{\pi}_j$ whenever $\pi_i < \pi_j$. That is, causal order in estimated graph is same that of the causal order in true graph. Hence, $Z = \{X_k | \hat{\pi}_k < \hat{\pi}_i\}$ is a valid adjustment set relative to (X_i, X_j), $\forall i, j$.

(ii) Proof of $B \Rightarrow A$: we prove the logical equivalent form of $B \Rightarrow A$ i.e., $\neg A \Rightarrow \neg B$, the contrapositive of $B \Rightarrow A$. To this end, assume $D_{\text{top}}(\hat{\pi}, A) \neq 0$, then there will be at least one edge $X_i \to X_j$ that cannot be oriented correctly due to the estimated topological order $\hat{\pi}$. i.e., $\hat{\pi}_j < \hat{\pi}_i$ but $\pi_j > \pi_i$. Hence, to find the causal effect of X_i on X_l; $l \neq j$, X_j is included in the back-door adjustment set Z relative to (X_i, X_l). Adding X_j to Z renders Z an invalid adjustment set because it violates the condition (i).

□

At the same time, our paper also uses another commonly used metric, SHD(structural hamming distance), for comparison. In our subsequent experiments, we found that in terms of SHD performance, our algorithm exhibits relatively mediocre performance. However, it still outperforms other algorithms. SHD measures how many edge changes are needed to transform the estimated causal graph into the true causal graph.

6 Experiments and Results

6.1 causal discovery comparison

To evaluate the performance of LLM-based algorithms in inferring causal order, we conducted experiments on several benchmark datasets. These include datasets from the Bayesian Network repository (Scutari and Denis (2014)), such as Earthquake, Cancer, Survey, Asia, Asia-M (the modified version of Asia), and Child. Additionally, we selected a medium-sized subset graph from the Neuropathic dataset (Tu et al. (2019)), which is used for pain diagnosis. These datasets cover a range of different applications, allowing us to comprehensively assess the accuracy and robustness of the LLM algorithms in various complex causal inference tasks.

Table 1: Mean and std dev of D_{top}

Dataset	**PC**	**SCORE**	**ICA LINGAM**	**Direct LINGAM**	**NOTEARS**	**Ours**
			$N = 250$			
Earthquake	1.80±0.22	5.00±0.00	4.20±0.36	3.00±0.20	1.80±1.80	**1.00±0.00**
Cancer	0.20±0.05	5.00±0.5	5.00±0.4	1.80±0.00	3.8±0.60	**1.00±0.00**
Survey	5.00±0.40	4.00±0.50	5.00±0.00	8.00±0.50	3.20±0.40	**0.00±0.00**
Asia	3.05±0.78	1.00±0.50	3.00±0.00	4.23±0.17	2.40±0.15	**0.83±0.72**
Asia-M	7.15±1.33	7.00±1.18	7.80±0.88	5.27±1.01	3.44±0.48	**3.00±0.00**
Child	11.00±1.90	8.00±0.40	8.00±2.02	9.00±0.42	**5.66±0.30**	6.05±0.04
Neuropathic	9.72±0.6	5.97±0.18	3.00±0.15	10.00±0.00	9.00±0.00	**1.33±0.20**
			$N = 1000$			
Earthquake	1.23±0.45	3.78±0.12	3.34±0.23	2.56±0.18	1.51±0.07	**0.00±0.00**
Cancer	**0.00±0.00**	3.92±0.15	4.67±0.28	3.50±0.00	2.50±0.20	**0.00±0.00**
Survey	3.80±0.21	3.55±0.19	4.22±0.14	3.60±0.05	**0.00±0.00**	**0.00±0.00**
Asia	2.75±1.04	3.99±0.00	3.15±0.61	2.80±0.12	1.12±0.55	**0.12±0.05**
Asia-M	4.33±0.92	3.88±0.00	2.72±0.09	3.12±0.08	2.47±0.38	**1.47±0.38**
Child	8.92±0.95	7.99±4.00	7.09±1.48	8.64±1.01	**5.56±0.43**	5.88±0.43
Neuropathic	8.75±0.34	5.62±0.09	2.87±0.73	9.55±0.14	8.23±1.98	**1.23±1.98**
			$N = 10000$			
Earthquake	1.23±0.35	3.57±0.22	2.45±0.12	2.22±0.30	0.09±0.06	**0.00±0.00**
Cancer	**0.00±0.00**	4.12±0.08	3.27±0.14	2.95±0.09	4.23±0.16	**0.00±0.00**
Survey	2.45±0.22	3.78±0.13	4.12±0.09	2.67±0.18	1.09±0.15	**0.00±0.00**
Asia	2.04±0.33	3.97±0.05	4.15±0.08	3.89±1.22	3.67±0.37	**0.02±0.00**
Asia-M	1.92±0.48	2.95±0.00	2.67±0.06	3.22±0.61	2.05±0.03	**1.05±0.03**
Child	7.32±0.58	7.05±0.06	6.87±0.76	7.23±0.32	**5.05±0.92**	5.15±0.11
Neuropathic	7.84±0.15	5.08±0.22	1.21±0.00	9.34±0.29	7.08±0.32	**1.08±0.32**

We found that our algorithm performs optimally in most cases, except for the Child dataset. We speculate that this is due to the inherent difficulty of the task, which may involve more complex patterns or high noise in the data. Nevertheless, even in this case, the foundational knowledge of the LLM allows our algorithm to perform with only a minimal gap compared to the optimal algorithm. When considering stability, our algorithm may even be superior in some cases, particularly across different datasets, offering more consistent and reliable results. This suggests that, even in the face of challenges, our algorithm demonstrates strong adaptability and is able to maintain efficient performance in various environments.

6.2 Ablation study

We first observe the number of cycles at different node counts using two prompt types: pair-wise and triplet. By comparing these two methods, we can assess their performance in generating cycle structures. Specifically, we examine how the number of cycles changes as the number of nodes increases, and analyze the impact of each method on the cycle count under different node conditions. We tested on the Neuropathic dataset.

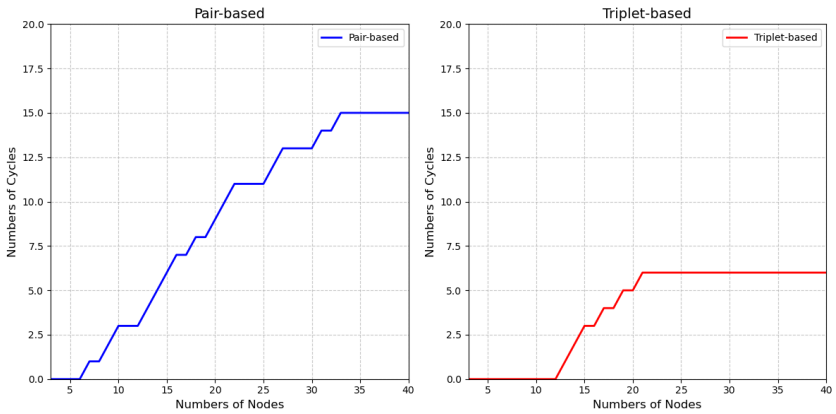

Figure 3: pair-wise and triplet prompts

Based on this figure, we can observe two points. The first is that triplet-based prompts can significantly reduce the number of cycles in the image, indicating that using triplet prompts is more effective than pair-based prompts in minimizing cycles. This may be because triplets provide more constraints, making the system's reasoning more stable. On the other hand, while LLMs can guide the process to some extent and help reduce the number of cycles, the presence of cycles remains unavoidable. This suggests that the reasoning capability of LLMs is still limited, and they cannot completely eliminate all possible cyclic structures. Alternatively, the inherent complexity of the task and the nature of the data might make cycle formation difficult to completely avoid.

Next, we examine the LLM-triplet and CAM algorithms, as well as the combined algorithm. We compare their SHD (Structural Hamming Distance) values, as LLM-triplet may still result in cycles.

Dataset	**LLM**	**CAM**	**CAM+LLM**
Earthquake	5	4	2
Cancer	5	4	4
Survey	5	2	2
Asia	8	2	1
Asia-M	3	6	3
Child	6	8	3
Neuropathic	8	5	3

Table 2: Ablation study

7 Conclusion

This paper explores novel approaches to leveraging Large Language Models (LLMs) for causal inference and discovery. We propose a triplet-based prompting strategy for extracting causal order from LLMs, demonstrating its effectiveness in causal graph discovery and causal effect estimation. Compared to directly inferring graph structures, this approach is more reliable as it is unaffected by other variables. Additionally, we introduce two algorithms that integrate LLM-inferred causal order with existing graph discovery techniques, enhancing their performance in causal inference tasks.

Our experimental results show significant improvements in causal order inference compared to traditional methods, highlighting the potential of LLMs as virtual domain experts in causal inference. We acknowledge the limitations of LLMs, such as the inability to completely eliminate cycles. Therefore, we combine LLMs with traditional graph discovery algorithms to further improve the accuracy and reliability of causal inference.

In summary, our research opens new avenues for utilizing LLMs in causal inference and discovery. Our approach not only enhances the accuracy of causal inference but also provides new insights for

developing more effective causal discovery algorithms. We believe that as LLMs continue to evolve, they will play an increasingly important role in the field of causal inference.

References

Ban, T., Chen, L., Wang, X., and Chen, H. (2023). From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10.

Gupta, S., Childers, D., and Lipton, Z. C. (2022). Local causal discovery for estimating causal effects. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*.

Huang, Y., Kleindessner, M., Munishkin, A., Varshney, D., Guo, P., and Wang, J. (2021). Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. *Frontiers in Big Data*, 4.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. (2020). Gradient-based neural dag learning. In *International Conference on Learning Representations (ICLR)*.

Long, S., Piché, A., Zantedeschi, V., Schuster, T., and Drouin, A. (2023). Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*.

Long, S., Schuster, T., and Piché, A. (2022). Can large language models build causal graphs? In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

Maathuis, M. H., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016). Distinguishing cause from effect using observational data: Methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Reisach, A., Seiler, C., and Weichwald, S. (2021). Beware of the simulated dag! causal discovery benchmarks may be easy to game. In *NeurIPS*, page 27772–27784.

Rolland, P., Cevher, V., Kleindessner, M., Russell, C., Janzing, D., Schölkopf, B., and Locatello, F. (2022). Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning (ICML)*.

Scutari, M. and Denis, J. (2014). *Bayesian Networks: With Examples in R*. Chapman & Hall/CRC Texts in Statistical Science.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. (2006). A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7(10).

Teyssier, M. and Koller, D. (2005). Ordering-based search: A simple and effective algorithm for learning bayesian networks. In *UAI*, volume 96, page 584–590.

Tu, R., Zhang, K., Bertilson, B., Kjellstrom, H., and Zhang, C. (2019). Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. In *Advances in Neural Information Processing Systems*, volume 32.

Willig, M., Zecević, M., Dhami, D. S., and Kersting, K. (2022). Probing for correlations of causal facts: Large language models and causality.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *NeurIPS*, 31.