Retrieving Classes of Causal Orders with Inconsistent Knowledge Bases

A Preprint

Federico Baldo
Sorbonne Université, INSERM,
Institut Pierre Louis d'Epidémiologie
et de Santé Publique,
F75012, Paris, France

Simon Ferreira
Sorbonne Université, INSERM,
Institut Pierre Louis d'Epidémiologie
et de Santé Publique,
F75012, Paris, France

Charles K. Assaad
Sorbonne Université, INSERM,
Institut Pierre Louis d'Epidémiologie
et de Santé Publique,
F75012, Paris, France

Abstract

Traditional causal discovery methods often rely on strong, untestable assumptions, which makes them unreliable in real applications. In this context, Large Language Models (LLMs) have emerged as a promising alternative for extracting causal knowledge from text-based metadata, which consolidates domain expertise. However, LLMs tend to be unreliable and prone to hallucinations, necessitating strategies that account for their limitations. One effective strategy is to use a consistency measure to assess reliability. Additionally, most text metadata does not clearly distinguish direct causal relationships from indirect ones, further complicating the discovery of a causal DAG. As a result, focusing on causal orders, rather than causal DAGs, emerges as a more practical and robust approach. We present a new method to derive a class of acyclic tournaments, which represent plausible causal orders, maximizing a consistency score derived from an LLM. Our approach starts by calculating pairwise consistency scores between variables, resulting in a semi-complete partially directed graph that consolidates these scores into an abstraction of the maximally consistent causal orders. Using this structure, we identify optimal acyclic tournaments, focusing on those that maximize consistency across all configurations. We subsequently show how both the abstraction and the class of causal orders can be used to estimate causal effects. We tested our method on both well-established benchmarks, as well as, real-world datasets from epidemiology and public health. Our results demonstrate the effectiveness of our approach in recovering the correct causal order.

Keywords Causal Inference, Large Language Models, Causal Orders

1 Introduction

Traditional causal discovery algorithms rely on observational data to uncover causal relationships. To do so, they often rely on strong assumptions Spirtes et al. (2001); Glymour et al. (2019); Peters et al. (2017); Assaad et al. (2022), such as causal sufficiency and faithfulness. The recent rise in popularity of Large Language Models (LLMs) offers a new tool to discover causal models Long et al. (2023a,b); Cohrs et al. (2023); Vashishtha et al. (2025); Kiciman et al. (2024). Indeed, unlike traditional causal discovery methods, LLM-aided approaches operate on textual data —leveraging pre-collected knowledge encoded in their training data.

Despite growing interest, initial attempts to extract reliable causal information from LLMs have met limited success. Indeed, LLMs have often been associated with ambiguous and inconsistent replies when queried on causal relationships Zečević et al. (2023). Moreover, we can argue that besides limited transfer of knowledge among different domains, LLMs have limited capabilities when it comes to uncovering new knowledge. To this end, most of their potential is grounded in the training data, which can include scientific literature and commonsense knowledge.

Most importantly, as underlined in Vashishtha et al. (2025), in natural language, direct and indirect causes are often conflated, making them difficult to distinguish. This ambiguity is evident in various domains such as philosophy, biology, and epidemiology. For instance, we commonly assert that a sedentary lifestyle causes type 2 diabetes, when in fact this link is fully mediated by obesity Li et al. (2022). More in general, in natural language, causal relations are frequently expressed as a simple relationship: ”X causes Y” or ”X affects Y” or ”X prevents Y”, etc. This oversimplification obscures the complex web of direct and indirect influences, including immediate ”parents” and distant ”ancestors” of a causal pathway. Given that LLMs are often characterized as ”causal parrots” Zečević et al. (2023) —meaning that they are just mimicking causal reasoning —we argue that they are more effective for identifying causal orders rather than constructing correct causal Directed Acyclic Grahps (DAGs) Vashishtha et al. (2025).

However, LLMs are notoriously unreliable, often producing hallucinated or inconsistent outputs. In this paper, we propose to quantify the reliability of the LLM and use such metric as a heuristic to identify classes of causal orders. The reliability of the LLM is represented by the self-consistency of the model when queried multiple times on pairwise causal relationships. The consistencies are then used to identify an abstraction, namely a Semi-Complete Partially Directed Graph (SCPDG), representing a compression of all maximally consistent causal orders. We show that the SCPDG can be transformed into a Maximally oriented Partially Directed Acyclic Graph (MPDAG) Perkovic (2020) to identify causal effects. Additionally, based on the SCPDG, we propose an exact method to derive all maximally consistent causal orders. We show that the class of orders can be used to estimate causal effects.

Unlike traditional causal discovery methods Spirtes et al. (2001); Glymour et al. (2019); Peters et al. (2017); Assaad et al. (2022), our approach does not rely on faithfulness, or any other parametric assumptions, but rather only on acyclicity and casual sufficiency. Additionally, we view the LLM as reasonably accurate knowledge base Zheng et al. (2024).

Contributions

- We provide an effective algorithm to find a class of causal orders maximally consistent with the knowledge provided by LLMs. Such method is based on a top-down search strategy that does not require any parametric assumptions or faithfulness.
- We demonstrate how to derive an MPDAG from abstractions of causal orders and how to identify causal effects using both the MPDAG and the class of causal orders.
- We offer a collection of realistic causal graphs from scientific literature that have not been utilized as benchmarks before.

The remainder of the paper is organized as follows: in Related Works, we review relevant literature related to LLM-aided causal discovery; in Background, we provide some of the basic notions used in the paper; the following section presents the main contribution of the paper and details regarding the proposed method; the last two sections provide a detailed description of experiments and discuss the results. All the proofs of Propositions and Theorems are provided in Appendix.

2 Related Work

Causal Inference with Background Knowledge. The use of expert knowledge in causal discovery has been a long-standing research topic, aimed at integrating domain-specific information to refine causal graphs. A notable first attempt at integrating background knowledge into causal representation has been made by Meek (1995), where a set of rules —known as Meek rules —were proposed to obtain maximally oriented Completed Partially Directed Acyclic Graph (CPDAG), representing a MEC, based on expert knowledge. These rules allow for the refinement of the CPDAG by orienting edges while preserving the conditional independence encoded in the graph and the acyclicity constraint. More recently, Maathuis and Colombo (2015); Perkovic et al. (2017); Perković et al. (2017); Perkovic (2020); Venkateswaran and Perković (2024) propose generalizations of identifiability by adjustment to abstraction of causal graphs, such as CPDAGs and Partially Directed Acyclic Graphs (PDAGs) and maximal PDAGs.

LLMs in Causal Discovery In causal discovery, LLMs are often viewed as expert, as they are trained on vast amounts of text data, including scientific literature and commonsense knowledge. In this context, numerous works attempted to integrate them to uncover causal DAGs Kiciman et al. (2024); Long et al. (2023a); Cohrs et al. (2023);

Vashishtha et al. (2025); Jiralerspong et al. (2024). However, LLMs have been shown to be unreliable and untrustworthy; their tendency to produce hallucinations is a notable example of this issue. Indeed, they often referred to as *imperfect experts* Long et al. (2023a); Vashishtha et al. (2025). Moreover, concerns regarding their capability to effectively reason have been raised, as they might be just capturing verbal pattern without actually learning the underlying reasoning Zečević et al. (2023). To tackle this problem most of the approaches proposed quantify the reliability of these models. As presented in Cohrs et al. (2025), there are two primary approaches to this task: the first is to compute the *uncertainty* of the LLM output using the probabilities associated to the tokens in the LLM's response; the second is to evaluate the *consistency*, i.e., the self-coherence, of the LLM output when queried multiple times. All of the LLM-aided causal disovery method start assuming to have a textual description of the variables involved, provided by a domain expert. In Cohrs et al. (2023), authors propose an LLM informed variant of the PC algorithm. Specifically, the PC algorithm is enhanced by incorporating LLMs to detect conditional independencies among variables. However, this method retains all the assumptions of the PC algorithm. In Long et al. (2023a), a pairwise prompt strategy is proposed to complete the orientation of edges in a given CPDAG. Each edge is associated to an uncertainty —based on the probabilities asigned to the response tokens —, then the Markov Equivalence Class (MEC) is refined through a Bayesian optimization process. This method relies on the assumption that a CPDAG is available, which is typically obtained through a traditional causal discovery algorithm. The pairwise prompt strategies, require a quadratic number of queries with respect to the number of variables. A more efficient approach to reduce the number of queries from quadratic to linear has been proposed in Jiralerspong et al. (2024). Starting from a set of variables deemed as prime causes, and provided explicitly by the LLM, the method learns the causal DAG through a BFS search. The method expands the causal DAG by identifying new variables that are influenced by the visited nodes.

LLMs and Causal Orders In Vashishtha et al. (2025), the authors propose a method to estimate causal orders using LLMs. The intuition is that LLMs can be more effective in identifying causal orders rather than full causal DAGs, given the inherent ambiguity of causal relationships in natural language. The method proposed estimates the topological order and then applies of the backdoor criterion Pearl (2009). The LLM is asked to provide a DAG for every triplet of variables, which allow to define a majority for the orientation of the edges. This method provide DAGs that are compatible with the causal order, but do not guarantee to estimate true causal graph.

In this work, as in Vashishtha et al. (2025), we start from the assumption that identifying causal orders rather than full causal DAGs presents a more direct task for LLMs. However, unlike Vashishtha et al. (2025), our approach focuses on: 1) using a pairwise prompt strategy to compute the self-consistency of the LLM, 2) identify an abstraction of the maximally consistent causal orders, namely a fully connected semi-complete partially directed graph, which can be used to obtain estimate of the causal effect, 3) further refines the abstraction to obtain a class of causal orders maximally consistent with the knowledge provided by the LLM —this class of causal orders can then be used to estimate the causal effects.

3 Background

Causal Graphs and Causal Orders A causal graph $\mathcal{G} = (\mathbb{V}, \mathbb{E})$ consists of a set of nodes $\mathbb{V}$ (or variables) and a set of directed edges $\mathbb{E}$. The existence of directed edge between two nodes indicates that there is a direct causal effect from X_i to X_j. Following standard causal assumptions Pearl (2009); Spirtes et al. (2001), we assume that $\mathcal{G}$ is a DAG, referred to as a *causal DAG*; additionally, we assume that all confounders are observed, i.e., causal sufficiency holds. Whenever we want to estimate the the total effect of a variable X on another variable Y, we denote it as $P(Y|do(X = x))$. In this context, the presence of confounding variables can lead to biased estimates of the causal effect. To address this issue, we can use the backdoor criterion, which provides a method for identifying sets of variables that, when controlled for, allow us to estimate the causal effect of X on Y without biasing the estimate Pearl (2009). The backdoor criterion is always satisfied if the conditioning set $\mathbb{Z}$ contains all the parents of X.

In this paper, we assume that the causal DAG is unknown and we rather focus on causal orders.

Definition 1 (Causal Order)**.** *Suppose a causal DAG $\mathcal{G}$. A causal order compatible with $\mathcal{G}$ is a bijective mapping $\pi : \mathbb{V} \mapsto \{1, \dots, d\}$ such that if Y is a descendant of X then $X \succ Y$, $\forall X, Y \in \mathbb{V}$.*

Most importantly, the backdoor criterion is always satisfied for causal orders by conditioning on all the predecessors of the treatment if there are no hidden confounders Pearl (2009); Vashishtha et al. (2025).

Maximally Partially Directed Graphs (MPDAGs) It is important to note that, in general, when causal sufficiency holds and only observational data are available, the best one can recover is a Complete Partially Directed Acyclic Graph (CPDAG) Spirtes et al. (2001). A CPDAG is a graph that represents the equivalence class of all DAGs that are

Markov equivalent to each other, meaning they encode the same conditional independence relations. In the presence of background knowledge — for instance provided by an expert — we can refine a CPDAG into an MPDAG Perkovic et al. (2017) by incorporating the additional edge orientations and then applying the Meek rules to propagate their implications. However, the backdoor criterion cannot be directly applied to these abstraction, which include undirected edges; to this end, the generalized backdoor criterion Maathuis and Colombo (2015); Perkovic et al. (2017), allows us to identify causal effects in CPDAGs.

Definition 2 (Generalized Backdoor Criterion)**.** *Let X, Y be sets of variables in a CPDAG $\mathcal{C}$. Then a set $\mathbb{Z}$ satisfies the generalized backdoor criterion relative to (X, Y) if:*

1. *$\mathbb{Z}$ does not contain possible descendants of X in $\mathcal{C}$;*
2. *Z blocks all directed paths from X to Y in $\mathcal{C}$.*

The generalized backdoor criterion extends to MPDAGs since they are enriched CPDAGs, in which all conditional independencies are preserved, and the orientation of edges is refined based on additional knowledge. If we can identify a backdoor set then the total effect is identifiable and can be estimated using the adjustment formula Pearl (2009):

$$P(Y|do(X = x)) = \sum_z P(Y|X = x, Z = z)P(Z = z) \tag{1}$$

4 Inconsistent Knowledge Base

The aim of this paper is to identify factual knowledge permeated in the LLMs from established literature regarding causal relationships in a specific domain. This knowledge is then used to identify causal orders maximally compatible with the information provided by the LLM. In this context, the LLM figures as a knowledge base Zheng et al. (2024) who: 1) has access to a large body of knowledge and 2) may provide incorrect responses, e.g., hallucinated. Measuring uncertainty in LLMs is a well-established practice, which aims at quantifying the reliability of the information provided by the model. In this paper, we use self-consistency as a proxy of uncertainty. Consistency is an effective measure for assessing the reliability of the information provided by LLMs; indeed, it has been shown to outperform other uncertainty metrics, such as entropy, confidence elicitation and token-level probabilities Manakul et al. (2023); Savage et al. (2024). Additionally, consistency has proven effective in reducing hallucinations Ji et al. (2023).

Following the approach adopted in Long et al. (2023a); Kadavath et al. (2022), we assume to have a set of variables $X_1, ..., X_d$ with a set of descriptive metadata associated to each variable (i.e. a textual description of the variable), $\mu_1, ..., \mu_d$. The consistency is the degree of agreement of the LLM when queried multiple times about the causal relationship between two variables X_i and X_j with semantically equivalent queries. The queries are generated by the LLM itself, which is asked to rephrase a starting prompt, such as *"Is X_i a cause of X_j?"*, into a set of semantically equivalent queries —more details on the prompts used in Appendix E. To reduce the number of incorrect responses, we reduce the possible answers to a `Yes` or a `No`. The consistency score is then computed as the proportion of `Yes` responses to the queries. Specifically, when an LLM is queried n times, the consistency score for $X_i \rightarrow X_j$ is calculated as:

$$C_{X_i \rightarrow X_j} = \frac{1}{n} \sum_{k=1}^{n} r_k$$

,where r_k is the response of the LLM to the k-th query, and $r_k = 1$ if the response is `True`, $r_k = 0$ if `False`, and $r_k = 0.5$ in case of non-admissible answers. It is important to note that $C_{X_i \rightarrow X_j}$ and $C_{X_j \rightarrow X_i}$ are computed independently of each other. We can define the following notions of consistency for an knowledge base:

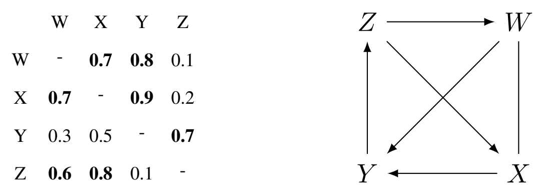
Definition 3. *A knowledge base is said to be consistent if, for every pair of variables X_i and X_j, such that $X_i \succ X_j$ the consistency score $C_{X_i \rightarrow X_j} \geq C_{X_j \rightarrow X_i}$.*

Definition 4. *An knowledge base is said to be strictly consistent if, for every pair of variables X_i and X_j, such that $X_i \succ X_j$ the consistency score $C_{X_i \rightarrow X_j} > C_{X_j \rightarrow X_i}$.*

It follows that any strictly consistent knowledge base is also consistent.

5 Maximally Consistent Causal Orders

Maximally Consistent Semi-Complete Partially Directed Graphs. The consistency matrix can be leveraged to construct a maximally consistent graph, $\mathcal{S}$. Specifically, a directed edge $X_i \rightarrow X_j$ is included in $\mathcal{S}$ if $C_{i \rightarrow j} \geq C_{j \rightarrow i}$. Note that this implies that when $C_{i \rightarrow j} = C_{j \rightarrow i}$ an undirected edge will be formed between X_i and X_j. Moreover, $\mathcal{S}$



	W	X	Y	Z
W	-	**0.7**	**0.8**	0.1
X	**0.7**	-	**0.9**	0.2
Y	0.3	0.5	-	**0.7**
Z	**0.6**	**0.8**	0.1	-

(a) (b)

Figure 1: (a) A consistency matrix for a semi-complete partially directed graph. The values in bold represent edges maximizing consistency. (b) A semi-complete partially directed graph.

can contain cycles, since there is no explicit mechanisms preventing them. A graph defined in this way is referred to as a *semi-complete partially directed graph*.

Definition 5 (Semi-Complete Partially Directed Graphs (SCPDG))**.** *A semi-complete partially directed graph is a dense graph where there is at least one arc, either directed or undirected, between each pair of its vertices.*

This graph represents an abstraction of the causal orders, capturing the directed relationships between nodes without committing to a fully resolved causal order or acyclic structure. By searching for such an abstraction as a starting point, we can harness its structure to identify or constrain the set of compatible causal orders. Moreover, we can show that the based on the knowledge base consistency the graph $\mathcal{S}$ is an MPDAG.

Proposition 1. *The maximally consistent SCPDG, $\mathcal{S}$, does not contain directed cycles if the consistency matrix is provided by a consistent knowledge base.*

The maximally consistent SCPDG, $\mathcal{S}$, obtained from a consistent knowledge base and by applying all the Meek rules is a MPDAG.

We should also note that the MPDAG obtained in such way is dense, and allows to determine identifiability efficiently.

Order Invariant Nodes. The graph $\mathcal{S}$ allows both for cycles and undirected edges. This implies the existence of the limit cases in which a node is connected to the rest of the graph only through undirected edges. These nodes are *order invariant* with respect to the causal order, since they can be placed in any position. Interestingly, this holds true for any node defined as follows:

Definition 6. *Let $\mathcal{S}$ be a maximally consistent SCPDG. We say that a vertex X in $\mathcal{S}$ is order invariant if X has an undirected edge between every other vertex in $\mathcal{S}$.*

These nodes do not provide additional information regarding the causal order, and can prevent the identification of the causal effects.

Maximally Consistent Acyclic Tournaments. The SCPDG, $\mathcal{S}$, captures the directed relationships between nodes, but it does not necessarily represent a valid causal order. To do so, we need to transform $\mathcal{S}$ into an acyclic tournament.

Definition 7 (Acyclic tournament)**.** *An acyclic tournament is a DAG with exactly one edge between each two vertices, in one of the two possible directions.*

An acyclic tournament provides a graphical representation that fully encodes a unique causal order. Specifically, the direction of the edge between any two nodes directly reflects their relative position in the causal order. To establish a connection between SCPDG and acyclic tournaments, we introduce the concept of compatibility, formalized in the following definition:

Definition 8 (Compatible Acyclic Tournament)**.** *Given a SCPDG, $\mathcal{S} = (\mathbb{V}, \mathbb{E})$, an acyclic tournament $\mathcal{T}$ is said to be compatible with $\mathcal{S}$ if it can be derived from $\mathcal{S}$ by reversing certain edges within each SCC of $\mathcal{S}$, while leaving all other edges unchanged.*

To obtain a class of plausible causal order, we need to find all maximally consistent acyclic tournaments compatible with the SCPDG.

Definition 9 (Maximally consistent acyclic tournament (MCAT) compatible with $\mathcal{S}$)**.** *An acyclic tournament compatible with $\mathcal{S}$ is said to be maximally consistent if it maximize the consistency score relative to all other acyclic tournaments compatible with $\mathcal{S}$.*

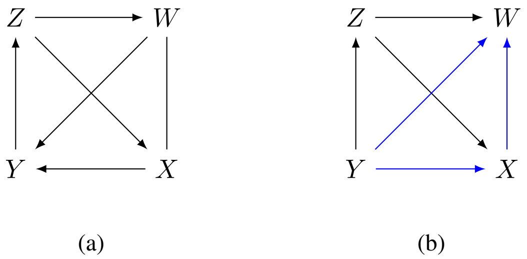


Figure 2: (a) an illustration of a semi-complete directed graph deprived of invariant nodes and (b) a compatible acyclic tournament with maximal consistency score.

Note that there might be multiples MCATs compatible with $\mathcal{S}$. To find them, we need to identify all acylic transformation of $\mathcal{S}$ leading to a compatible MCAT. We identify two main scenarios:: 1) $\mathcal{S}$ is a MCAT and maximally consistent by construction. 2) we need to find the minimal set of edges to reverse in $\mathcal{S}$ to eliminate all cycles while maximizing the consistency score.

The minimal set of edges to remove from a graph to transform it into an acyclic one is known as the Feedback Arc Set (FAS). A FAS is a smallest set of directed edges in a graph that when removed eliminates all cycles from it. However, since we are searching for a tournament rather than just any DAG, we require the resulting graph to remain fully connected. As shown in Barthélemy et al. (1995), reversing all edges in the FAS also results in an acyclic graph, and in our case, an acyclic tournament. Moreover, in our context, we are not merely interested in any acyclic transformation of $\mathcal{S}$, but rather the one that maximizes the consistency score. Therefore, we focus on a weighted version of FAS, where the goal is to find a FAS that maximizes the sum of the weights associated to the edges. To this end, instead of considering consistencies scores as weights, we define a cost score, $B_{i\to j}$, for each edge, $X_i \to X_j$, in $\mathcal{S}$.

$$B_{i\to j} = \xi - C_{i\to j} + C_{j\to i}.$$

where ξ is the total consistency score of $\mathcal{S}$, calculated as the sum of the consistency scores of all edges. In this context, the cost score $B_{i\to j}$ represents the cost of reversing the edge $X_i \to X_j$ in $\mathcal{S}$. The goal is to find a FAS that maximizes the total cost score, which is equivalent to maximizing the total consistency score of the resulting acyclic tournament. To find a FAS from $\mathcal{S}$ and the cost B, we use an exact solution, denoted as ExactFAS based on an integer programming formulation which is guaranteed to yield an optimal result. However, since, finding a FAS is NP-complete, ExactFAS can be too slow for very for large graphs. Nevertheless, assuming that the true causal graph is acyclic, we expect the graphs to be relatively small, allowing for efficient detection of a FAS.

Finding All Maximally Consistent Acyclic Tournaments. Obtaining the FAS maximizing the total cost score does not guarantee to find all the MCATs compatible with $\mathcal{S}$. Indeed, there might be multiple MCATs compatible with $\mathcal{S}$. To this end, we proposed a method to find all the maximal weighted acyclic tournaments given a matrix of weights W, denoted as MATS (**M**aximal Weighted **A**cyclic **T**ournaments **S**earch). The intuition is to constrain the solution space based on the previous solution of ExactFAS, which is guaranteed to yield a maximal acyclic tournament. More specifically, let A^* be a solution obtained by the ExactFAS, meaning that a there is a maximally weighted acyclic tournament, $\mathcal{T} = (V, E_{\mathcal{T}})$, such that $E_{\mathcal{T}} = (E_{\mathcal{S}}\ \mathbb{A}^*) \cup \mathbb{A}^{*T}$. We want to find all the acyclic tournaments $\mathcal{T}' = (V, E'_{\mathcal{T}})$ such that $E'_{\mathcal{T}} = (E_{\mathcal{S}}\ \mathbb{A}') \cup \mathbb{A}^{T'}$ and $\mathbb{A}' \neq \mathbb{A}^*$. This can be obtained excluding subsets of edges $\mathbb{F}$ from the admissible edges, such that $\mathbb{F} \in \mathcal{P}(\mathbb{A}^*)$. If the exclusion of $\mathbb{F}$ leads to a suboptimal solution, i.e., of non-maximal score, then we exclude from the search all the subsets containing $\mathbb{F}$, such that $\mathbb{F} \subset \mathbb{F}'$, since they all lead to suboptimal solutions The process is repeated for every optimal solution of the ExactFAS.

Theorem 1. *The MATS algorithm is sound, complete and terminates.*

Proposition 2. *If the consistency matrix is provided by a consistent knowledge base., the MATS algorithm is guaranteed to return a class of acyclic tournaments containing the true causal order.*

Proposition 3. *If the consistency matrix is provided by a strictly consistent knowledge base., the MATS algorithm is guaranteed to return the true causal order.*

6 Reasoning with Classes of Causal Orders

Reasoning with Semi-Complete Partially Directed Graphs. To reason with SCPDG, we rely on the *generalized backdoor criterion* Maathuis and Colombo (2015), which applies to MPDAGs Perković et al. (2017). Following the

Corollary B, if the maximally consistent SCPDG, $\mathcal{S}$, is derived from a consistent knowledge base it will always be an MPDAG. However, this does not hold in presence of inconsistent knowledge bases, where $\mathcal{S}$ may have a directed cycle. Without loss of generality, we can transform the SCPDG into a MPDAG by: 1) substituting the directed edges contained in a cycle with an undirected ones and 2) enforce the Meek rules. It is important to note that since the graph is dense, only the rule enforcing acyclicity will be applied, since there are not colliders in the graph. Unfortunately, this transformation may lead to a loss of information.If we assume to have only one treatment variable X and an outcome variable Y, we can still reason with the MPDAG Perkovic et al. (2017). In this context, since the MPDAG is fully connected, the set of parents of the treatment variable X is a valid adjustment set, and it is equivalent to the backdoor set. Moreover, in this paper we provide a simpler characterization of identifiability for dense MPDAGs; indeed, every time an undirected edge is related to the treatment variable X, it indicates that the total effect is not identifiable by adjustment.

Proposition 4. *Given a dense MPDAG and the treatment variable X, the total effect is not identifiable, using (1), iff there are no undirected edge related to X.*

Given a dense MPDAG, if it contains an invariant node, any causal effect between two variable X and Y is not identifiable.

Reasoning with Acyclic Tournaments. Identification over the dense MPDAG derived from the SCPDG might not be possible in some cases, e.g., in the SCPDG the treatment is part of a cycle. To address this, we can reason using the the maximally consistent causal orders compatible with $\mathcal{S}$. Given an estimated causal order, if we assume that the treatment X is sorted before the outcome Y, $X \succ Y$, then we can estimate the causal effect of X on Y adjusting for any predecessor of the treatment X. Additionally, reasoning over classes of causal orders allows us to provide uncertainty over the estimation. Indeed, we can identify orders with the same causal effect, i.e., if $\mathbb{Z} \succ X \succ \mathbb{M} \succ Y$ and $\mathbb{Z} \succ X \succ \mathbb{M}' \succ Y$ are both in the class of MCATs compatible with $\mathcal{S}$, then their effect is the same. For every pair of orders in the class for which the backdoor set is the same, the causal effect of X on Y is the same. Additionally, this allows us to weight the estimation based on the number of orders that contain the same causal effect, which in turns identifies a probability distribution over the causal effect of X on Y.

7 Experimental Results

The code[1] is implemented in `python 3.10`. We relied on multiple LLMs as inconsistent knowledge bases. More specifically, results presented in Table 1 and 2 were obtained using `gpt-4.1-nano` a fast and lightweight version of `gpt`. However, the method is designed to be compatible with many open-source LLMs, particularly those available through the `ollama` platform—a library that facilitates the management of multiple LLMs. Among these, we tested our method on: `mistral` with 7 billion parameters.All results pertaining to the use of `mistral` are available in Appendix. The implementation of graphs relies on the `igraph`, a `C++` library that offers an implementation of ExactFAS —more details in Appendix.

Baselines. Concerning LLM-aided approaches, we compare to a state-of-the-art method for discovering causal orders, proposed in (Vashishtha et al., 2025), which we refer to as Triplets[2]. Additionally, to provide a comprehensive overview of the method's potential, we compare to more traditional causal discovery approaches. Specifically, we conducted experiments using the PC algorithm with Fisher's conditional independence test[3], and a linear version of NOTEARS[4]. Finally, we present results from the hybridization of MATS with the PC algorithm, where the estimated orders are used to orient the edges in the graph's skeleton.

Datasets We tested the method on 11 causal graphs, which included a minimum of 3 nodes and a maximum of 8 nodes. Further details can be found in the Appendix. Among these, 2 are well-known causal DAGs included in the `bnlearn` library, namely Asia Lauritzen and Spiegelhalter (2018) and Cancer Korb and Nicholson (2004).Additionally, we utilized 9 real-world causal graphs, primarily sourced from scientific literature in the fields of epidemiology and public health. These causal DAGs include Covid 1, Covid 2, Covid 3 Griffith et al. (2020), Covid 4 Glemain et al. (2024), Genetic Palmer et al. (2012), MSU Piccininni et al. (2023), Neighborhood Chaix et al. (2009), Climate Guevara et al. (2024), Supermarket Chaix et al. (2012). Using real-world benchmarks is essential to validate the effectiveness of our method in practical scenarios. The benchmarks included in the `bnlearn` library are widely recognized and commonly used for testing, which increases the likelihood that the LLM has effectively learned the causal graph.

[1] An anonymized version of our code is available at https://github.com/LCOSKDD/LCOS.git

[2] We provide an implementation of the method in at https://github.com/LCOSKDD/LCOS.git

[3] https://github.com/py-why/causal-learn

[4] https://github.com/xunzheng/notears

	Data-Driven				**Text-Driven**		**Text-Driven + Data-Driven**
Dataset	**PC**		**NOTEARS**		**Triplets**	**MATS (Ours)**	MATS+PC (Ours)
	$\mathcal{D}_{top}$	SHD	$\mathcal{D}_{top}$	SHD	$\mathcal{D}_{top}$	$\mathcal{D}_{top}$	SHD
Asia	1.5 ± 0.8	3.0 ± 1.7	2.0	2.0	7.0	0.0 ± 0.0	0.0 ± 0.0
Cancer	0.0 ± 0.0	0.0 ± 0.0	2.0	3.0	0.0	0.0 ± 0.0	0.0 ± 0.0
Climate	0.0 ± 0.0	1.0 ± 0.0	1.0	1.0	6.0	0.0 ± 0.0	1.0 ± 0.0
Covid 1	0.0 ± 0.0	0.0 ± 0.0	1.0	1.0	1.0	0.0 ± 0.0	0.0 ± 0.0
Covid 2	0.5 ± 0.5	1.0 ± 1.0	0.0	0.0	1.0	0.0 ± 0.0	0.0 ± 0.0
Covid 3	0.5 ± 0.5	1.0 ± 1.0	0.0	1.0	0.0	1.5 ± 0.5	3.0 ± 1.0
Covid 4	1.5 ± 0.5	3.0 ± 1.4	1.0	3.0	6.0	0.0 ± 0.0	0.0 ± 0.0
Genetic	0.0 ± 0.0	0.0 ± 0.0	3.0	4.0	5.0	0.0 ± 0.0	0.0 ± 0.0
MSU	1.0 ± 0.0	2.0 ± 0.0	2.0	4.0	0.0	2.0 ± 0.0	4.0 ± 0.0
Neighbor	2.5 ± 0.8	4.0 ± 1.7	2.0	4.0	5.0	1.0 ± 0.0	3.0 ± 0.0
Supermarket	4.0 ± 0.0	8.0 ± 1.0	2.0	6.0	10.0	3.0 ± 0.0	7.0 ± 0.0

Table 1: $\mathcal{D}_{top}$ (↓) of the estimated causal orders. SHD (↓) of estimated causal graphs. The values in blue represent the best performing method for each dataset.

For data-driven methods, we generated synthetic data based on the true causal DAG, which was also used to estimate causal effects —more details in Appendix.

Evaluation Classical metrics for evaluating the performance of causal discovery methods include the Structural Hamming Distance (SHD), which counts the modification to transform the estimated in the graph into the true one. However, traditional metrics are not as effective for evaluating the error on estimated causal orders. Indeed, there can be multiple causal orders that are consistent with the same causal DAG. To this end, we rely on a metric proposed in Ruiz et al. (2022); Rolland et al. (2022), that measures the number of parents sorted after children in the estimated order —compared to the true DAG, $\mathcal{G}$. Thus, for each node $Y \in \mathcal{V}$, a correct causal order assumes that for all parents of Y, $X \in Parent(Y)$, $X \succ Y$. The metric is then defined as follows:

$$\mathcal{D}_{top} = \sum_{u \in V} \sum_{v \in Desc_{\mathcal{G}}(u)} \mathbf{1}(v \notin Desc_{\hat{\mathcal{G}}}(u))$$

where $\mathcal{G}$ is the true causal DAG and $\hat{\mathcal{G}}$ is the estimated DAG (for our method, the DAG is an acyclic tournament), and $\mathbb{E}$ is the set of edges in $\mathcal{G}$. To present results in Table 1 we relied on both SHD and $\mathcal{D}_{top}$. It is worth noting that MATS and PC output a class of graphs, rather than a single graph. In such cases, we compute the average and standard error across all graphs in the class.

We compute the total effect based on the MPDAG derived from a SCPDG and then use the estimated causal orders. The estimation is performed through a linear regression on the linear synthetic data and compared to the true total effect with the Absolute Error (AE).

Results As shown in Table 1, the MATS algorithm consistently outperforms all other methods, achieving the lowest error across the majority of the benchmarks. In particular, in 7 out of the 11 datasets MATS recovers a class of causal orders containing exclusively correct causal orders. Moreover, when MATS is not the best-performing method, namely in the Covid 3, MSU, and Supermarket causal graphs, the error is still consistently low. The hybrid model combining MATS and PC remains the best performing, on average, closely followed by PC. It is worth noting that orienting the edges of the skeleton based on the estimated orders can lead to a conflation into a single causal DAG. In general, the results obtained from data-driven methods are relative to data generated assuming linearity. In contrast, text-based methods do not require parametric assumptions to remain consistent even in non-linear settings. Most interestingly, the worst case of MATS is always better than the worst of every other method.

In Table 2, we present the AE relative to the estimation of the total effect. We evaluate the error separately based on the MPDAG obtained from the SCPDG, as well as the estimated causal orders. In general, we observe a low AE for most cases concerning the estimation based on causal orders. However, this is not the case for MPDAGs, where the effect is not identifiable in most cases. Notably, the highest AE observed is consistent with errors in the estimation of the causal orders —Covid 3, MSU, and Supermarket benchmarks.

Dataset	**MPDAG**$_{MATS}$	**Causal Orders**$_{MATS}$	$\mathcal{D}_{top}$
Asia	-	0.0006 ± 0.0	0.0 ± 0.0
Cancer	-	0.001 ± 0.0	0.0 ± 0.0
Climate	-	0.035 ± 0.005	0.0 ± 0.0
Covid 1	0.013	0.013 ± 0.0	0.0 ± 0.0
Covid 2	0.0027	0.0027 ± 0.0	0.0 ± 0.0
Covid 3	-	0.09 ± 0.06	1.0 ± 0.5
Covid 4	0.042	0.042 ± 0.0	0.0 ± 0.0
Genetic	0.027	0.027 ± 0.0	0.0 ± 0.0
MSU	-	0.15 ± 0.0	2.0 ± 0.0
Neighbor	0.0094	0.012 ± 0.002	1.0 ± 0.0
Supermarket	-	0.62 ± 0.0	3.0 ± 0.0

Table 2: AE (↓) of total effect estimates for the MPDAG and Causal Orders obtained by the MATS algorithm. If non identifiable the value is set to '-'. For reference, the last column shows the $\mathcal{D}_{top}$ of the estimated causal orders.

8 Discussion & Conclusion

The method described in this paper offers an effective approach to using LLMs as a knowledge base for retrieving classes of causal orders. This approach relies solely on textual descriptions of the variables and does not require observational data. A key insight behind the method is that natural language often leaves causal mechanisms implicit. This implies that some causal relationships may not be explicitly stated in the text, yet the causal order remains intact.

We compared our methods with data-driven methods, namely PC and NOTEARS, as well as, a LLM-aided method for causal order discovery Vashishtha et al. (2025). The results point to the fact that our approach can provide accurate estimation most of the time outperforming the other methods. Moreover, PC and NOTEARS have been applied to data generated assuming linearity. In contrast, text-based methods do not make parametric assumptions.

Memorization. A crucial aspect to take into account when using LLMs in causal inference tasks is *memorization*. Memorization refers to the ability of LLMs to recall specific information from their training data Carlini et al. (2023). Indeed, the scientific literature from which we derived the benchmark graphs may have been part of the data used to train the LLM. In this sense, we are actually looking for a trade-off: on the one hand, we want an LLM that has encoded knowledge regarding a specific causal link; on the other hand, we do not want the LLM to directly derive the causal graph from its training data, as this would be a poor generalization. Assessing if the LLM has memorized the causal graph is a hard task. A naive attempt to evaluate if the model has memorized the graph would be to ask directly. In most cases, when we asked, the language model failed to recognize the scientific papers from which the real-world graphs were derived. However, it was aware of the graphs included in the `bnlearn` library, but it could not specify them. This does not guarantee that the LLM has not memorized the causal graph, but it suggests that it is not directly aware of it.

Limitations. The MATS algorithm has several limitations that should be taken into account. First, the accuracy of the estimation strongly relies on the consistency of the LLM. In the presence of an inconsistent knowledge bases, we cannot guarantee the correctness of the class of orders retrieved. Second, the computational complexity of the method can increase significantly with larger graphs. This is primarily due to the calculation of the consistency matrix, which has a quadratic complexity in relation to the number of nodes, and the ExactFAS algorithm, which is an NP-hard problem.

Future Works. Extensions of this work will focus on reducing computational complexity. Efficiently computing the consistency matrix could be achieved by parallelizing the queries to the LLM for a specific causal link. Furthermore, we did not explore the use of chain-of-thought based LLMs Wei et al. (2023), which could potentially improve the accuracy of the method. In situations where a collection of documents relevant to a specific domain is available, we can enhance the reliability of the knowledge base by using Retrieval-Augmented Generation (RAG) Lewis et al. (2020). This approach enables the retrieval of pertinent information from a document corpus to effectively answer questions.

Acknowledgments

This work was supported by the CIPHOD project (ANR-23-CPJ1-0212-01).

References

Charles K. Assaad, Emilie Devijver, and Eric Gaussier. 2022. Survey and Evaluation of Causal Discovery Methods for Time Series. *J. Artif. Intell. Res.* 73 (2022), 767–819. https://doi.org/10.1613/jair.1.13428

Jean-Pierre Barthélemy, Olivier Hudry, Garth Isaak, Fred S. Roberts, and Barry Tesman. 1995. The reversing number of a diagraph. *Discrete Applied Mathematics* 60, 1 (1995), 39–76. https://doi.org/10.1016/0166-218X(94)00042-C

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying Memorization Across Neural Language Models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=TatRHT_1cK

Basile Chaix, Kathy Bean, Mark Daniel, Shannon Zenk, Yan Kestens, Hélène Charreire, Cinira Leal, Frédérique Thomas, Noëlla Karusisi, Christiane Weber, Jean-Michel Oppert, Chantal Simon, Juan Merlo, and Bruce Pannier. 2012. Associations of Supermarket Characteristics with Weight Status and Body Fat: A Multilevel Analysis of Individuals within Supermarkets (RECORD Study). *PloS one* 7 (04 2012), e32908. https://doi.org/10.1371/journal.pone.0032908

Basile Chaix, Cinira Leal, and David Evans. 2009. Neighborhood-level Confounding in Epidemiologic Studies Unavoidable Challenges, Uncertain Solutions. *Epidemiology (Cambridge, Mass.)* 21 (11 2009), 124–7. https://doi.org/10.1097/EDE.0b013e3181c04e70

Kai-Hendrik Cohrs, Emiliano Diaz, Vasileios Sitokonstantinou, Gherardo Varando, and Gustau Camps-Valls. 2023. Large Language Models for Constrained-Based Causal Discovery. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*. https://openreview.net/forum?id=NEAoZRWHPN

Kai-Hendrik Cohrs, Emiliano Diaz, Vasileios Sitokonstantinou, Gherardo Varando, and Gustau Camps-Valls. 2025. Large language models for causal hypothesis generation in science. *Machine Learning: Science and Technology* 6, 1 (Jan. 2025), 013001. https://doi.org/10.1088/2632-2153/ada47f Publisher: IOP Publishing.

Benjamin Glemain, Charles Assaad, Walid Ghosn, Paul Moulaire, Xavier de Lamballerie, Marie Zins, Gianluca Severi, Mathilde Touvier, Jean-François Deleuze, SAPRIS-SERO study group, Nathanaël Lapidus, and Fabrice Carrat. 2024. Does hospital overload increase the risk of death when infected by SARS-CoV-2? *medRxiv* (2024). https://doi.org/10.1101/2024.08.26.24312569

Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics* 10 (2019). https://doi.org/10.3389/fgene.2019.00524

Gareth J. Griffith, Tim T. Morris, Matthew J Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma C. Sharp, Jonathan A. C. Sterne, Tom M. Palmer, George Davey Smith, Kate Tilling, Luisa Zuccolo, Neil Martin Davies, and Gibran Hemani. 2020. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Communications* 11 (2020). https://api.semanticscholar.org/CorpusID:218937284

Laura Andrea Barrero Guevara, Sarah C Kramer, Tobias Kurth, and Matthieu Domenech de Cellès. 2024. Causal inference concepts can guide research into the effects of climate on infectious diseases. (2024). arXiv:2402.12507 [q-bio.PE] https://arxiv.org/abs/2402.12507

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1827–1843. https://doi.org/10.18653/v1/2023.findings-emnlp.123

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. 2024. Efficient Causal Graph Discovery Using Large Language Models. In *ICLR 2024 Workshop: How Far Are We From AGI*. https://openreview.net/forum?id=5RBUTx75yr

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. arXiv:2207.05221 [cs.CL] https://arxiv.org/abs/2207.05221

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=mqoxLkX210 Featured Certification.

Kevin B. Korb and Ann E. Nicholson. 2004. Bayesian Artificial Intelligence. (2004). https://api.semanticscholar.org/CorpusID:203667732

S. L. Lauritzen and D. J. Spiegelhalter. 2018. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society: Series B (Methodological)* 50, 2 (12 2018), 157–194. https://doi.org/10.1111/j.2517-6161.1988.tb01721.x arXiv:https://academic.oup.com/jrsssb/article-pdf/50/2/157/49097926/jrsssb_50_2_157.pdf

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Dan-dan Li, Yang Yang, Zi-yi Gao, Li-hua Zhao, Xue Yang, Feng Xu, Chao Yu, Xiu-lin Zhang, Xue-Qin Wang, Li-hua Wang, and Jian-Bin Su. 2022. Sedentary lifestyle and body composition in type 2 diabetes. *Diabetology & Metabolic Syndrome* 14 (01 2022). https://doi.org/10.1186/s13098-021-00778-6

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. 2023a. Causal Discovery with Language Models as Imperfect Experts. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*. https://openreview.net/forum?id=RXlvYZAE49

Stephanie Long, Tibor Schuster, and Alexandre Piché. 2023b. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279* (2023).

Marloes H. Maathuis and Diego Colombo. 2015. A generalized back-door criterion. *The Annals of Statistics* 43, 3 (2015), 1060 – 1088. https://doi.org/10.1214/14-AOS1295

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

Christopher Meek. 1995. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (Montréal, Qué, Canada) *(UAI'95)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 403–410.

Tom M. Palmer, Deborah A. Lawlor, Roger M. Harbord, Nuala A. Sheehan, Jonathan H. Tobias, Nicholas John Timpson, George Davey Smith, and Jonathan A. C. Sterne. 2012. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research* 21 (2012), 223 – 242. https://api.semanticscholar.org/CorpusID:14863122

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2 ed.). Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511803161

Emilija Perkovic. 2020. Identifying causal effects in maximally oriented partially directed acyclic graphs. 124 (03–06 Aug 2020), 530–539. https://proceedings.mlr.press/v124/perkovic20a.html

Emilija Perković, Markus Kalisch, and Maloes H Maathuis. 2017. Interpreting and using CPDAGs with background knowledge. *Association for Uncertainty in Artificial Intelligence (UAI)* (2017).

Emilija Perkovic, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. 2017. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *J. Mach. Learn. Res.* 18, 1 (Jan. 2017), 8132–8193.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Marco Piccininni, Tobias Kurth, Heinrich J Audebert, and Jessica L Rohmann. 2023. The effect of mobile stroke unit care on functional outcomes: an application of the front-door formula. *Epidemiology* 34, 5 (2023), 712–720.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russel, Bernhard Schölkopf, Dominik Janzing, and Francesco Locatello. 2022. Score matching enables causal discovery of nonlinear additive noise models. (2022). arXiv:2203.04413 [cs.LG] https://arxiv.org/abs/2203.04413

Gabriel Ruiz, Oscar Hernan Madrid Padilla, and Qing Zhou. 2022. Sequentially learning the topological ordering of causal directed acyclic graphs with likelihood ratio scores. (2022). arXiv:2202.01748 [stat.ME] https://arxiv.org/abs/2202.01748

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2024. Large language model uncertainty measurement and calibration for medical diagnosis and treatment. *medRxiv* (2024), 2024–06.

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

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. 2025. Causal Order: The Key to Leveraging Imperfect Experts in Causal Inference. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=9juyeCqL0u

Aparajithan Venkateswaran and Emilija Perković. 2024. Towards Complete Causal Explanation with Expert Knowledge. arXiv:2407.07338 [stat.ML] https://arxiv.org/abs/2407.07338

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=tv46tCzs83

Danna Zheng, Mirella Lapata, and Jeff Z. Pan. 2024. Large Language Models as Reliable Knowledge Bases? *CoRR* abs/2407.13578 (2024). https://doi.org/10.48550/arXiv.2407.13578

A Appendix

B A. Proofs

Proposition 5. *The maximally consistent semi-complete partially directed graph $\mathcal{S}$ does not contain directed cycles if the consistency matrix is provided by a consistent knowledge bases.*

Proof. If the knowledge bases is consistent, all orientation in $\mathcal{S}$ preserve the true causal order, thus it cannot introduce be directed cycles. □

The maximally consistent semi-complete partially directed graph $\mathcal{S}$ obtained from a consistent knowledge bases and by applying all the orientation rules is a MPDAG.

Proof. The proof follows from Proposition 5. If the knowledge bases is consistent, the maximally consistent semi-complete partially directed graph $\mathcal{S}$ does not contain directed cycles. Additionally, since the graph is fully connected, there cannot be unshielded colliders. Thus, all orientation rules can be applied and the resulting graph is a MPDAG. □

Theorem 2. *The MATS algorithm is sound, complete and terminates.*

Proof. We define the following notation in reference to Algorithm 1:

- `maxScore` is the maximal consistency score; thus, given a maximally consistent acyclic tournament $\mathcal{T} = (\mathbb{V}, \mathbb{E}_{\mathcal{T}})$:
$$\texttt{maxScore} = \sum_{(i,j)\in\mathbb{E}_{\mathcal{T}}} W[i,j]$$
- The method `ExactFAS` returns a FAS of a directed graph $\mathcal{G}$, which is a set of edges that can be reversed to make $\mathcal{G}$ acyclic and of maximal weight. If applied to a maximally consistent semi-complete partially directed graph $\mathcal{S}$, it returns a set $\mathbb{A}_i$ for $\mathcal{S}$ with respect to the consistency matrix W. If $\mathbb{A}_i$ is optimal, meaning that leads to a maximally consistent acyclic tournament, then it holds that:
$$\mathcal{T}_i = (\mathbb{V}, (\mathbb{E}_{\mathcal{S}} \setminus \mathbb{A}_i) \cup \mathbb{A}_i^T)$$
where $\mathbb{A}_i^T$ is the transpose of $\mathbb{A}_i$.
$$\texttt{score}(\mathcal{T}_i) = \texttt{maxScore}$$

Soundness. *Every tournament $\mathcal{T}$ in `Results` is a maximally consistent acyclic tournament.*

- **Initialization:** The algorithm starts by computing the maximally consistent semi-complete partially directed graph $\mathcal{S}$. Then, it computes a FAS $\mathbb{A}_0$ of $\mathcal{S}$ with respect to the consistency matrix W. By definition of `ExactFAS`, a FAS $\mathbb{A}_0$ is a set of edges that can be reversed to transform $\mathcal{S}$ into a maximally weighted acyclic tournament, $\mathcal{T}^0$; the tournament is then added to `Results` and the maximal consistency score `maxScore` is computed.
- **Iteration:** At every iteration, we compute a FAS of the semi-complete partially directed graph $\mathcal{S}$ with respect to the consistency matrix W' in which some edges have been excluded from the solution space. a FAS A is computed by $\texttt{ExactFAS}(\mathbb{E}_{\mathcal{S}}, W')$. The acyclic tournament $\mathcal{T}_A$ and its score are then computed. If the score of $\mathcal{T}_A$ is equal to `maxScore`, it is maximally consistent, thus it is added to `Results`.

Completeness. *Every maximally consistent acyclic tournament $\mathcal{T}$ is in `Results`.*

The solution space is explored by exclusion. The idea is that any maximally consistent solution is unique, thus removing from the solution space subsets of optimal FAS will force `ExactFAS` to search for other maximally consistent solutions. This is achieved by iterating over the power set of the the otimal FAS, contained in `maximalFAS`. To exclude a set of edges $\mathbb{F}$ from the admissible solutions we set the cost of reversing edges in $\mathbb{F}$ to $-\infty$, making them suboptimal by construction. So we can build W' as a copy of W where $W[j,i] = -\infty$ for every edge $(i,j) \in \mathbb{F}$. If there is another maximally consistent solution, A', such that $\mathbb{F} \not\subset \mathbb{A}'$, then it will be found in the next iteration of the algorithm by $\texttt{ExactFAS}(\mathcal{S}, W')$. We define $\mathbb{F}$ as the union of subsets of optimal solutions, $\mathbb{A}_i \in$ `maximalFAS`.

$$\mathbb{F} = \mathbb{F}_0 \cup \mathbb{F}_1 \cup \ldots \cup \mathbb{F}_k$$

where $\mathbb{F}_i \subset \mathbb{A}_i$. Given an undiscovered a set $\mathbb{F}$ of forbidden edges:

- $\mathbb{F} = \emptyset$, which corresponds to the case where the algorithm has not yet explored any subset of the optimal FAS. The algorithm will compute the first FAS $\mathbb{A}_0$ and add it to `Results`.
- $\mathbb{F} \neq \emptyset$, meaning that the algorithm has already explored some subsets of the optimal FAS. In this case, the algorithm will compute a FAS $\mathbb{A}_i$ that is either: 1) optimal, meaning that it is a maximally consistent acyclic tournament, or 2) suboptimal, meaning that it is not a maximally consistent acyclic tournament. In the first case, $\mathbb{A}_i$ will be added to `Results`; moreover, we will add to the `Queue` the combinatorial union between $\mathbb{F}$ all the subsets of $\mathbb{A}_i$, which represents a tigheter constraint on the admissible solutions. Eventually this will lead to the discovery of a new maximally consistent acyclic tournament.

Termination. *The MATS algorithm terminates.*

MATS sistematically explores the power set of the optimal FAS, which is finite. Thus, the algorithm will eventually terminate. Additionally, the algorithms adopts a caching mechanism to avoid exploring the same subsets multiple times. Also, any set leading to suboptimal solution allows to identify sets of edges that can be excluded from the search space. Indeed, any set containing a subset of edges that leads to a suboptimal solution will not be explored.

□

Proposition 6. *If the consisntency matrix is provided by a consistent knowledge bases, the MATS algorithm is guaranteed to return a class of acyclic tournaments containing the true causal order.*

Proof. We assume that the ground truth graph $\mathcal{G} = (\mathbb{V}, \mathbb{E})$ is a causal DAG. We have that an edge in the maximally consistent semi-complete partially directed graph, $\mathcal{S}$, is oriented if and only if the consistency score $C_{X_i \to X_j} > C_{X_j \to X_i}$ and that $X_i \succ_{\mathcal{G}} X_j$. If not, the edge is undirected. In this context, the only cycles that can be present in $\mathcal{S}$ are those that are not oriented, meaning that they are composed only of undirected edges.

□

If the consistency matrix is provided by a strictly consistent knowledge bases, the MATS algorithm is guaranteed to return the true causal order.

Proof. The proof follows from Proposition 6. A strictly consistent knowledge bases guarantees that for every pair of variables X_i and X_j, such that $X_i \succ X_j$, it holds that $C_{X_i \to X_j} > C_{X_j \to X_i}$. Meaning that the maximally consistent semi-complete partially directed graph $\mathcal{S}$ is an acyclic tournament since there cannot be undirected edges. □

C B. Algorithms

Algorithm 1 Maximally Weighted Acyclic Tournaments Search (MATS)

```
Input: $\mathbb{V}$, variables; $W$, weights
Output: $Results$, a set of maximally weighted acyclic tournaments
$\mathcal{S} \leftarrow$ MAXIMALLYWEIGHTEDGRAPH($\mathbb{V}, W$)
$(\mathbb{V}, \mathbb{E}_{\mathcal{S}}) \leftarrow \mathcal{S}$
$\mathbb{A} \leftarrow$ EXACTFAS($\mathbb{E}_{\mathcal{S}}, W$)                ▷ Find feedback arc set
$\mathbb{E}_{\mathcal{T}} \leftarrow (\mathbb{E}_{\mathcal{S}} \setminus \mathbb{A}) \cup \mathbb{A}^T$
$maxScore \leftarrow \sum_{(i,j)\in\mathbb{E}_{\mathcal{T}}} W[i,j]$
$Queue \leftarrow \mathcal{P}(\mathbb{A})$                ▷ Power set of $\mathbb{A}$
$maximalFAS \leftarrow \{\mathbb{A}\}$
$Results \leftarrow \{\mathbb{E}_{\mathcal{T}}\}$
$Cache \leftarrow \emptyset$
while $Queue \neq \emptyset$ do
    $\mathbb{F} \leftarrow$ POP($Queue$)
    $Cache \leftarrow Cache \cup \mathbb{F}$
    $W' \leftarrow$ COPY($W$)
    for $(i,j) \in \mathbb{F}$ do
        $W'[j,i] \leftarrow -\infty$
    $\mathbb{A} \leftarrow$ EXACTFAS($\mathbb{E}_{\mathcal{S}}, W'$)
    $\mathbb{E}_{\mathcal{T}} \leftarrow (\mathbb{E}_{\mathcal{S}} \setminus \mathbb{A}) \cup \mathbb{A}^T$
    $newScore \leftarrow \sum_{(i,j)\in\mathbb{E}_{\mathcal{T}}} W[i,j]$
    if $newScore = maxScore$ and $\mathbb{A} \notin maximalFAS$ then
        $maximalFAS \leftarrow maximalFAS \cup \{\mathbb{A}\}$
        $Results \leftarrow Results \cup \{\mathbb{E}_{\mathcal{T}}\}$
        for $\mathbb{F}' \in \mathcal{P}(\mathbb{A})$ do
            if $\mathbb{F}' \cup \mathbb{F} \notin Cache$ and $\mathbb{F}' \cup \mathbb{F} \notin Queue$ then
                PUSH(Queue, $\mathbb{F}' \cup \mathbb{F}$)
    else
        for $\mathbb{F}' \in Queue$ do
            if $\mathbb{F} \subset \mathbb{F}'$ then
                REMOVE(Queue, $\mathbb{F}'$)
            $Cache \leftarrow Cache \cup \{\mathbb{F}'\}$
return $Results$
```

D C. Implementation Details

Efficient FAS Computation. The implementation of the `ExactFAS` algorithm is based on a method implemented in the `igraph` library. This library is implemented in C, thus providing an efficient implementation of the algorithm. However, the solution to the FAS is computed on the whole graph rather than its strongly connected components (SCCs). To address this, we implemented a preprocessing step that identifies the SCCs of the semi-complete partially directed graph $\mathcal{S}$. The `ExactFAS` algorithm is then applied to each SCC separately, and the results are combined to obtain the final acycluc transformation.

Undirected Edges. To reduce computational time, undirected edges are handled separately. Before applying Algorithm 1, all undirected edges are removed from the graph $\mathcal{S}$. After generating a class of acyclic tournaments, each undirected edge is reintroduced as a directed one, oriented in one of its possible directions, as long as it doesn't create new cycles. Since undirected edges contribute equally to consistency, their orientation does not change the consistency score of the maximal acyclic tournament, but reduces the size of the SCCs processed by `ExactFAS`.

Data Generation. For each node x in the DAG,

$$x = f_x(Parents(x)) + \epsilon_x,$$

where $Parents(x)$ is the set of parents of x in the graph, f_x is a non-linear function, picked randomly from $\{sin(.)$, $cos(.)$, $square(.)\}$, and ϵ_x is sampled from a Gaussian distribution. We want to emphasize that we used non-linear functions to generate the data, but most of the data-driven algorithms tend to perform better when linearity holds. Indeed, we have created a hard experimental condition for conventional discovery methods. However, this choice is based on two considerations: first, non-linear relationships are prevalent in real-world data, and second, we aim to highlight the effectiveness of text-driven algorithms, which are not constrained by observational data and their parametric assumptions. Nevertheless, to provide a more complete overview of the results associated with the data-driven methods, we also report the results obtained by data-driven algorithms on linear synthetic data.

E D. Prompts

You are an expert in the field of $\{field\}$.
The task is to provide causal relationships between the variables.
Keep your answers concise and to the point.
Does $\{var_i\}$ causes $\{var_j\}$?
(A) Yes
(B) No
The answer is:

Table 3: Prompt used to query the LLM for causal relationships.

Provide me with $\{n\_rephrase\}$ rephrased versions of the following sentence: $\{sentence\}$
The rephrased sentences should preserve the semantic meaning, even if absurd, of the original one.
The answers should be in the following format:
1. rephrased sentence 1
2. rephrased sentence 2
3. rephrased sentence 3

Table 4: Prompt used to query the LLM for rephrasing.