Causal Graph Discovery with Retrieval-Augmented Generation based Large Language Models

Anonymous ACL submission

Abstract

Causal graph recovery is traditionally done using statistical estimation-based methods or based on individual's knowledge about variables of interests. They often suffer from data collection biases and limitations of individuals' knowledge. The advance of large language models (LLMs) provides opportunities to address these problems. We propose a novel method that leverages LLMs to deduce causal relationships in general causal graph recovery tasks. This method leverages knowledge compressed in LLMs and knowledge LLMs extracted from scientific publication database as well as experiment data about factors of interest to achieve this goal. Our method gives a prompting strategy to extract associational relationships among those factors and a mechanism to perform causality verification for these associations. Comparing to other LLM-based methods that directly instruct LLMs to do the highly complex causal reasoning, our method shows clear advantage on causal graph quality on benchmark datasets. More importantly, as causality among some factors may change as new research results emerge, our method show sensitivity to new evidence in the literature and can provide useful information for updating causal graphs accordingly.

1 Introduction

Estimating causal effect between variables from observational data is a fundamental problem to many domains including medical science (Höfler, 2005), social science (Angrist et al., 1996), and economics (Imbens and Rubin, 2015; Yao et al., 2021). It enables reliable decision-making from complex data with entangled associations.

While it is usually expensive and infeasible to investigate causal effects by the golden standard—randomized experiments—researchers employ causal inference (Pearl, 2010) to estimate causal effects from observational data. There are two main frameworks for causal inference: the potential outcome framework (Rubin, 1974) and the structural causal model (SCM) (Pearl, 1995). Priori causal structures, usually represented as Directed Graphical Causal Models (DGCMs) (Pearl, 2000; Spirtes et al., 2001), are often used to represent and analyze the causal relationships. These causal graphs help disentangle the complex interdependencies and facilitate the analysis of causal effects. Recovering causal graphs often relies on experts' knowledge or statistical estimation on experimental data (Spirtes and Glymour, 1991). Causal Discovery (CD) algorithms (Spirtes and Glymour, 1991) are the main statistical estimation-based methods that use conditional independence tests to assess associational relationships (called associational reasoning) for inferring causal connections (Spirtes et al., 2001; Chickering, 2002; Shimizu et al., 2006; Sanchez-Romero et al., 2018).

Consequently, the reliability of these algorithms is affected by the quality of data, which can be compromised by issues such as measurement error (Zhang et al., 2017), selection bias (Bareinboim et al., 2014) and unmeasured confounders (Bhattacharya et al., 2021) (See Example A.1 in Appendix A.1). Additionally, CD algorithms often assume certain distribution, such as Gaussian about data, which may fail to accurately reflect the complexity of real-world scenarios.

To mitigate the limitations, Large Language Models (LLMs) (Zhao et al., 2023) have recently been employed for causal graph recovery (Zhou et al., 2023). There are two main streams of these work: 1) directly outputting causal graphs (Choi et al., 2022; Long et al., 2022; Kıcıman et al., 2023); 2) assisting in refining causal graphs generated by statistical estimation-based methods Vashishtha et al. (2023); Ban et al. (2023). Most work have a straightforward way of using LLMs. They directly query the causal relationship between each pair of variables (Choi et al., 2022; Long et al., 2022; Kıcıman et al., 2023) by prompting LLMs with the definition of causality, task details and description of the variables of interest. They require LLMs to have extensive domain knowledge and capabilities to perform complex causal reasoning. Whether LLMs have sufficient knowledge in specific domains or whether they have causal reasoning capabilities are questionable(Kandpal et al., 2023; Zečević et al., 2023).

An alternative approach is to exploit LLMs' capabilities on associational reasoning, e.g., querying the conditional independences (CIs) and recover causal graphs based on extracted associations using CD algorithms(Cohrs et al., 2023). However, it remains difficult for LLMs to understand the CIs between variables, especially the independence conditioned on a large set of variables. (Jiralerspong et al., 2024) tries to inject statistical CI results into LLMs to improve direct causal relationship query results, but the efficacy varies among datasets.

We propose the LLM Assisted Causal Recovery (LACR) method to address these challenges. LACR enhances the knowledge base of LLMs with Retrieval Augmented Generation (RAG) (Lewis et al., 2020; Borgeaud et al., 2022) for reliable associational reasoning. We retrieve highly related knowledge base from a large scientific corpus that contains valuable insight hidden in datasets about associational/causal relationships among variables. We further enhance the accuracy of LCAR's causal recovery results by aggregating the collective extracted information from related literature according the Wisdom of the Crowd principle (Grofman et al., 1983). LACR also uses an associaitonal reasoning-based causal recovery prompt strategy which elaborately instructs the LLMs the mathematical intuitions behind conditional independence, and builds a surjection from conditional independences extracted by LLMs to causal relationships between variables. LACR is data-driven and dos not rely on task-specific knowledge for document retrieval or prompt design. It can serve as a causal graph recovery tool for generic tasks.

Our methodology provides a structured and systematic approach to inferring causal relationships, as it is grounded in a broader evidentiary base and subject to systematic validation. As LACR conducts associational reasoning on a reliable knowledge base, most of which provide evidences based on experimental data analysis, LACR largely overcomes the collection bias problem in statistical estimation-based CD algorithms. We discuss this in detail in Section 4 by pointing out the causal conflict between the well-known causal discovery results and recent research results extracted by LACR.

Our Contributions:

• We introduce a novel RAG-based causal graph recovery method that achieves better associational reasoning. The method shows its potential in accurate causal graph construction and overcoming data collection bias issues in traditional methods.

• We design an associational reasoning-based prompting strategy that reduce LLMs' task complexity to simple associational reasoning to improve the reliability of LLMs' results. The reliability gain further improve the quality of recovered causasl graphs.

• We conduct experiments in several well-known real-world causal graphs and demonstrate the efficacy of LACR. More importantly, based on the scientific evidence returned by our method, we show bias exists in the validation datasets widely used in the CD community, and suggest ways to improve.

2 Background

In this section, we introduce the preliminaries of the *directed graphical causal models* (DGCM) and the *causal graph recovery* problem.

2.1 Directed Graphical Causal Models

A *Directed Graphical Causal Model* (DGCM) is a tuple $M = \langle G, P \rangle$. In the model, $G = \langle V, E \rangle$ is a Directed Acyclic Graph (DAG), also known as a *causal graph*, where the set of nodes $V = \{v_1, \cdots, v_n\}$ represents random variables (with $|V| = n$), and $E \subseteq \{(v_i, v_j) \mid v_i, v_j \in V, v_i \neq v_j\}$ is a set of directed edges, also called *causal edges*, that encode *causal relationships*. Let $\bar{G} = \langle V, \bar{E} \rangle$ be the *skeleton* of DAG G, where each $(v_i, v_j) \in \bar{E}$ is an undirected edge, and it indicates that one of (v_i, v_j) and (v_j, v_i) is in E. Let a sequence of distinct nodes $\ell = (v_{j_1}, v_{j_2}, \cdots, v_{j_m})$ denote a *path*, such that for each $i \in \{1, 2, \cdots, m-1\}$, $(v_{j_i}, v_{j_{i+1}}) \in \bar{E}$. A path is a *causal path* from v_{j_1} to v_{j_m} if for each $i \in \{1, 2, \cdots, m-1\}$, $(v_{j_i}, v_{j_{i+1}}) \in E$. The joint probability distribution of all variables is denoted by P. Note that we do not consider any variable other than those in V, that is, we assume there is no so-called latent or exogenous variable.

Constraints of causal graphs. A causal graph is subject to a series of constraints on variables' *associational relationships*. Especially, the causal edges specify the causal relationships between variables. Given $(v_i, v_j) \in E$, v_i is a *direct cause* of v_j. That is, when holding the other variables constant, varying the value of v_i triggers a corresponding change in the value of v_j, but not vice versa. This causal relationship thus entails the associational relationship between the variables, i.e., their marginal probability distributions $P(v_i)$ and $P(v_j)$ are associated (or correlated), which does not have the direction attribute. Notice that two variables can be associated even though they do not have a direct causal relationship between each other. Typical examples are that two variables linked by a causal path, and two variables pointed to by two causal paths that have the same starting node (which is usually called a covariate). The precise constraints follow an assumption of the causal graph called the Causal Markov Assumption.

Assumption 2.1 (Causal Markov Assumption)**.** *In any causal graph, each variable is independent of its non-descendants conditioned on its parents in the causal graph.*

Therefore, the structure of a causal graph implies graphical constraints called *d-separation* (Pearl, 2000) that specify a conditional associational relationship between variables. In the rest of this paper, for any given variable pair $v_i, v_j \in V$, we constantly use V' to denote an arbitrary subset of $V \setminus \{v_i, v_j\}$, unless otherwise specified.

Definition 2.2 (d-separation)**.** *A variable set V' blocks a path ℓ if (i) ℓ contains at least one arrow-emitting variable belonging to V', or (ii) ℓ contains at least one collider (variable v_i is a collider if $(v_{j_{i-1}}, v_{j_i}), (v_{j_{i+1}}, v_{j_i}) \in E$) that does not belong to V' and has no descendant belonging to V'. If V' blocks all paths from v_i to v_j, V' is said to d-separate v_i and v_j.*

If V' d-separates v_i and v_j, then the joint probability distribution P encodes that the two variables are independent conditioned on V'.

Assumption 2.1 is a necessary condition for the encoding of the associaitonal relationship constraints in P. On the other hand, the following *faithfulness assumption* is a sufficient condition that P encodes such constraints.

Assumption 2.3 (Causal Faithfulness Assumption)**.** *A joint distribution P does not encode additional conditional associational relationships other than those consistent with G's d-separation information. We call such P is faithful to G.*

We now formally define the constraints that follow distribution P faithful to causal graph G. Let $\alpha(ij \mid V') \in \{0, 1\}$ be the *conditional associational relationship* between variables $v_i, v_j \in V$ conditioned on variable set V'. $\alpha(ij \mid V') = 0$ denotes that v_i and v_j are independent conditioned on V' according to P, and $\alpha(ij \mid V') = 1$ denotes associated. We write $\alpha(ij)$ when $V' = \emptyset$.

Then, by Assumptions 2.1 and 2.3 and Definition 2.2, we have that for $v_i, v_j \in V$:
1. V' d-separates v_i and $v_j \implies \alpha(ij \mid V') = 0$;
2. $\alpha(ij) = 1$ and $(v_i, v_j) \notin \bar{E} \implies \exists V'$ s.t. $\alpha(ij \mid V') = 0$;
3. $(v_i, v_j) \in \bar{E} \implies \nexists V'$ s.t. $\alpha(ij \mid V') = 0$.

3 Methodology

We now start to introduce our LLM-based method, called *large language model assisted causal recovery* (LACR), that uses a prompt strategy elaborately designed following the process of a statistical estimation-based CD method, called the *constraint-based causal graph construction* (CCGC). We first show how CCGC works.

3.1 Constraint-based Causal Graph Construction: From Data to Causation

Based on Assumptions 2.1 and 2.3, we are able to partially construct the causal graph G from a *knowledge base* KB that is faithful to G by a statistical estimation-based method. In a nutshell, KB can be but not limited to data, the LLM's background knowledge, and external documents. For more details, see Section 3.2.1. We take data as the KB in CCGC. A KB is called faithful to G if it estimates a joint distribution that is faithful to G.

The process of CCGC can be divided into two phases: the *edge existence verification* phase, which first constructs the skeleton, and the *orientation* phase, which determines the direction of each undirected edge. LACR only uses the CCGC-based prompt strategy to conduct the edge existence verification, and therefore, we only introduce the first phase of CCGC.

For each pair of variables $v_i, v_j \in V$, we verify the existence of the undirected edge in between (i.e., whether $(v_i, v_j) \in \bar{E}$ or not) by statistically testing whether v_i and v_j can be d-separated by any variable set V'. Let $\hat{\alpha}_{\texttt{KB}}(ij \mid V') \in \{0, 1\}$ be an estimator of $\alpha(ij \mid V')$, based on KB. Next, based on a given KB that is faithful to G, we define $\zeta_{\texttt{KB}} : V \times V \to \{0, 1\}$ as the *causal* *edge existence* mapping, such that $\zeta_{\texttt{KB}}(ij) = 0$ if $\exists V'$ s.t. $\hat{\alpha}_{\texttt{KB}}(ij \mid V') = 0$, otherwise $\zeta_{\texttt{KB}}(ij) = 1$. $\zeta_{\texttt{KB}}(ij) = 0$ implies that we estimate there is no edge between v_i and v_j, since the pair of variables can be d-separated by at least one variable set. See Appendix A.1 for an example of CCGC's process.

Compared with LACR, most existing LLM-based causal graph construction methods directly query LLMs the causal relationships. With the introduction of CCGC, we next illustrate the limited reliability of such methods.

Limited Reliability of Direct Causal Prompt. We name the prompt used in such direct query of causal relationships as the *direct causal prompt*. Examples include "Is A a cause of B?" and "Does the change of A cause the change of B", which are wildly used in related work (Kıcıman et al., 2023; Choi et al., 2022; Long et al., 2022). Such prompt directly queries the causal edge existence ($\zeta_{\texttt{KB}}(\cdot)$) and the causal direction.We argue that such direct prompting requires extensive causal reasoning capability from LLMs. The following proposition (see proof in Appendix C.1) shows the high complexity hidden behind a direct causal prompt.

Proposition 3.1. *Assuming that estimating $\hat{\alpha}_{\texttt{KB}}(ij \mid V')$ for a given V' needs $O(1)$ time, inferring $\zeta_{\texttt{KB}}(ij)$ requires $O(2^{n-2})$, where $n = |V|$.*

Proof. The proof is illustrated in Section C.1. □

We now start formally introducing the *large language model assisted causal recovery* (LACR) method, which first extracts the conditional associational relationships between variables, and determines the causal relationships following the process of CCGC (see Section 3.1). We implement such a process by a series of separated queries using the constraint-based causal prompt. The LACR consists of two steps: the edge existence verification (LACR 1) and orientation (LACR 2).

3.2 LACR 1: Edge Existence Verification

In this phase, we construct the skeleton of the causal graph, i.e., verifying the existence of each edge without clarifying its direction. We use LLMs to mine the statistical evidence to verify the conditional associational relationship between each pair of variables and determine the existence of a causal edge (recall Section 3.1), from the retrieved scientific documents (corresponding to document-based query), LLMs' internal knowledge (corresponding to background-based query), and statistical estimation-based output. To achieve this target, we design a prompt strategy that encodes the statistical principles of CCGC.

3.2.1 Constraint-Based Causal (CC) Prompt

In LACR, KB can be the LLM's background knowledge, external documents, and datasets. For each variable pair, namely v_i and v_j, we clarify their conditional associational relationship by mining the statistical evidence from KB. We conduct a chain of 4 queries to determine the final opinion of each piece of KB, namely, the **background reminder**, the **association verifier**, the **association type verifier**, and the **association rechecker**. However, if any KB does not contain sufficient information to determine the value of $\hat{\alpha}_{\texttt{KB}}(ij \mid V')$, we ask the LLM to give an answer UNKNOWN. We classify such knowledge bases as *unusable*, and they are discarded during the decision-making phase. See the original prompt in Appendix C.6.

Background reminder. This prompt component helps the LLM to understand the full picture of the task, and avoid misinterpretation of variables' meaning. We aim to provide minimum external information about the task other than the names of the variables to the LLM. Therefore, we only give the full FACTOR list (i.e., the names of the variables), and specify the DOMAINS from which the variables are. For example, in the ASIA experiment dataset (see Section 4), all variables are from the domains of MEDICAL, BIOLOGY, AND SOCIAL SCIENCE. Finally, we ask the LLM to specify the meaning of each variable, as well as the interaction among them.

Association verifier. The component utilizes KR to verify the zero-order associational relationship between v_i and v_j, i.e., $\hat{\alpha}_{\texttt{KB}}(ij)$. The LLM is provided with an ASSOCIATION CONTEXT (an instruction of how to determine whether v_i and v_j is associated or not) and KB. Then, the LLM determines the relationship as ASSOCIATED (if $\hat{\alpha}_{\texttt{KB}}(ij) = 1$), INDEPENDENT (if $\hat{\alpha}_{\texttt{KB}}(ij) = 0$), or UNKNOWN, based on the statistical evidence extracted from KB. If the decision is an association, the LACR goes to the next query.

Association type verifier. Upon determining ASSOCIATED between v_i and v_j, we further need to determine whether this association is "indirect" or "direct", i.e., whether there exists V' that can d-separate v_i and v_j. Based on the given KB and reasoning, the LLM is asked to read an ACCUSATION TYPE CONTEXT (an instruction of how to judge whether the association is indirect or direct based on KB). Intuitively, the ACCUSATION TYPE CONTEXT illustrates that if the association between v_i and v_j is mediated by variables from V' , then, the association is indirect, otherwise it is direct. To align precisely with CCGC, the ACCUSATION TYPE CONTEXT further explains that "the association mediated by third variables" means that the association is eliminated if we control the third variables constantly. The LACR goes to the final query if the decision is an INDIRECTLY ASSOCIATED.

Association rechecker. Considering the potential that the LLM can return INDIRECTLY ASSOCIATED because it judges that the association between v_i and v_j is mediated by external variables that are not from V. Since we do not consider external variables, we ask the LLM to verify whether the set of mediating variables includes any from $V \setminus \{v_i, v_j\}$. If yes, the association type should be corrected to DIRECTLY ASSOCIATED.

3.2.2 The CC Prompt is Deterministic

Using the above prompt strategy, we demonstrate that the LLM's return can determine the existence of causal edges based on a given KB. We first specify that the LLM's return based on the above prompt must be one from set {INDEPENDENT, DIRECTLY ASSOCIATED, INDIRECTLY ASSOCIATED, UNKNOWN}. If the return is UNKNOWN, the KB is unusable. Then, for each *usable* KB, we have the following proposition (see proof in Appendix C.2).

Proposition 3.2. *For each variable pair v_i and v_j, the mapping from the conditional associational relationship space of v_i and v_j to the return set of each usable* KB *is a surjection, and the mapping from the return set of each usable* KB *to the range of ζ_{ij}, i.e., $\{0, 1\}$, is also a surjection.*

Proof. The proof is illustrated in Section C.2. $\square$

3.2.3 LACR 1

With the above CC prompt, we are ready to introduce LACR 1 (Algorithm 1). We initialize the algorithm by setting the skeleton graph $\bar{G}$ as a complete undirected graph $\bar{G}^c$, giving each variable pair v_i, v_j a pre-retrieved set of k relevant scientific documents as the document-based knowledge base $\mathbf{DOC} = \{\mathbf{DOC}_{ij} = \{\mathtt{DOC}^1_{ij}, \cdots, \mathtt{DOC}^k_{ij}\}\}_{v_i, v_j \in V, \text{ s.t., } v_i \neq v_j}$. Then, for each variable pair, we query the LLM by the CC prompt to estimate $\hat{\zeta}_{\mathtt{KB}}(ij)$ based on each of the given documents provided in $\mathbf{DOC}_{ij}$ and the LLM's background knowledge BG. If the decision of the LLM is INDIRECTLY ASSOCIATED or INDEPENDENT, i.e., $\hat{\zeta}_{\mathtt{KB}}(ij) = 0$, by Proposition 3.2, we add -1 point to the score S, if the decision is DIRECTLY ASSOCIATED (i.e., $\hat{\zeta}_{\mathtt{KB}}(ij) = 1$), we add 1 point to S, otherwise, we do not change S if LLM answers UNKNOWN based on KB. After considering all of the LLM's decisions for v_i and v_j, if the final score $S > 0$, we keep the undirected edge (v_i, v_j); otherwise, we remove it from $\bar{G}$. Finally, the algorithm returns the skeleton after querying each variable pair based on all KBs.

Algorithm 1 LACR 1

Input: $\bar{G} \leftarrow \bar{G}^c$, $\mathbf{DOC}$, $\mathbf{D}$
for $\forall v_i, v_j \in V$, s.t., $v_i \neq v_j$ **do**
  $S = 0$
  for $\mathtt{KB} \in \mathbf{DOC}_{ij} \cup \{\mathtt{BG}\}$ **do**
    if $\hat{\zeta}_{\mathtt{KB}}(ij) = 1$ **then**
      $S += 1$
    else if $\hat{\zeta}_{\mathtt{KB}}(ij) = 0$ **then**
      $S += -1$
  if $S \leq 0$ **then**
    $\bar{G} \leftarrow \bar{G} \backslash (v_i, v_j)$
Return: $\bar{G}$

Note that using the score S to aggregate each KB's "opinion" for each variable pair is equivalent to making the collective decision of $\hat{\zeta}(ij)$ by the simple majority voting rule (Brandt et al., 2016). We slightly bias the decision towards $\hat{\zeta}_{\mathtt{KB}}(ij) = 0$ by the setting of removing an edge if $S \leq 0$, since generally, the LLM's decision biases towards DIRECTLY ASSOCIATED. We use this biased setting because (1) almost KBs cannot load the evidence showing $\hat{\alpha}_{\mathtt{KB}}(ij \mid V')$ for all possible V', and (2) if most retrieved documents are unusable (no research report on v_i and v_j's association), then, it is more possible that v_i and v_j are not associated. By the theory of the Wisdom of the Crowd, LACR's decision tends to be more accurate than querying a single knowledge base, and it can be improved by adding more relevant documents (see a detailed description in Appendix B).

3.3 LACR 2: Orientation

Starting at the skeleton output by LACR 1, we continue to determine the direction of each edge in the skeleton. In LACR 2, we simply utilize direct query

to LLM for the orientation task due to LLMs' high performance on causal orientation tasks (Kıcıman et al., 2023). For each pair of adjacent variables in the skeleton, we use a two-step prompt strategy:
Background reminder. Similar to LACR 1, we provide the main variables and the domain information of the task, and ask the LLM to clarify the variables' meanings, as well as their interaction.
Orienting. With the above clarification, we ask the LLM to thoroughly understand the given KB and a CAUSAL DIRECTION CONTEXT, that specifies that if variable A is the cause of variable B, then, the change of A's value causes a change of B's value, but not vice versa. Then, we ask the LLM to give its decision based on all of the above information.

4 Experiments

In this section, we first introduce the ground truth datasets and how we collect three research literature pools. Then we introduce the settings of our solution and baselines. Finally, we evaluate the pruning and orienting results, respectively.

4.1 Experiment Data

Validation datasets. We validate our method on four datasets (namely, ASIA, SACHS, and CORONARY). All datasets have reported causal graphs (see Appendix C.4) based on real-world data. It is worth noting that, we only limit the selection of validation datasets to real-world datasets because LACR uses a realistic knowledge base.
ASIA (lau, 1988). The ASIA dataset has 8 nodes (from domains of medical, biology, and social science) and 8 edges, revealing the potential reasons and symptoms of lung diseases.
SACHS (Sachs et al., 2005). The SACHS dataset has 11 nodes (from the medical and biological domains) and 16 edges. It uncovers the interaction among proteins related to several human diseases.
CORONARY (Reinis et al., 1981). The CORONARY dataset has 6 nodes (from the medical and biological domains) and 9 *undirected* edges, revealing the causal relationship among several potential reasons of coronary heart disease. We only use it to validate LACR 1 because the edges are undirected.

4.2 Experimental Settings

We use GPT-4o in the following experiments.
Research document pool construction. In our experiment, we automatically build the pre-retrieved document set for each variable pair (**Initialization** in Algorithm 1) in two steps:
(1) Relevant paper search: We search 20 paper titles by querying "name[v_i] and name[v_j]" to the Google Scholar engine using the SerpApi (SerpApi), and rank the papers by Google Scholar's default relevance ranking.
(2) Paper download: Based on the aforementioned ranked paper title list, we use the PubMed API[1] to download the papers. For each paper title, we prioritize downloading the full document from the PubMed Central (PMC) database, and only download the abstract document from the PubMed database if the full version is not available in PMC. for each variable pair, we download up to 10 documents from the top of the ranked title list (note that some papers are unavailable in PubMed).
Statistical causal discovery method. In the validation of LACR 1, additional to LLM, we also test the impact of injecting statistical estimation-based results into the decision-making phase. That is, adding point 1 (resp. -1) to score S if the statistical estimation-based method determines $\hat{\zeta}_{\texttt{KB}}(ij) = 1$ (resp. $\hat{\zeta}_{\texttt{KB}}(ij) = 0$) in Algorithm 1, where KB is numerical data. We use the Peter-Clark (PC)(Spirtes et al., 2001) algorithm as the statistical estimation-based method. We import the data from the bnlearn package (Scutari et al., 2019).
Baseline methods. We survey recent LLM-based causal graph construction methods, and for each dataset, we select the baseline method with the best performance. For each dataset, we present two types of baseline LLMs: baseline LLM1, which is a pure LLM-based method, and baseline LLM2, which is a hybrid method combining a statistical estimation-based and an LLM-based method. We do not compare LACR to any baseline method on the CORONARY dataset as the dataset's absence in such methods' validation.
Validation metrics. We measure LACR 1 and LACR 2 by different metrics. For LACR 1, we show the the adjacency precision (AP), the adjacency recall (AR), the F1 score, and the Normalized Hamming Distance (NHD), as follows.

First, we count three attributes of each graph: true positive (TP): the number of edges that are successfully recovered, false positive (FP): the number of edges that are recovered but different from the ground truth graph, and false negative (FN): the number of edges that exist in the ground truth but not recovered in our constructed graph. Then, we compute AP: $\frac{\texttt{TP}}{\texttt{TP}+\texttt{FP}}$, AR: $\frac{\texttt{TP}}{\texttt{TP}+\texttt{FN}}$, F1: $\frac{2AP*AR}{AP+AR}$, and

[1] https://www.ncbi.nlm.nih.gov/home/develop/api/

	Dataset	AP	AR	F1	SHD
ASIA	LACR 1 (`BG`)	**1**	**1**	**1**	0
	LACR 1 (`DOC`)	0.571	**1**	0.727	0.122
	LACR 1 (`PC`)	**1**	0.75	0.857	0.041
	Baseline LLM1	**1**	0.88	0.93	0.016
	Baseline LLM2	0.8	**1**	0.89	0.031
CORO	LACR 1 (`BG`)	0.625	0.625	0.625	0.167
	LACR 1 (`DOC`)	0.667	0.75	0.706	0.139
	LACR 1 (`PC`)	**0.778**	**0.875**	**0.824**	0.083
SACHS	LACR 1 (`BG`)	**0.8**	0.5	0.615	0.083
	LACR 1 (`DOC`)	0.467	**0.875**	**0.609**	0.149
	LACR 1 (`PC`)	0.421	0.5	0.457	0.157
	Baseline LLM1	N/A	N/A	0.31	0.63
	Baseline LLM2	0.59	N/A	0.56	0.12

Table 1: Performances of our solution LACR 1 with different `KB`. We test the performance across three datasets, and compare to baseline methods: ASIA: LLM1: (Jiralerspong et al., 2024), LLM2: (Jiralerspong et al., 2024), SACHS: LLM1: (Zhou et al., 2024), LLM2: (Takayama et al., 2024).

NHD: $\frac{\texttt{FP}+\texttt{FN}}{n^2}$, where n is the number of variables. Intuitively, NHD is the number different edges between two graphs, normalized by n^2.

In the validation of LACR 2, we simply compute the True Edge Accuracy (TEA), i.e., the ratio of correctly oriented edges among all true positive edges in LACR 1's output skeleton.

4.3 Evaluation

We now first present observations based on experimental results for three datasets, which contains Edge Existence Verification (Section 4.3.1) and Orientation (Section 4.3.2), followed by a comprehensive analysis of the overall results (Section 4.3.3).

4.3.1 Observation on Edge Existence Verification

We present the performance of LACR 1 on causation existence verification with different knowledge bases `KB` in the section. The orienting performance will be shown in the next section. Table 1 lists the performance of all compared methods, where `BG` denotes only LLM's background knowledge, `DOC` denotes both LLM's background knowledge and the fixed number of documents, and `PC` denotes `DOC` plus the results output by the PC algorithm. We have the following observations:

ASIA. We have three observations from the experimental results on the ASIA dataset. First, LACR 1 achieves the best performance when relying solely on `BG`. It successfully recovers the full skeleton and outperforms the high performance of the pure LLM method in (Jiralerspong et al., 2024). Second, adding retrieved documents into `KB` reduces performance (AP from 1 to 0.57, and F1 score from 1 to 0.73) according to the given ground truth in (lau, 1988). Third, by further aggregating the output of the PC algorithm, the F1 score increases from 0.73 to 0.86 compared to the ground truth in (lau, 1988).

CORONARY (CORO). The results differ notably from those based on the ASIA dataset, LACR 1 with only the LLM's background knowledge achieves the worst performance, with values of 0.625 for all of AP, AR, and F1 scores. By adding documents and the PC algorithm into `KB`, all metrics increase, reaching 0.875.

SACHS. We have three observations from the results on the SACHS dataset. First, the best performance of LACR 1 is achieved using only the LLM's background knowledge, outperforming the baseline method (combined method of LLM and hybrid statistical methods of DirectLiNGAM). Second, adding documents as the knowledge base slightly decreases the F1 score by 0.01. Third, with the PC algorithm's output, the F1 score reduces to 0.46, which is worse than the baseline method.

4.3.2 LACR 2: Orientation

The results (Table 3 in Appendix C.5) of LACR 2 show TEA is 1 (i.e., orienting all `TP` edges correctly) on both ASIA and SACHS, based on all `KB`. We can observe that, upon successfully recovering causal edges by LACR 1, the orientation accuracy is high, reaching 1 for all knowledge bases and all datasets. It demonstrates the efficacy of the orientation prompt as well as LLM's capability for causal orientation reasoning. We conjecture that the success of this task strongly depends on the rich evidence stored in the scientific literature, and the easy understandability of such evidence, compared to the extraction of associational relationship.

4.3.3 Overall Results Analysis

By summarizing the overall performance of LACR, it is worth noticing the following points:

LACR's performance tends to monotonically increase by taking more `KB` with high quality and readability. Observing LACR's performance on ASIA and CORONARY, we notice that our methods generally perform better with high-quality and readability of the input documents and statistical results. While we discuss the performance drop of LACR1 (`DOC`) on the ASIA dataset later, the overall trend coincides with the Condorcet theorem (Grofman et al., 1983) in voting theory, which suggests that aggregating diverse, high-quality inputs leads to better outcomes.

LACR performs differently on tail and non-tail data. We observe that LACR performs better on ASIA and CORONARY datasets compared to the SACHS dataset. This is because the terms in ASIA and CORONARY are more common to LLMs during training. In contrast, SACHS mainly contains symbols with specific meanings in a specific area. Despite feeding scientific documents to LACR on all datasets, the lack of prior knowledge of these symbols in the training phase limits LLMs' understanding of their meanings, resulting in hallucinations. This has been observed in RAG-based legal research tools (Magesh et al., 2024).

Updating on the current ground truth causal graphs is necessary. We found, through LACR's responses, strong evidence from domain research (see details in Appendix C.3) indicating that an update to the ground truth causal graphs is necessary. For example, on the ASIA dataset, the ground truth being outdated led to reduced performance when using additional documents. Similarly, for the CORONARY dataset, the improvement in performance with added documents and the PC algorithm suggests that the ground truth for CORONARY is more current compared to ASIA.

4.4 Refining the Ground Truth (ASIA and CORONARY)

The ground truth causal graph needs refinement because it is outdated and significantly differs from current SOTA domain knowledge. Additionally, we aim to determine if the LLM can identify new causal relationships based on SOTA literature. In this part, we present the observations and analyses on the refined ground truth causal graphs based on the refined ASIA and CORONARY datasets. We will discuss how refined graph truth affects performance from three perspectives: causal inferring based on LLM'S background knowledge `BG`, external literature `DOC`, and statistic data `PC`. Due to the page limitation, the details of refining are illustrated in Section C.3.

Background Knowledge `BG`. The performance relying solely on background knowledge `BG` varies between the two datasets. In the ASIAN dataset, all results slightly drop down, whereas in the dataset, results improve. A possible reason is that the background knowledge `BG` related to ASIA is outdated, while the knowledge `BG` for CORONARY is more current. Consequently, when the ground truth is updated based on new domain-specific knowledge, the outcomes are different significantly between the datasets.

	Dataset	AP	AR	F1	SHD
ASIA	LACR 1 (`BG`)	1	0.8	0.889	0.041
	LACR 1 (`DOC`)	0.714	1	0.833	0.082
	LACR 1 (`PC`)	1	0.6	0.75	0.082
CORO	LACR 1 (`BG`)	0.75	0.75	0.75	0.111
	LACR 1 (`DOC`)	0.778	0.875	0.824	0.083
	LACR 1 (`PC`)	0.667	0.75	0.706	0.139

Table 2: Performances of LACR 1 on the refined ground truth with different `KB`, comparing to baseline LLM-powered methods.

External Literature `DOC`. Incorporating `DOC` significantly increases performance for both datasets, as the refind ground truth better aligns with SOTA research trends, demonstrating the importance of up-to-date and relevant domain knowledge in improving model accuracy.

Statistic Data `PC`. Different from the results of incorporating `DOC`, adding `PC`'s results consistently worsens performance, highlighting the PC-based solution is using an outdated dataset compared to updated SOTA knowledge. As the ground truth evolves to reflect current research advancements, the relative performance of `PC`-based results diminishes. These findings emphasize the importance of up-to-date domain-specific knowledge for accurate causal graph recovery.

5 Conclusion

In this paper, we proposed a novel LLM-based causal graph construction method called LACR which uses the constraint-based causal prompt strategy designed according to the constraint-based causal graph construction (CCGC) method. Comparing to most existing LLM-based causal graph construction methods, that use the direct causal prompt to query LLMs to do highly complex causal reasoning, LACR mainly relies on LLMs to do low-complexity associational reasoning, and follows the process of CCGC to determine the causal relationships. For accurate associational reasoning, we utilize LLMs' RAG feature to extract statistical evidence with high relevance and quality from a large scientific corpus. Lastly, we validate LACR's efficacy on several well-known datasets and show LACR's outstanding performance among LLM-based methods. More importantly, LACR's responses show the conflict between the ground truths and SOTA domain research, which requests a refinement of the validation ground truths.

Limitations

We first address three technical limitations of the current version of LACR. The first is the paper search accuracy. The pre-retrieved document set needs high quality and relevance to provide relevant evidence. Therefore, we conjecture that using refined queries and other search engines can enhance the performance. The second limitation is LLMs' understanding on highly professional documents. Through our experiments, we found that LLMs' poor comprehension capability on specific domains, e.g., the SACHS dataset, limits LACR's performance. An optional solution is to fine-tune LLMs to better understand such documents. The third is the complexity of LACR. The method needs to query each variable pair ($O(n^2)$), and for each variable pair, multiple documents need to be queried.

We then address other practical limitations. What comes first is the need of up-to-data practical validation datasets and causal graphs in causal discovery community. Many validation datasets are synthesized, which are not usable in such practical knowledge-based methods. The second practical limitation is the access of scientific papers. In our experiment, we focus on biomedical datasets for the accessibility of research papers in PubMed, however, the full contexts of most of the papers are not open accessible. It would open the possibility of overall better understanding of causal relationships if full documents are accessible in more research domains and broader scientific databases.

References

1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194.

Roya Alavi-Naini, Batool Sharifi-Mood, and Maliheh Metanat. 2012. Association between tuberculosis and smoking. *International journal of high risk behaviors & addiction*, 1(2):71.

Genet A Amere, Pratibha Nayak, Argita D Salindri, KM Venkat Narayan, and Matthew J Magee. 2018. Contribution of smoking to tuberculosis incidence and mortality in high-tuberculosis-burden countries. *American journal of epidemiology*, 187(9):1846–1855.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data.

Elias Bareinboim, Jin Tian, and Judea Pearl. 2014. Recovering from selection bias in causal and statistical inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

ELIZABETH Barrett-Connor and K-T Khaw. 1984. Family history of heart attack as an independent predictor of death due to cardiovascular disease. *Circulation*, 69(6):1065–1069.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. 2021. Differentiable causal discovery under unmeasured confounding. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 2314–2322. PMLR.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. *Handbook of computational social choice*. Cambridge University Press.

Kai-Chi Chen, Hong-Ren Yu, Wei-Shiang Chen, Wei-Che Lin, Yi-Chen Lee, Hung-Hsun Chen, Jyun-Hong Jiang, Ting-Yi Su, Chang-Ku Tsai, Ti-An Tsai, et al. 2020. Diagnosis of common pulmonary diseases in children by x-ray images and deep learning. *Scientific Reports*, 10(1):17374.

David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Kristy Choi, Chris Cundy, Sanjari Srivastava, and Stefano Ermon. 2022. Lmpriors: Pre-trained language models as task-specific priors. *arXiv preprint arXiv:2210.12530*.

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?"*.

David Edwards. 2000. *Introduction to graphical modelling*. Springer Science & Business Media.

Bernard N. Grofman, Guillermo Owen, and Scott L. Feld. 1983. Thirteen theorems in search of the truth. *Theory and Decision*, 15:261–278.

Himanshu Gupta, Sanjeev Mahajan, Mohan Lal, Adarshjot Kaur Toor, Shyam Sunder Deepti, and Naresh Chawla. 2022. Prevalence of tobacco consumption and smoking and its effect on outcome among microbiologically confirmed new pulmonary tuberculosis patients on daily regimen of dots in amritsar city. *Journal of Family Medicine and Primary Care*, 11(5):2150–2156.

T Hintsa, M Shipley, D Gimeno, M Elovainio, T Chandola, M Jokela, L Keltikangas-Järvinen, J Vahtera, MG Marmot, and M Kivimäki. 2010. Do family history of chd, education, paternal social class, number of siblings and height explain the association between psychosocial factors at work and coronary heart disease? the whitehall ii study. *Occupational and environmental medicine*, 67(5):330.

Marc Höfler. 2005. Causal inference based on counterfactuals. *BMC medical research methodology*, 5(1):1–12.

David J Horne, Monica Campo, Justin R Ortiz, Eyal Oren, Matthew Arentz, Kristina Crothers, and Masahiro Narita. 2012. Association between smoking and latent tuberculosis in the us population: an analysis of the national health and nutrition examination survey. *PloS one*, 7(11):e49050.

Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Xuefeng Jin, Caiyun Zhang, Chao Chen, Xiaoning Wang, Jing Dong, Yuanyuan He, and Peng Zhang. 2023. Tropheryma whipplei-induced plastic bronchitis in children: a case report. *Frontiers in Pediatrics*, 11:1185519.

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

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

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

Seung Hoon Kim, Yong-Moon Park, Kyungdo Han, Seung Hyun Ko, Shin Young Kim, So Hyang Song, Chi Hong Kim, Kyu Yeon Hur, and Sung Kyoung Kim. 2022. Association of weight change following smoking cessation with the risk of tuberculosis development: A nationwide population-based cohort study. *Plos one*, 17(4):e0266262.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Ryan P Lindsay, Sanghyuk S Shin, Richard S Garfein, Melanie LA Rusch, and Thomas E Novotny. 2014. The association between active and passive smoking and latent tuberculosis infection in adults and children in the united states: results from nhanes. *PloS one*, 9(3):e93137.

Stephanie Long, Tibor Schuster, and Alexandre Piché. 2022. Can large language models build causal graphs? In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*.

Prince Ntiamoah, Sanjay Mukhopadhyay, Subha Ghosh, and Atul C Mehta. 2021. Recycling plastic: diagnosis and management of plastic bronchitis among adults. *European Respiratory Review*, 30(161).

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

Judea Pearl. 2000. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York.

Judea Pearl. 2010. Causal inference. *Causality: objectives and assessment*, pages 39–58.

Z Reinis, J Pokornỳ, V Bazika, J Tiserova, K Gorican, D Horakova, E Stuchlikova, T Havranek, and F Hrabovskỳ. 1981. Prognostic value of the risk profile in the prevention of ischemic heart disease. *Bratislavske Lekarske Listy*, 76(2):137–150.

Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Ruben Sanchez-Romero, Joseph D Ramsey, Kun Zhang, MR K Glymour, Biwei Huang, and Clark Glymour. 2018. Causal discovery of feedback networks with functional magnetic resonance imaging. *bioRxiv*, page 245936.

Marco Scutari, Maintainer Marco Scutari, and Hiton-PC MMPC. 2019. Package 'bnlearn'. *Bayesian network structure learning, parameter learning and inference, R package version*, 4(1).

SerpApi. Google scholar api.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. 2006. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Peter Spirtes and Clark Glymour. 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search*. The MIT Press.

Masayuki Takayama, Tadahisa Okuda, Thong Pham, Tatsuyoshi Ikenoue, Shingo Fukuma, Shohei Shimizu, and Akiyoshi Sannai. 2024. Integrating large language models in causal discovery: A statistical causal approach. *arXiv preprint arXiv:2402.01454*.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. 2023. Causal inference using llm-guided discovery.

Jianming Wang and Hongbing Shen. 2009. Review of cigarette smoking and tuberculosis in china: intervention is needed for smoking cessation among tuberculosis patients. *BMC public health*, 9:1–9.

Ming-Gui Wang, Wei-Wei Huang, Yu Wang, Yun-Xia Zhang, Miao-Miao Zhang, Shou-Quan Wu, Andrew J Sandford, and Jian-Qing He. 2018. Association between tobacco smoking and drug-resistant tuberculosis. *Infection and drug resistance*, pages 873–887.

Caroline E Wright, Katie O'Donnell, Lena Brydon, Jane Wardle, and Andrew Steptoe. 2007. Family history of cardiovascular disease is associated with cardiovascular responses to stress in healthy young men and women. *International Journal of Psychophysiology*, 63(3):275–282.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46.

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*.

Kun Zhang, Mingming Gong, Joseph Ramsey, Kayhan Batmanghelich, Peter Spirtes, and Clark Glymour. 2017. Causal discovery in the presence of measurement error: Identifiability conditions. *arXiv preprint arXiv:1706.03768*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Guanglin Zhou, Shaoan Xie, Guangyuan Hao, Shiming Chen, Biwei Huang, Xiwei Xu, Chen Wang, Liming Zhu, Lina Yao, and Kun Zhang. 2023. Emerging synergies in causality and deep generative models: A survey.

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. 2024. Causalbench: A comprehensive benchmark for causal learning capability of large language models. *arXiv preprint arXiv:2404.06349*.

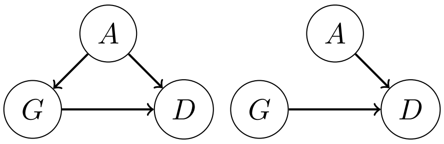

Figure 1: Causal graphs in Example A.1: left-the truth causal graph; right-recovered causal graph by the biased data.

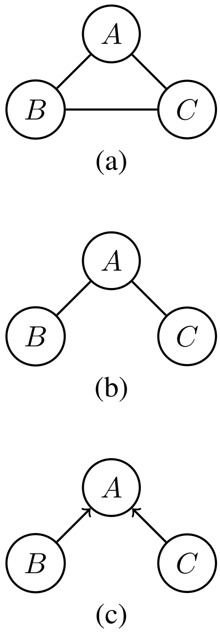

Figure 2: PC algorithm's process.

A Appendix

A.1 Examples

As follows, we first show an example of statistical estimation-based methods' vulnerability to a type of data bias, the so-called selection bias (Bareinboim et al., 2014).

Example A.1. *Consider that we would like to investigate the causal relationship of three variables: A (human age), G (human gender), and D (some disease). Assume that the true causal graph is the left figure in Figure 1.*

Generally speaking, human age and gender are associated because female has a longer average lifespan. Assuming that this association is only significant for $A \geq 60$*. However, if each point in a dataset has age under* 60*, we cannot observe significant difference between the population of male and female. Then, we would recover the causal graph as the right figure in Figure 1.*

The second example shows the processing of a well-known constraint-based causal graph discovery algorithm called PC algorithm.

Example A.2. *Consider a causal discovery task for three variables* A*,* B*, and* C*, and two different joint probability distributions* P^1 *and* P^2*. We start with a complete undirected graph Figure (a) 2.*

Then, by P^1*, we conduct the zero-order independence tests and obtain:* $\hat{\alpha}(AB) = 1$*,* $\hat{\alpha}(AC) = 1$*, and* $\hat{\alpha}(BC) = 0$*. Then, we keep edges* (A, B) *and* (A, C)*, and remove* (B, C)*, and obtain Figure (b) 2, since* B *and* C *are not a cause of each other, otherwise they must be associated. Based on the zero-order tests, we can already determine the causal graph as Figure (c) 2, as* A *must be a collider since* B *and* C *are d-separated by* $\emptyset$*.*

On the other hand, if we consider P^2*, we first have zero-order tests showing all pairs are associated, and we cannot remove any edge in Figure (a) 2. We then conduct first-order tests, and obtain:* $\hat{\alpha}(AB \mid C) = 1$*,* $\hat{\alpha}(AC \mid B) = 1$*, and* $\hat{\alpha}(BC \mid A) = 0$*. Therefore, we can remove the edge* (B, C) *from Figure (a) 2, and obatin Figure (b) 2. However, we cannot determine the directions of the edges because all directions of* $A \rightarrow B \rightarrow C$*,* $A \leftarrow B \leftarrow C$*,* $A \leftarrow B \rightarrow C$ *indicate the conditional independences consistent with* P^2*.*

B Enhancing Skeleton Estimation Accuracy by LACR

The theory of Wisdom of the Crowd (Grofman et al., 1983) states that if (1) each individual voter can make the correct decision better than random decision (e.g., by a toss), and (2) voters make their decision independently, then, the accuracy of the collective decision made by simple majority monotonically increases with the number of voters. In LACR, each `KB` can be seen as a voter. Generally the above conditions tend to be guaranteed because (1) both `BG` and `DOC` have high quality and the delivered information is better than random information, and (2) different research papers deliver their results in a relatively independent way because of scientific integrity. Therefore, LACR's decision tends to be more accurate than querying single knowledge base, and it can be improved by adding more relevant documents.

C Proofs

C.1 Proof of Proposition 3.1

Proof. To verify if $\zeta_{\texttt{KB}}(ij) = 0$, by the definition of $\zeta(ij)$, we need to check whether there exists a variable set V' such that V' d-separates v_i and

v_j. That is, $\alpha_{\texttt{KB}}(ij \mid V') = 0$. Then, the worst case is that we need to check every combination of $V' \subseteq V \backslash \{v_i, v_j\}$, which needs $O(2^{n-2})$ time. □

C.2 Proof of Proposition 3.2

Proof. For each usable knowledge base KB, any possible return through the CC prompt must from the set {DIRECTLY ASSOCIATED, INDIRECTLY ASSOCIATED, INDEPENDENT}.

We first show the first half of the proposition, i.e., the mapping from the conditional associational relationship space between v_i and v_j, i.e., $(\hat{\alpha}_{\texttt{KB}}(ij \mid V'))_{V' \subseteq V \backslash \{v_i, v_j\}}$, to LLM's return space based on each usable KB, i.e., {INDEPENDENT, DIRECTLY ASSOCIATED, INDIRECTLY ASSOCIATED} is a surjection. Note that $(\hat{\alpha}_{\texttt{KB}}(ij \mid V'))_{V' \subseteq V \backslash \{v_i, v_j\}}$ forms a $2^{|V|-2}$-dimensional vector, recording $\alpha_{\texttt{KB}}(ij \mid V') \in \{0, 1\}$ for all possible V'. We discuss three exclusive cases:

1. $\hat{\alpha}_{\texttt{KB}}(ij) = 0$. That is, the zero-order conditional associational relationship between v_i and v_j is independent. This case is mapped to LLM return INDEPENDENT.

2. For all possible V', $\hat{\alpha}_{\texttt{KB}}(ij \mid V') = 1$. The case denotes that v_i and v_j are always associated conditioned on any possible V', and therefore, this case is mapped to LLM return DIRECTLY ASSOCIATED.

3. $\hat{\alpha}_{\texttt{KB}}(ij) = 1$ and $\exists V'$ such that $|V'| \geq 1$ and $\hat{\alpha}_{\texttt{KB}}(ij \mid V') = 0$. In this case, controlling variables in V', the statistical association between v_i and v_j is eliminated, and then it is mapped to LLM return INDIRECT ASSOCIATED.

We then show the second half of the proposition, i.e., the mapping from {INDEPENDENT, DIRECTLY ASSOCIATED, INDIRECTLY ASSOCIATED} to $\{\zeta_{\texttt{KB}}(ij) = 0, \zeta_{\texttt{KB}}(ij) = 1\}$ is a surjection. If the return is DIRECTLY ASSOCIATED, the KB specifies that v_i and v_j cannot be d-separated, and therefore, it is mapped to $\zeta_{\texttt{KB}}(ij) = 1$. On the other hand, if LLM return is INDEPENDENT or INDIRECTLY ASSOCIATED, then, it indicates that V' exists that can d-separate v_i and v_j, where INDEPENDENT corresponds to $V' = \emptyset$. Therefore, these two last cases correspond to $\zeta_{\texttt{KB}}(ij) = 0$. □

C.3 Evidences of Ground Truth Refinement

C.3.1 ASIA

Smoking and Tuberculosis In the documents (Wang and Shen, 2009; Horne et al., 2012; Kim et al., 2022; Wang et al., 2018; Gupta et al., 2022; Lindsay et al., 2014; Amere et al., 2018; Alavi-Naini et al., 2012) fed into LLM as the KB, strong evidence shows that Smoking and Tuberculosis are associated, and the association cannot be eliminated by controlling the other variables in the ASIA dataset. This conflicts against the conditional associational relationship between these two variables in the ground truth causal graph (Appendix C.4), since both of the only two paths have a collider, which indicates that Smoking and Tuberculosis are independent from each other. Based on the scientific evidence returned by LACR, a causal link exists between the two factors.

Bronchitis and X-ray Documents based on LACR's response, (Jin et al., 2023; Ntiamoah et al., 2021) show that an association exists between Bronchitis and Positive X-ray report. (Chen et al., 2020) further develops a deep-learning-based method to detect bronchitis directly from X-ray reports for children with age from 1-17 years old. The evidence shows an association between the two variables, and the association is not mediated by the variable "Smoking" as indicated by the causal graph. Therefore, we add a causal link between Bronchitis and X-ray in the ground truth causal graph.

C.3.2 CORONARY

Strenuous Mental Work and Family Anamnesis Of Coronary Heart Disease According to the ground truth causal graph skeleton (Reinis et al., 1981), there is a direct causal relationship between variable Strenuous Mental Work and variable Family Anamnesis Of Coronary Heart Disease. However, according to the evidence returned by our method and the intuitive description in (Reinis et al., 1981), we observe that this causal linkage should be removed with high probability. By (Edwards, 2000) (p.26), this edge is not intuitively expected though it is recovered by the Bayesian learning method. Additionally, non of LACR's responses suggests the direct association between the two variables, and moreover, it returns evidences showing the positive probability to remove the edge. (Wright et al., 2007) shows that people with Family Anamnesis Of Coronary Heart Disease are easier to

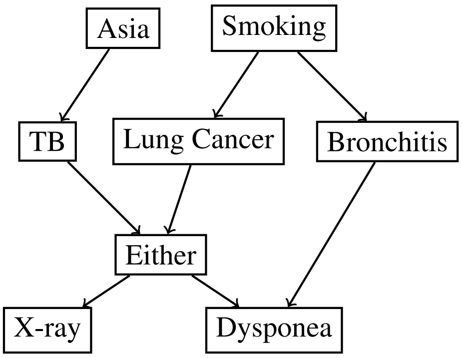

Figure 3: Ground truth causal graph of ASIA in (lau, 1988).

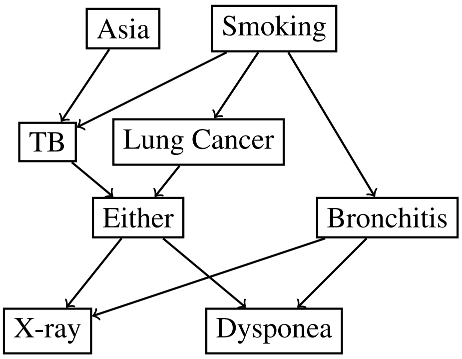

Figure 4: Refined ground truth causal graph of ASIA by LACR.

react to mental stress from work by higher Systolic Blood Pressure. (Hintsa et al., 2010) shows that the association between psychosocial factors at work and coronary heart disease is largely independent from the Family Anamnesis Of Coronary Heart Disease.

Systolic Blood Pressure and Family Anamnesis Of Coronary Heart Disease LACR also returns evidence (Barrett-Connor and Khaw, 1984) showing that Family Anamnesis of Coronary Heart Disease and Systolic Blood Pressure are associated even after the adjustment of several variables including Smoking. We therefore also add this edge between the two variables.

C.4 Additional Experiment Details

The ground truth causal graphs of all datasets in Section 4.

C.5 Additional Experimental Results

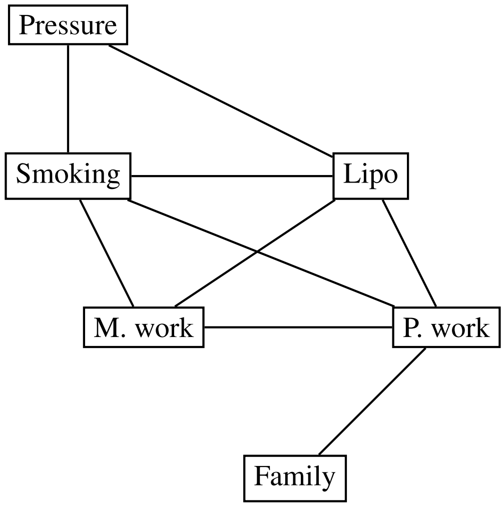

Figure 5: Original ground truth causal graph of CORONARY in (Reinis et al., 1981).

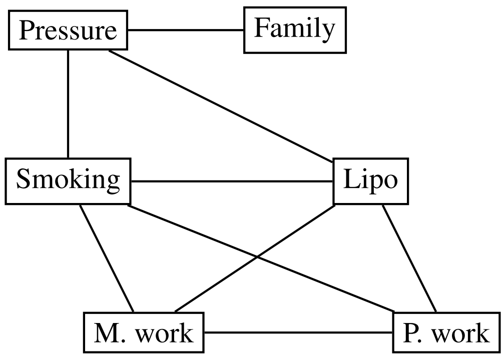

Figure 6: Refined ground truth causal graph of CORONARY by LACR.

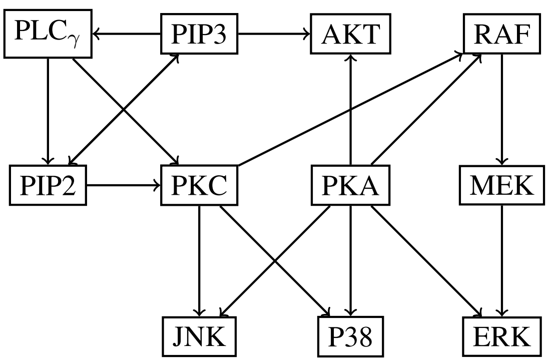

Figure 7: Biological ground truth causal graph in (Sachs et al., 2005).

	ASIA	SACHS
LACR2 (BG)	1	1
LACR2 (DOC)	1	1
LACR2 (PC)	1	1

Table 3: The TEA of LACR 2 on datasets of ASIA, SACHS, based on LACR 1's output skeleton on KB of BG, DOC, and PC, respectively.

C.6 Prompts

C.6.1 Association Context

````
The association relationship between two factors A
and B can be associated or independent, and this
association relationship can be clarified by the
following principles:

1. If A and B are statistically associted or
correlated, they are associated, otherwise they are
independent.
2. The association relationship can be strongly
clarified if there is statistical evidence
supporting it.
3. If there is no obvious statistical evidence
supporting the association relationship between A
and B, it can also be clarified if there is any
evidence showing that A and B are likely to be
associated or independent statistically.
4. If there is no evidence to clarify the
association relationship between A and B, then it is
 unknown.
````

C.6.2 Association Type Context

````
If two factors A and B are associated, they may be
directly associated or indirectly associated with
respect to a set of Given Third Factors, and it can
be clarified by the following principle:

1. The first principle is to try to find statistical
 evidence from the given knowledge to clarify the
following association types. If you cannot find
statistical evidence, at lease find evidence that is
 likely to be able to statistically clarify the
association type between A and B. If no obvious
evidence can be found, the association type is
unknown.
2. If the evidence shows that any factors from the
Given Third Factors mediate the association between
A and B, then A and B are indirectly associated via
these factors.
3. If the evidence shows that by controlling any
factors from the Given Third Factors, A and B are
not associated any more, then A and B are associated
 indirectly.
4. If the evidence shows that A and B are still
associated even if we control any of the given third
 factors, then A and B are directly associated.
5. If you think A and B are indirectly associated
via any of the given third factors, it must be true
that: (1) A and the third factors are directly
associated; (2) B and the third factors are directly
 associated.
````

C.6.3 Association Background Reminder

````
As a scientific researcher in the domains of {domain
}, you need to clarify the statistical relationship
between some pairs of factors. You first need to get
 clear of the meanings of the factors in {factors},
which are from your domains, and clarify the
interaction between each pair of those factors.
````

C.6.4 LLM Association Query (with documents)

````
Your task is to thoroughly read the given 'Document
'. Then, based on the knowledge from the given '
Document', try to find statistical evidence to
clarify the association relationship between the
pair of 'Main factors' according to the 'Association
 Context' (delimited by double dollar signs).
Consider the given document and the association
context. Answer the 'Association Question', write
your thoughts, and give the reference in the given
document. Respond according to the first expected
format (delimited by double backticks).

Document:
{document}

Main factors:
{factorA} and {factorB}

Association Context:
$$
{association_context}
$$

Association Question:
Are {factorA} and {factorB} associated?

First Expected Response Format:
``
Document Identifier: XXX

Thoughts:
[Write your thoughts on the question]

Answer:
(A) Associated
(B) Independent
(C) Unknown

Reference:
[Skip this if you chose option C above. Otherwise,
provide a supporting sentence from the document for
your choice]
``
````

C.6.5 LLM Association Type Query (with documents)

````
Read and understand the Association Type Context.
Consider carefully the role of any of the third
factors appearing according to the Association Type
Context. Then, based on your thoughts so far, answer
 the 'Association Type Question' with the 'Given
Third Factors', write your thoughts, and give your
reference in the given document. Respond according
to the expected format (delimited by triple
backticks)

Association Type Context:
$$$
{association_type_context}
$$$

Given Third Factors:
{factors} except for {factorA} and {factorB}

Association Type Question: Are {factorA} and {
factorB} directly associated or indirectly
associated?

Second Expected Response Format:
```
Thoughts:
[Write your thoughts on the question]

Answer:
(D) Directly Associated
(E) Indirectly Associated
(C) Unknown

Reference:
[Skip this if you chose option C above. Otherwise,
provide a supporting sentence from the document for
your choice]

Intermediary Factors:
[Skip this if you did not choose D or C above.
Otherwise list all factors involved in this indirect
 association relationship, each separated by a comma
]
```
````

C.6.6 LLM Association Query (with background knowledge)

````
Your task is to thoroughly use the knowledge in your
 training data to solve a task. Your task is: based
on your background knowledge, try to find
````

~~~
statistical evidence to clarify the association
relationship between the pair of 'Main factors'
according to the 'Association Context' (delimited by
 double dollar signs).
Consider your background knowledge and the
association context. Answer the 'Association
Question', and write your thoughts. Respond
according to the 'First Expected Format' (delimited
by double backticks).

Main factors:
{factorA} and {factorB}

Association Context:
$$
{association_context}
$$

Association Question:
Are {factorA} and {factorB} associated?

First Expected Response Format:
``
Thoughts:
[Write your thoughts on the question]

Answer:
(A) Associated
(B) Independent
(C) Unknown
``
~~~

C.6.7 LLM Association Type Query (with background knowledge)

~~~
Read and understand the 'Association Type Context'.
Consider carefully the role of any of the third
factors appearing according to the Association Type
Context. Then, based on your thoughts so far, answer
 the 'Association Type Question' with the 'Given
Third Factors', and write your thoughts. Respond
according to the Second Expected Format (delimited
by triple backticks)

Association Type Context:
$$$
{association_type_context}
$$$

Given Third Factors:
{factors} except for {factorA} and {factorB}

Association Type Question: Are {factorA} and {
factorB} directly associated or indirectly
associated?

Second Expected Response Format:
```
Thoughts:
[Write your thoughts on the question]

Answer:
(D) Directly Associated
(E) Indirectly Associated
(C) Unknown

Intermediary Factors:
[Skip this if you did not choose D or C above.
Otherwise list all factors involved in this indirect
 association relationship, each separated by a comma
]
```
~~~

C.6.8 LLM Rethink Query

~~~
If none of the Intermediary Factors you found is not
 in the Given Third Factor list, then, the
association type between A and B is direct
association.
Check your above response, and answer the
Association Type Question again. Respond according
to the Second Expected Format (delimited by triple
backticks).

Given Third Factors:
{factors} except for {factorA} and {factorB}

Association Type Question: Are {factorA} and {
factorB} directly associated or indirectly
associated?

Second Expected Response Format:
```
Thoughts:
[Write your thoughts on the question]

Answer:
(D) Directly Associated
(E) Indirectly Associated
(C) Unknown

Intermediary Factors:
[Skip this if you did not choose D or C above.
Otherwise list all factors involved in this indirect
 association relationship, each separated by a comma
]
```
~~~

C.6.9 Causal Background Reminder

~~~
As a scientific researcher in the domains of {domain
}, you need to clarify the statistical relationship
between some pairs of factors. You first need to get
 clear of the meanings of {factorA} and {factorB},
which are from your domains, and clarify the
interaction between them.
~~~

C.6.10 LLM Causal Direction Query (with background knowledge)

~~~
Your task is to thoroughly use the knowledge in your
 training data to solve a task. Your task is: based
on your background knowledge, try to find
statistical evidence to clarify the direction of the
 causal relationship between the pair of 'Main
factors' according to the 'Causal direction context'
 (delimited by double dollar signs).
Consider according to your background knowledge and
the 'Causal direction context'. Answer the 'Causal
direction question', and write your thoughts.
Respond according to the 'Expected Format' (
delimited by double backticks).

Main factors:
{factorA} and {factorB}

Causal direction context:
$$
{causal_direction_context}
$$

Causal direction question:
Is {factorA} the cause of {factorB}, or {factorB}
the cause of {factorA}?

First Expected Response Format:
``
Thoughts:
[Write your thoughts on the question]

Answer:
(A) {factorA} is the cause of {factorB}
(B) {factorB} is the cause of {factorA}
(C) Unknown
```
~~~

C.6.11 LLM Causal Direction Query (with documents)

~~~
Your task is to thoroughly read the 'Given document'
 to solve a task. Your task is: based on the 'Given
document', try to find statistical evidence to
clarify the direction of the causal relationship
between the pair of 'Main factors' according to the
'Causal direction context' (delimited by double
dollar signs).
~~~

```
First thoroughly read and understand the Given
document and the 'Causal direction context'. Then,
Answer the 'Causal direction question', and write
your thoughts. Respond according to the 'Expected
Format' (delimited by double backticks).

Given document:
{document}

Main factors:
{factorA} and {factorB}

Causal direction context:
$$
{causal_direction_context}
$$

Causal direction question:
Is {factorA} the cause of {factorB}, or {factorB}
the cause of {factorA}?

First Expected Response Format:
``
Thoughts:
[Write your thoughts on the question]

Answer:
(A) {factorA} is the cause of {factorB}
(B) {factorB} is the cause of {factorA}
(C) Unknown

Reference:
[Skip this if you chose option C above. Otherwise,
provide a supporting sentence from the document for
your choice]
```