IRIS: An Iterative and Integrated Framework for Verifiable Causal Discovery in the Absence of Tabular Data

Tao Feng[1], Lizhen Qu[1] *, Niket Tandon[2], Gholamreza Haffari[1]
[1]Monash University, [2]Microsoft Research, India
[1] firstname.lastname@monash.edu, [2] nikett@gmail.com

Abstract

Causal discovery is fundamental to scientific research, yet traditional statistical algorithms face significant challenges, including expensive data collection, redundant computation for known relations, and unrealistic assumptions. While recent LLM-based methods excel at identifying commonly known causal relations, they fail to uncover novel relations. We introduce IRIS (**I**terative **R**etrieval and **I**ntegrated **S**ystem for Real-Time Causal Discovery), a novel framework that addresses these limitations. Starting with a set of initial variables, IRIS automatically collects relevant documents, extracts variables, and uncovers causal relations. Our hybrid causal discovery method combines statistical algorithms and LLM-based methods to discover known and novel causal relations. In addition to causal discovery on initial variables, the missing variable proposal component of IRIS identifies and incorporates missing variables to expand the causal graphs. Our approach enables real-time causal discovery from only a set of initial variables without requiring pre-existing datasets.[1]

1 Introduction

A fundamental task in various disciplines of science, including biology, economics and healthcare, is to identify and utilize underlying causal relations (Kuhn, 1962). Although interventional experiments are ideal for discovering causal relations, they are often impractical due to ethical, financial, or logistical constraints. Therefore, researchers develop statistical methods to infer causal relations from purely observational tabular data (Pearl, 2009; Spirtes et al., 2000), though such data is often *not* available for a wide range of NLP applications.

Statistical and large language model (LLM)-based causal discovery algorithms face distinct challenges that limit their applicability in real-world scenarios. First, traditional statistical algorithms predominantly require high-quality structured tabular data, which is notoriously difficult to obtain. In contrast, LLM-based methods can consistently estimate causal relations explicitly present in their training data without relying on tabular data. However, these models encounter significant limitations when attempting to uncover causal relationships that were not previously documented (Feng et al., 2024b; Zečević et al., 2023). Second, statistical causal discovery algorithms require predefined sets of random variables as input, a constraint that significantly limits their flexibility. LLMs, however, demonstrate the capability to reliably extract and identify concepts and entities as variables directly from texts (Zhang et al., 2011; Glymour et al., 2019). Third, most statistical algorithms are theoretically grounded and mathematically verifiable, but operate under assumptions that rarely hold in real-world scenarios, such as the *causal sufficiency* assumption (*i.e.,* the absence of unobservable variables in the causal graph) and *acyclicity* assumption (*i.e.,* the absence of cycles in the causal graph) (Pearl, 2009; Neal, 2020). In contrast, the verification of LLMs' predictions in causal discovery remains an open challenge.

To address these limitations, we propose IRIS, **I**terative **R**etrieval and **I**ntegrated **S**ystem for verifiable causal discovery, in the absence of tabular data for statistical methods. To leverage the strengths of both statistical methods and LLMs, our framework takes a *hybrid* causal discovery approach, combining statistical methods with LLM-based causal relation extraction and verification techniques. This hybrid strategy allows us to leverage known causal relations and uncovering novel causal relations. IRIS begins with a set of initial random variables, which are sent as queries to retrieve a collection of relevant documents. Consequently, LLMs are applied to map the unstructured texts into structured tabular

*Corresponding author
[1]Our code and data are available at https://github.com/WilliamsToTo/iris

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9400–9428

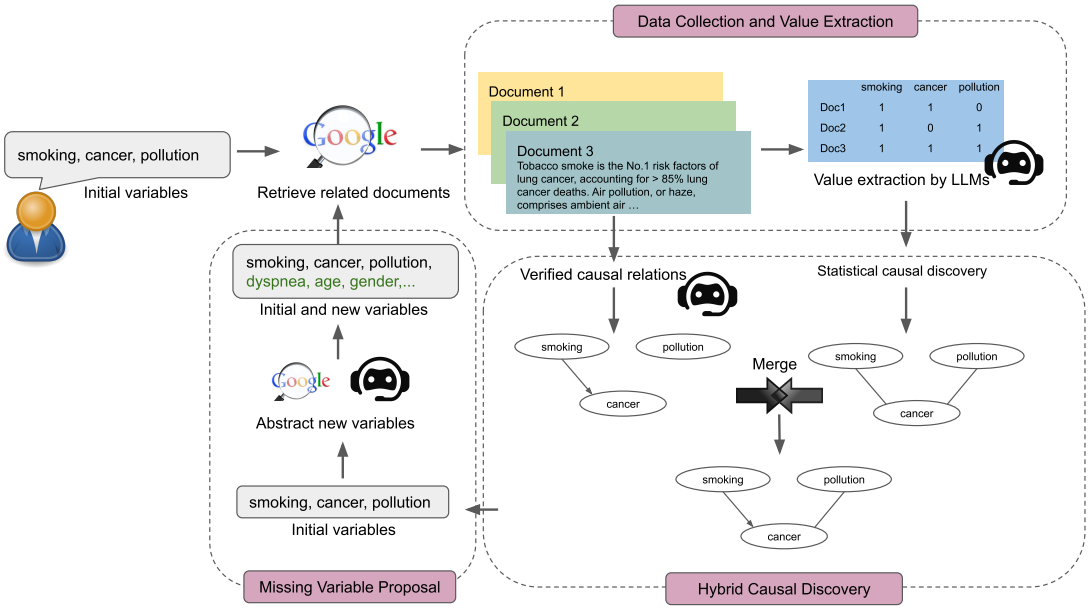


Figure 1: Illustration of IRIS. Given initial variables, we use the Google Search API and LLMs to *collect relevant documents and extract variable values*, then form structured data. For *hybrid causal discovery*, the statistical branch uses the structured data, while the causal relation extraction branch uses the retrieved documents. Their results are merged into the final causal graph. The *missing variable proposal* component identifies new variables, which are iteratively fed into our framework to expand the causal graphs.

data, which is utilized by an appropriate statistical method to perform causal discovery. Its results are further merged with the causal relations predicted and verified by LLMs. This hybrid approach allows cycles in causal graphs, thereby relaxing the *acyclicity* assumption. Additionally, we introduce a variable proposal component to identify new variables that have causal relations with the initial variables. This component allows us to relax the *causal sufficiency* assumption. We then iteratively use the expanded variables as input to our framework, further expanding the causal graphs.

Our experimental results demonstrate that IRIS significantly surpasses strong baselines across all datasets and scales effectively from small (4 initial variables) to large causal graphs (27 initial variables), as detailed in Section 4.1. Evaluations of individual components reveal that each component outperforms its corresponding baselines. Specifically, the evaluation of value extraction component shows that IRIS with GPT-4o exceeds the strong baselines, which also utilizes GPT-4o (Section 4.2). Our hybrid causal discovery method consistently outperforms both statistical algorithms and LLM-based approaches (Section 4.3). Lastly, our variable proposal component is more effective compared to prompt-based baselines (Section 4.4).

Primary contributions of IRIS are as follows: 1) We introduce an automatic sample collection and value extraction component that significantly reduces the manual labor for data collection in causal discovery tasks. 2) We propose a hybrid causal discovery method that leverages existing causal relations and uncovers novel causal relations. Our method permits cycles in causal graphs, thus relaxing the *acyclicity* assumption. 3) We develop a missing variable proposal component that identifies new variables that may have causal relations with the initial variables, relaxing the *causal sufficiency* assumption. 4) Experimental results demonstrate that IRIS consistently outperforms its baselines, with each component of IRIS also surpassing corresponding baselines.

2 Background

Causal discovery focuses on uncovering causal relations within a set of variables. Given a pair of variables (X, Y), the objective is to determine whether $X \leftarrow Y$, $Y \leftarrow X$, or no causal influence between them, where $\leftarrow$ denotes causal direction. A key distinction between causal discovery and relation extraction in NLP is that causal discovery can reveal unknown causal relations, whereas relation extraction focuses on transforming relations in free text into structured relational tuples.

Although randomized controlled trials and A/B

testing are the gold standard for causal discovery (Fisher, 1935), these experimental approaches are often impractical due to ethical or financial limitations. Thus, researchers turn to rely on statistical analysis of observational data to infer causal relations.

Statistical approaches to causal discovery can be broadly classified into: constraint-based methods, such as Peter and Clark (PC) (Spirtes et al., 2000) and inductive causation (IC) (Pearl, 2009); score-based methods (Heckerman et al., 1995; Chickering, 2002; Koivisto and Sood, 2004; Mooij et al., 2016); and functional methods (Shimizu et al., 2006; Hyvärinen et al., 2010). These methods employ statistical measures from observational data to construct causal graphs but have notable limitations. First, they require resource-intensive and extensive data collection. Second, theoretically, they cannot precisely identify ground-truth causal graphs but instead yield an equivalence class of true causal graphs (Spirtes et al., 2000; Pearl, 2009).

Furthermore, many statistical approaches, such as PC and Greedy Equivalence Search (GES), operate under assumptions. *Causal sufficiency* assumption posits that all variables are observed and included, neglecting the potential unobserved variables (Neal, 2020). Some algorithms, such as Tetrad condition-based (Silva et al., 2006; Kummerfeld and Ramsey, 2016) and high-order moments-based approaches (Adams et al., 2021; Chen et al., 2022) focus on only uncover specific types of unobserved variables, such as latent confounders (i.e., common causes). However, our work aims to identify more general unobserved variables, including confounders, mediators, causes, or effects of observed variables. *Acyclicity* assumption states that causal graphs contain no cycles, which allows causal discovery to align with Bayesian network and simplifies mathematical challenges. However, this assumption often contradicts real-world phenomena. Many causal graphs are known to contain feedback loops, such as the poverty cycle: poverty $\rightarrow$ limited access to education $\rightarrow$ low-paying jobs $\rightarrow$ poverty, (Banerjee and Duflo, 2012; De Weiss and Sirkin, 2010) and the predator-prey cycle: increase in predator population $\rightarrow$ decrease in prey population $\rightarrow$ decrease in predator population (Schmitz, 2017; Abrams, 2001). In contrast to prior work, our causal discovery framework allows for the inclusion of unobserved variables and permits cycles within causal graphs to align with real-world scenarios.

The advent of LLMs provides new opportunities to address causal discovery (Kıcıman et al., 2023; Zečević et al., 2023; Long et al., 2022). These approaches require LLMs to determine the causal relation between a given pair of variable names. However, the reliability of such methods is under scrutiny. Zečević et al. (2023) argue that LLMs may function as *"causal parrots"*, which depend on *memorization* to recall the causal relations present in their training data rather than infer causal relations. This raises concerns about LLMs' *generalization* to identify causal relations that are rare or absent in pre-training data. Feng et al. (2024b) presents empirical evidence that suggests while LLMs excel at reproducing frequent causal relations in pre-training data, they struggle to uncover novel causal relations.

In contrast to approaches that directly employ LLMs for causal discovery, Liu et al. (2024) utilize LLMs to extract variables and their values from collected documents, then apply statistical methods to uncover causal relations among these variables. Our work diverges from this approach by only taking a set of initial variables as input and employing an automated process to collect relevant documents. After variable value extraction, we implement a hybrid causal discovery approach, which integrate both statistical and LLM-based methods. Furthermore, our framework is capable of identifying new variables that exhibit causal relations with the initial set, thereby enabling an iterative process of data collection and causal discovery on an expanded variables set. This iterative method allows for a comprehensive exploration of the causal relations surrounding the initial variables.

3 Methodology

We introduce a real-time causal discovery framework, IRIS. Our method differs from prior causal discovery algorithms in three key aspects. First, IRIS does not rely on pre-existing observational data; instead, it automatically collects and extracts observational data related to the initial variables. Second, our hybrid causal discovery component can utilize known causal relations and uncover novel causal relations. Third, our approach relaxes the *acyclicity* and *causal sufficiency* assumptions.

3.1 Problem Definition

Given a set of initial variables, $\mathbb{Z} = \{z_1, z_2, ..., z_N\}$, where each z_i represents one

variable, the goal of real-time causal discovery is to automatically collect relevant unstructured data $\mathbb{D}$ and extract variable values to form structured data $\mathbb{X}$, which enables the discovery of causal relations through unstructured and structured data. After identifying causal relations among initial variables, the process involves identifying new variables causally related to the initial variables, resulting in an expanded set of variables $\mathbb{Z}_m$. The final output is an expanded causal graph $\mathcal{G} = (\mathbb{Z}_m, \mathbb{R})$, where $\mathbb{R} = \{\boldsymbol{r}_1, ..., \boldsymbol{r}_l\}$ represents the set of causal relations.

3.2 Data Collection and Value Extraction

The first step of IRIS comprises two main steps: collection of relevant documents and extraction of variable values. The detailed procedure is outlined in Algorithm 1 in Appendix A.3.

Retrieval of Relevant Documents We retrieve relevant documents using the Google API [2]. To maximize the relevance to initial variables, we create search queries using a stepwise removal approach: 1) Begin with queries containing all variable names (*e.g.,* "smoking" AND "cancer" AND "pollution"). 2) Progressively remove one variable (*e.g.,* "smoking" AND "cancer"). 3) Stop with single-variable queries (*e.g.,* "smoking"). We also use synonyms of variables to enhance coverage. We select the top-k retrieved documents for each query. To ensure relevance to most variables, k is higher for queries containing more variables. The retrieval process continues until the total number of collected documents reaches a predefined threshold. The resulting document set is denoted as $\mathbb{D} = \{d_1, .., d_T\}$, where d_i represents one document.

Extraction of Variable Values We use LLMs to extract variable values from collected documents $\mathbb{D}$. Given an LLM $\boldsymbol{M}$, we design a prompt l including a document d_i and a description of one variable z_j. The variable description includes its name and the meaning of its values. We guide the LLM to generate responses following multiple thinking steps, simulating human expert reasoning, and provide the final answer in a specific format (Lin et al., 2024). This generation process can be denoted as $o_{ij} = \boldsymbol{M}(l(d_i, \mathrm{z}_j))$, where o_{ij} is LLM's response regarding the value of variable z_j in document d_i. We then extract the value v_{ij} from response o_{ij}. By iterating through all variables and documents, we construct a structured data $\mathbb{X}$ where each column represents a variable and each row represents a document.

3.3 Hybrid Causal Discovery

We employ a hybrid causal discovery approach, leveraging both statistical methods and LLM-based relation extraction techniques. The detailed process of our hybrid causal discovery method is outlined in Algorithm 2 in Appendix A.3.

Statistical Causal Discovery For structured data $\mathbb{X}$, we employ statistical causal discovery algorithms including PC (Spirtes et al., 2000), GES (Chickering, 2003), and NOTEARS (Zheng et al., 2018). For instance, the PC algorithm performs conditional independence tests between variable pairs, progressively expanding the conditioning sets to determine the presence of causal relations. These algorithms process structured data $\mathbb{X}$ to produce a causal graph $\hat{\mathcal{G}}_s$ as the output.

LLM-based Causal Relation Extraction We introduce a novel causal relation extraction method inspired by causal relation verification (Si et al., 2024; Wadden et al., 2022). We treat each potential causal relation as a claim (*e.g.,* "smoking causes lung cancer") and find documents containing both the cause and effect terms (*e.g.,* "smoking" AND "lung cancer"). To ensure the trustworthiness of retrieved documents, we restrict the search domain to reputable academic repositories [3]. We then employ LLMs to assess whether each document supports or refutes or not relates with the causal relation using a carefully designed prompt. If a majority of documents support the causal relation, we incorporate it into a causal graph $\hat{\mathcal{G}}_v$. Otherwise, it is excluded.

Graph Merging The two branches of our hybrid method produce two causal graphs: $\hat{\mathcal{G}}_s$ from statistical methods and $\hat{\mathcal{G}}_v$ from the LLM-based approach. To merge them into the final causal graph $\hat{\mathcal{G}}$, we post-process the causal graph $\hat{\mathcal{G}}_s$ by adding high-confidence causal relations from $\hat{\mathcal{G}}_v$ and removing those strongly refuted by the verification process. This merging strategy is employed for two reasons: (1) the structured data $\mathbb{X}$ from the value extraction phase might contain noise; (2) causal relations that are widely supported or refuted by trustworthy documents can be treated as known knowledge.

[2] https://developers.google.com/custom-search/docs/overview

[3] Our search is limited to the following academic website domains: jstor.org, springer.com, ieee.org, ncbi.nlm.nih.gov, sciencedirect.com, scholar.google.com, arxiv.org.

3.4 Missing Variable Proposal

This step aims to identify missing variables not included in the initial set but potentially causally related to them, and append these to $\mathbb{Z}_m$, as outlined in Algorithm 3 in Appendix A.3.

Variable Abstraction We first use LLMs to abstract missing variables from the retrieved documents $\mathbb{D}$. For each document, LLMs are instructed to analyze the content of each document, identify variables that could influence or be influenced by the initial variables, and then provide the most possible variable in a specified format.

Variable Selection To select the most promising variables from all abstracted variables, we employ a dual approach combining causal relation verification and statistical measures. *Causal Relation Verification*: Using the method described in Section 3.3, we verify whether each new variable has a confirmed causal relation with any initial variable. Variables supported by the majority of documents are added to $\mathbb{Z}_m$. *Statistical Measure*: We compute the Pointwise Mutual Information (PMI) between each new variable and the initial variables to quantify their dependence, with higher PMI scores indicating stronger potential causal association. The PMI between two variables $(\mathrm{z}_i, \mathrm{z}_j)$ is defined as:

$$
\begin{aligned}
PMI(\mathrm{z}_i, \mathrm{z}_j) &= \log \frac{p(\mathrm{z}_i, \mathrm{z}_j)}{p(\mathrm{z}_i)p(\mathrm{z}_j)} = \log \frac{\frac{o(\mathrm{z}_i, \mathrm{z}_j)}{C}}{\frac{o(\mathrm{z}_i)}{C}\frac{o(\mathrm{z}_j)}{C}} \\
&= \log \frac{o(\mathrm{z}_i, \mathrm{z}_j)}{o(\mathrm{z}_i)o(\mathrm{z}_j)} + \cancel{\log C}
\end{aligned} \tag{1}
$$

where $o(\mathrm{z}_i, \mathrm{z}_j)$ is the count of documents where $(\mathrm{z}_i, \mathrm{z}_j)$ co-occur, $o(\mathrm{z}_i)$ is the count where z_i appears, and C is the total number of retrievable documents. Since C is constant, $\log C$ is ignored. These counts are obtained by the Google Search API. We compute the PMI score of each abstracted variable with the initial variables and append the top k variables with the highest aggregate PMI scores to $\mathbb{Z}_m$.

With the expanded variables $\mathbb{Z}_m$, we can iterate the data collection, value extraction, and causal discovery processes to generate an expanded causal graph $\mathcal{G} = (\mathbb{Z}_m, \mathbb{R})$ that incorporates these missing variables and new causal relations.

4 Experiments

4.1 Evaluation of the IRIS Framework

Datasets. The initial variables are from: Cancer (Korb and Nicholson, 2010), Respiratory Disease, Diabetes, Obesity (Long et al., 2022), Alzheimer's Disease Neuroimaging Initiative (ADNI) (Shen et al., 2020), and Insurance (Binder et al., 1997). For more details, see Appendix A.7.

Our Method and Baselines. We employ GPT-4o due to its superior performance across all components of IRIS (see Sections 4.2, 4.3, and 4.4). All prompts in IRIS are designed using the Chain-of-Thought (CoT) (Wei et al., 2022) strategy and incorporate retrieved documents. Detailed prompt engineering for IRIS and baselines is provided in Appendix A.4. For the statistical causal discovery algorithms, we utilize the Greedy Equivalence Search (GES) algorithm because it achieves the highest average F1 score and Normalized Hamming Distance (NHD) ratio across all datasets, as demonstrated in Section 4.3.

We consider the following baselines: *0-shot* relies solely on a zero-shot prompt. *CoT* enhances prompts through a step-by-step reasoning process, mimicking human thought patterns. *Retrieval-Augmented Generation (RAG)* (Lewis et al., 2020) incorporates retrieved documents into CoT prompt. Both baselines and human annotation determine causal relations among expanded variables from our missing variable proposal component.

Evaluation. We hire three domain experts to independently annotate ground-truth expanded causal graphs. Edges are included if at least two annotators agree. Inter-annotator agreement is high, with a Krippendorff's alpha of 0.88 (Krippendorff, 2011). The detailed annotation instruction is in Table 7 in Appendix A.5. The expanded causal graphs are illustrated in Figures 3 - 7 in Appendix A.6. Following Kıcıman et al. (2023); Feng et al. (2024b), we evaluate the results of causal discovery using precision, recall, F1 score, and the Ratio of Normalized Hamming Distance (NHD) to baseline NHD. The ratio is defined as ratio $= \frac{\text{NHD}}{\text{baseline NHD}}$, where the baseline NHD is derived from the worst-performing causal graph that has the same number of edges as the predicted graph. A lower ratio signifies a more accurate predicted causal graph. All results are averaged over three independent runs per causal graph.

Experimental Results and Analysis. Table 1 demonstrates that IRIS consistently outperforms all baselines across all datasets, achieving the highest F1 scores and lowest NHD ratios. A paired t-test (Ross and Willson, 2017) confirms that the performance differences between IRIS and the baselines (in both F1 and NHD ratio) are statistically signifi-

Method	P	R	F1↑	Predict Edge	NHD Ratio↓
Cancer					
0-shot	0.64	0.32	0.43	14	0.57
CoT	0.67	0.38	0.48	18	0.54
RAG	0.70	0.44	0.54	17	0.49
IRIS	0.89	0.57	**0.70**	18	**0.30**
Respiratory Disease					
0-shot	0.67	0.36	0.47	12	0.53
CoT	0.64	0.4	0.49	12	0.51
RAG	0.64	0.45	0.53	16	0.47
IRIS	0.67	0.55	**0.60**	18	**0.40**
Diabetes					
0-shot	0.70	0.46	0.56	17	0.45
CoT	0.66	0.48	0.55	16	0.46
RAG	0.73	0.47	0.57	16	0.43
IRIS	0.76	0.50	**0.60**	17	**0.39**
Obesity					
0-shot	0.57	0.33	0.42	14	0.58
CoT	0.59	0.38	0.46	25	0.54
RAG	0.62	0.45	0.52	19	0.49
IRIS	0.67	0.58	**0.62**	21	**0.38**
ADNI					
0-shot	0.47	0.29	0.36	17	0.64
CoT	0.46	0.31	0.37	21	0.62
RAG	0.50	0.34	0.40	19	0.60
IRIS	0.50	0.36	**0.42**	20	**0.58**
Insurance					
0-shot	0.35	0.38	0.36	69	0.65
CoT	0.41	0.38	0.39	65	0.61
RAG	0.44	0.40	0.42	67	0.57
IRIS	0.61	0.46	**0.53**	49	**0.47**

Table 1: Evaluation results of the complete framework.

cant (p-value ≤ 0.05). For both the baselines and IRIS, the variance across all metrics is below 0.05, likely due to the consistency of the retrieved documents and the stability of GPT-4o's responses. In terms of precision and recall, while some baselines (e.g., RAG in ADNI) achieve comparable precision to IRIS, none match its recall. This highlights IRIS's ability to uncover a greater number of true causal relations through its hybrid causal discovery approach. Among the datasets, ADNI exhibits the lowest overall performance for both methods, likely due to the inherent complexity of Alzheimer's disease causal relations. Meanwhile, the Insurance dataset, which contains the most complex causal graph (expanding from 27 initial variables to 35 variables and 67 edges), showcases the scalability of IRIS. Among the baselines, RAG performs better than others, underscoring the effectiveness of integrating retrieved documents with reasoning steps for causal discovery.

4.2 Evaluation of Value Extraction

Datasets. We evaluate the value extraction component of our method using two table-to-text datasets:

Method	P	R	F1
AppleGastronome			
COAT (GPT-4o)	0.74	0.76	0.75
IRIS (Llama)	0.71	0.72	0.71
IRIS (GPT-3.5)	0.75	0.77	0.76
IRIS (GPT-4o)	**0.79**	**0.82**	**0.79**
Neuropathic			
COAT (GPT-4o)	0.72	0.80	0.79
IRIS (Llama)	**0.76**	0.82	0.79
IRIS (GPT-3.5)	0.71	0.89	0.79
IRIS (GPT-4o)	0.73	**1.0**	**0.84**

Table 2: Evaluation results of value extraction. Llama represents Llama-3.1-8b-instruct.

AppleGastronome and Neuropathic (Liu et al., 2024). These datasets are particularly suitable for our task as they provide tabular data where columns represent variables and rows represent samples. Each row is associated with a corresponding textual description. The datasets are structured as follows: AppleGastronome contains 7 variables and 100 samples. Variable values are -1, 0, or 1. Neuropathic contains 7 variables and 100 samples. Variable values are 0 or 1.

Our Method and Baselines. We utilize state-of-the-art LLMs for our method: Llama-3.1-8b-Instruct (Meta, 2024), GPT-3.5-turbo (OpenAI, 2022), GPT-4o (OpenAI, 2024). Additionally, we compare our method with COAT, which also utilizes an LLM to extract values of variables from documents (Liu et al., 2024). To ensure a fair comparison, we use GPT-4o in both our method and the COAT implementation.

Metrics. Given that variable values are categorical, we frame the value extraction task as a classification problem, predicting the value of a variable in a given document. Therefore, we employ standard classification metrics: precision, recall, and F1.

Experimental Results and Analysis. Table 2 presents the evaluation results of the value extraction component on the AppleGastronome and Neuropathic datasets. Our method's superior performance with GPT-4o, compared to COAT using the same LLM, indicates that our approach is more effective than COAT under identical LLM. In both datasets, we observe a consistent trend of improvement from Llama-3.1-8b-Instruct to GPT-3.5, and further to GPT-4o when using our method. This progression aligns with the general understanding that more advanced LLMs tend to perform better on

complex tasks. Overall, the models perform better on the Neuropathic dataset compared to AppleGastronome. This could be attributed to the simpler binary values of the Neuropathic dataset (values 0 or 1) compared to the ternary values in AppleGastronome (-1, 0, 1). The additional complexity in AppleGastronome might introduce more opportunities for misclassification. The high performance of GPT-4o suggests that it could be highly effective for value extraction in our framework.

4.3 Evaluation of Causal Discovery

Datasets. We evaluate our hybrid causal discovery component on: Cancer (Korb and Nicholson, 2010), Respiratory Disease, Diabetes, Obesity (Long et al., 2022), Alzheimer's Disease Neuroimaging Initiative (ADNI) (Shen et al., 2020), and Insurance (Binder et al., 1997). These causal graphs are annotated by domain experts. The ground-truth causal graphs are presented in Figure 9 in Appendix A.7.

Our Method and Baselines. In our hybrid causal discovery, for statistical algorithms, we utilize PC (Spirtes et al., 2000), GES (Chickering, 2003), and NOTEARS (Zheng et al., 2018). Among the three statistical methods, we select the one that demonstrates the best performance for hybrid causal discovery. Based on the value extraction results (see Table 2), we use GPT-4o, which demonstrated the best performance, as the LLM for both our method and the baseline approaches. To illustrate how different LLMs affect the performance of our method, we employ the Llama-3.1-8b-instruct model as a counterpart. We compare our method against several baselines: 1) Pairwise-LLM (Feng et al., 2024b) calls LLMs for each pair of variables to determine causal relations. 2) BFS-LLM (Jiralerspong et al., 2024) employs a breadth-first search with LLMs to determine causal relations. 3) COAT extracts values using LLMs and discovers causal relations with the PC algorithm (Liu et al., 2024).

Metrics. We evaluate predicted causal graphs using precision, recall, F1, and NHD ratio as detailed in Section 4.1.

Experimental Results and Analysis. The evaluation results of the causal discovery component across datasets are presented in Figure 2. More detailed are presented in Table 10 - 15 in Appendix A.8. In these results, our hybrid method consistently outperforms baselines across all datasets. This highlights the effectiveness of combining statistical algorithms with LLM-based methods.

We observe that the performance of individual statistical algorithms (GES, NOTEARS, PC) varied across datasets. PC excels in Respiratory Disease and Obesity. GES demonstrates optimal performance on Diabetes and Obesity. NOTEARS performs best on Cancer and ADNI but struggles significantly with Diabetes and Obesity, achieving a 0 F1 score and a 1 NHD ratio. This variation highlights the importance of selecting statistical algorithms based on the characteristics of the observational data, which presents a compelling area for further research. From our experiments, both GES and PC exhibit strong performances; however, GES outperforms PC with a 0.09 higher average F1 score and a 0.09 lower average NHD ratio. Given these results, GES is recommended as the primary choice when the suitability of the algorithm is uncertain. When comparing the performance of Llama-3.1-8b-instruct and GPT-4o, GPT-4o consistently outperforms Llama-3.1-8b-instruct across all datasets, with a particularly significant gap observed in the ADNI dataset. We believe this discrepancy arises because ADNI involves specialized knowledge that is less commonly represented in the pre-training data of Llama-3.1-8b-instruct.

Pairwise-LLM and BFS-LLM show competitive performance on simpler datasets. They perform well on the Cancer and Respiratory Disease datasets. However, their performance degrades on more complex datasets like ADNI. This suggests that while LLMs have potential in causal discovery, they may struggle with more complex causal relations, possibly due to the lower occurrence of such domain-specific causal relations in their training data (Feng et al., 2024b). The COAT method yields results similar to IRIS- PC because both approaches extract values from documents and discover causal relations through the PC algorithm.

4.4 Evaluation of Missing Variable Proposal

Datasets. Evaluating the missing variable proposal component presents a unique challenge: the ground-truth missing variables are inherently unknown in real-world scenarios. To address this, we simulate missing variables and assess our method's ability to identify them. We start with complete, ground-truth causal graphs and remove variables to create incomplete graphs. We employ the initial causal graphs from Cancer, Respiratory Disease, Diabetes, Obesity, ADNI, and Insurance. For each causal graph, we iteratively remove one variable at a time, creating multiple test cases per graph. We

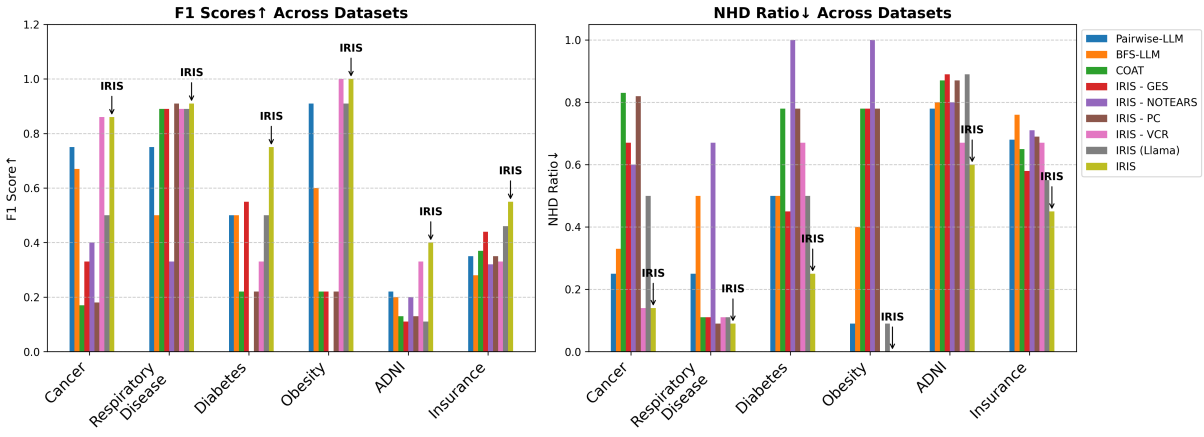


Figure 2: Evaluation results of causal discovery component on five datasets. A higher F1 score indicates better performance, while a lower NHD ratio reflects better performance. VCR refers to verified causal relations that are extracted from relevant academic documents and validated by LLMs. "Llama" refers to the use of the Llama-3.1-8b-instruct model as a substitute for GPT-4o in our method.

then apply our missing variable proposal component to these incomplete graphs, aiming to identify the removed variables.

Our Method and Baselines. For our method and the baseline, we use GPT-4o as the primary LLM. To assess the impact of different LLMs, we also replace GPT-4o with Llama-3.1-8b-instruct in our method. We compare with the following baselines: *0-shot*, which generates new variables using a zero-shot prompt; *CoT*, which enhances the prompt with reasoning steps; and *RAG*, which proposes new variables based on retrieved documents, similar to our method but relying solely on prompting to select the final variables. Prompts of baselines are provided in Appendix A.4.

Metrics. We evaluate the performance using a *success rate* metric, calculated as follows: 1) For each incomplete causal graph, we check if our method successfully proposes the removed variable in its final set of proposed variables $\mathbb{Z}_m$. 2) We count a "success" for each correctly proposed variable. 3) The success rate is computed as: Success Rate = Number of Successes / Total Number of Incomplete Graphs. For instance, in a causal graph with five variables, we create five different incomplete graphs by removing each variable. If our method correctly proposes the removed variable in three of these five graphs, the success rate would be 0.6 (=3/5). For the statistical approach, we select the top-5 variables based on their PMI scores.

Experimental Results and Analysis. The evaluation results of our Missing Variable Proposal (MVP) component are presented in Table 3. The

Method	Cancer	Resp. Disease	Diabetes	Obesity	ADNI	Insurance
0-shot	0.40	0.25	0.50	0.25	0.25	0.22
CoT	0.40	0.50	0.50	0.75	0.25	0.30
RAG	0.60	0.75	0.75	0.75	0.38	0.41
MVP	**0.80**	**0.75**	**1.00**	**1.00**	**0.50**	**0.59**
- VCR	0.60	0.75	0.50	0.75	0.25	0.48
- Stats	0.60	0.75	0.75	1.00	0.38	0.52
↔ Llama	0.40	0.50	0.25	0.50	0.13	0.45

Table 3: Success rate of the missing variable proposal (MVP) component. -VCR omits verified causal relation extraction; -Stats omits statistical approaches; ↔ Llama uses Llama-3.1-8b-instruct instead of GPT-4o.

MVP method consistently outperforms other baselines and ablation variants across all datasets. This demonstrates the effectiveness of combining verified causal relation extraction (VCR) with statistical approach (Stats) in identifying missing variables. Ablation studies indicate that both VCR and statistical approaches play a crucial role in enhancing the success rate of the MVP. The performance gap between MVP and MVP ↔ Llama indicates the superior capability of GPT-4o in understanding and reasoning about causal relations. All baselines consistently underperform compared to our MVP, indicating that relying solely on the textual knowledge from documents and LLMs is not enough for proposing missing variables.

5 Conclusion

In this paper, we introduce IRIS, a novel framework that addresses several longstanding challenges in causal discovery. By integrating automated data collection, hybrid causal discovery methods, and a variable proposal components, IRIS significantly

advances our ability to uncover causal relations in real-world scenarios. Our approach not only reduces the reliance on extensive manual data collection but also leverages existing knowledge in order to facilitate the discovery of novel causal relations with novel variables. Our experimental results show that IRIS consistently outperforms competitive baselines. Future work could aim to enhancing the scalability of IRIS for larger and more complex causal graphs by integrating causal relations extracted from texts with the ones identified through statistical algorithms.

Limitations

Our approach to uncovering causal relations using retrieved documents and LLMs has certain limitations. A primary challenge is the potential bias inherent in both the data and the LLMs. Retrieved documents may contain sampling biases, inaccuracies, or incomplete coverage of causal relations. Likewise, LLMs may inherit biases from their pretraining data or face limitations in generalization, potentially affecting their interpretation of causal relationships. To mitigate these issues, we retrieve documents from reliable academic websites, and leverage state-of-the-art LLMs like GPT-4o.

Another limitation is that the number of queries to the LLM grows quadratically with the number of variables. On average, our method takes approximately 15 hours to run, which is about three times slower than the zero-shot baseline. However, all LLM-related processes can be parallelized. For instance, in causal relation extraction, each causal relation can be independently queried in parallel to determine whether the relation holds. This parallel processing significantly mitigates the computational overhead and ensures that the framework remains scalable even as the number of variables increases.

Finally, the energy consumption of LLM inference presents an environmental challenge. While optimizing efficiency in LLM inference is an important research direction, it is beyond the scope of this work.

Ethics Statement

We acknowledge the importance of ACL Code of Ethics and agree with it. We ensure that our study is compatible with the provided code.

Our work involves uncovering causal relations using retrieved documents and LLMs, and we acknowledge the ethical considerations associated with this approach. The potential biases inherent in both the retrieved data and the LLMs pose a significant challenge. To mitigate these risks, we prioritize retrieving data from credible sources, such as academic publications and verified websites, to ensure the reliability of the input data. Additionally, we employ state-of-the-art LLMs, like GPT-4, which are designed to provide high-quality and robust outputs. However, we recognize that no system is entirely free from bias, and users of this framework should exercise caution in interpreting its results.

The evaluation of our method involves hiring human experts to annotate causal graphs. We have ensured that the annotation process adheres to ethical guidelines, including providing fair compensation for their contributions. Rigorous measures have been taken to thoroughly anonymize the causal graphs, which do not contain any personally identifiable information or sensitive data related to the contributors. The causal graphs were compiled with contributions from PhD students, which may inherently introduce biases influenced by their demographic backgrounds. We advise researchers utilizing this dataset to carefully account for these potential biases, particularly in studies related to AI fairness, bias, and safety.

References

Peter A Abrams. 2001. Predator-Prey Interactions. In *Evolutionary Ecology: Concepts and Case Studies* . Oxford University Press.

Jeffrey Adams, Niels Richard Hansen, and Kun Zhang. 2021. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. In *Advances in Neural Information Processing Systems*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Ananth Balashankar, Sunandan Chakraborty, Samuel Fraiberger, and Lakshminarayanan Subramanian. 2019. Identifying predictive causal factors from news streams. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2338–2348, Hong Kong, China. Association for Computational Linguistics.

A.V. Banerjee and E. Duflo. 2012. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. PublicAffairs.

Adrián Bazaga, Pietro Lio, and Gos Micklem. 2024. Unsupervised pretraining for fact verification by language model distillation. In *The Twelfth International Conference on Learning Representations*.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Computing Surveys (CSUR)*, 55(1):1–35.

John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. 1997. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244.

Quoc-Chinh Bui, Breanndán Ó Nualláin, Charles A Boucher, and Peter MA Sloot. 2010. Extracting causal relations on hiv drug resistance from literature. *BMC bioinformatics*, 11:1–11.

Du-Seong Chang and Key-Sun Choi. 2006. Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. *Information Processing and Management*, 42(3):662–678.

Zhengming Chen, Feng Xie, Jie Qiao, Zhifeng Hao, Kun Zhang, and Ruichu Cai. 2022. Identification of linear latent variable model with arbitrary distribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6350–6357.

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

David Maxwell Chickering. 2003. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 306–316, Melbourne, Australia. Association for Computational Linguistics.

Susan Pick De Weiss and Jenna Sirkin. 2010. *Breaking the poverty cycle: The human basis for sustainable development*. Oxford University Press.

Tao Feng, Lizhen Qu, and Gholamreza Haffari. 2023. Less is more: Mitigate spurious correlations for open-domain dialogue response generation models by causal discovery. *Transactions of the Association for Computational Linguistics*, 11:511–530.

Tao Feng, Lizhen Qu, Xiaoxi Kang, and Gholamreza Haffari. 2025. CausalScore: An automatic reference-free metric for assessing response relevance in open-domain dialogue systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2351–2369, Abu Dhabi, UAE. Association for Computational Linguistics.

Tao Feng, Lizhen Qu, Zhuang Li, Haolan Zhan, Yuncheng Hua, and Reza Haf. 2024a. IMO: Greedy layer-wise sparse representation learning for out-of-distribution text classification with pre-trained models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2625–2639, Bangkok, Thailand. Association for Computational Linguistics.

Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, and Gholamreza Haffari. 2024b. From pre-training corpora to large language models: What factors influence llm performance in causal discovery tasks? *Preprint*, arXiv:2407.19638.

R. A. Fisher. 1935. *The Design of Experiments*. Oliver and Boyd.

Daniela Garcia. 1997. Coatis, an nlp system to locate expressions of actions connected by causality links. In *Proceedings of the 10th European Workshop on Knowledge Acquisition, Modeling and Management*, EKAW '97, page 347–352, Berlin, Heidelberg. Springer-Verlag.

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

David Heckerman, Dan Geiger, and David M Chickering. 1995. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197–243.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3023–3030, New York, NY, USA. Association for Computing Machinery.

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. 2004. A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

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

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

Christopher S. G. Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 336–343, Hong Kong. Association for Computational Linguistics.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *Preprint*, arXiv:2402.07401.

Mikko Koivisto and Kismat Sood. 2004. Exact bayesian structure discovery in bayesian networks. *The Journal of Machine Learning Research*, 5:549–573.

Kevin B. Korb and Ann E. Nicholson. 2010. *Bayesian Artificial Intelligence, Second Edition*, 2nd edition. CRC Press, Inc., USA.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Thomas S. Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Erich Kummerfeld and Joseph Ramsey. 2016. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1655–1664, New York, NY, USA. Association for Computing Machinery.

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

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.

Yufeng Li, Rrubaa Panchendrarajan, and Arkaitz Zubiaga. 2024. Factfinders at checkthat! 2024: Refining check-worthy statement detection with llms through data pruning. *Preprint*, arXiv:2406.18297.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. 2024. Discovery of the hidden world with large language models. *Preprint*, arXiv:2402.03941.

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*.

Meta. 2024. Meet llama 3.1.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *J. Mach. Learn. Res.*, 17(1):1103–1204.

Brady Neal. 2020. *Introduction to Causal Inference from a Machine Learning Perspective*.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2024. Hello gpt-4o.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Amanda Ross and Victor L. Willson. 2017. *Paired Samples T-Test*, pages 17–19. SensePublishers, Rotterdam.

Oswald Schmitz. 2017. Predator and prey functional traits: understanding the adaptive machinery driving predator–prey interactions. *F1000Research*, 6.

Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. 2020. Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology. *Scientific reports*, 10(1):2975.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030.

Jiasheng Si, Yibo Zhao, Yingjie Zhu, Haiyang Zhu, Wenpeng Lu, and Deyu Zhou. 2024. CHECKWHY: Causal fact verification via argument structure. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15636–15659, Bangkok, Thailand. Association for Computational Linguistics.

Ricardo Silva, Richard Scheine, Clark Glymour, and Peter Spirtes. 2006. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(8):191–246.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, Prediction, and Search*. MIT press.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186.

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, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2011. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA. AUAI Press.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Comput. Surv.*, 56(11).

Youwen Zhao, Xiangbo Yuan, Ye Yuan, Shaoxiong Deng, and Jun Quan. 2023. Relation extraction: advancements through deep learning and entity-related features. *Social Network Analysis and Mining*, 13(1):92.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. Dags with no tears: continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9492–9503, Red Hook, NY, USA. Curran Associates Inc.

A Appendix

A.1 Related Work

Causal Discovery Causal discovery aims to uncover causal structures among variables, distinguishing itself from relation extraction in NLP by revealing novel causal relations rather than merely extracting known relations. While experimental approaches such as randomized controlled trials are gold standard methods(Fisher, 1935), practical limitations often necessitate statistical methods using observational data. These include constraint-based and score-based approaches (Spirtes et al., 2000; Pearl, 2009; Heckerman et al., 1995). However, statistical methods face challenges in data collection and theoretical limitations. Recent advancements in LLMs have introduced new possibilities for causal discovery without direct data access (Kıcıman et al., 2023; Zečević et al., 2023; Long et al., 2022). However, concerns about LLMs functioning as "causal parrots" and their ability to generalize to novel relations have been raised (Zečević et al., 2023; Feng et al., 2024b). Alternative approaches, such as using LLMs for variable proposer and combining them with statistical methods (Feng et al., 2023, 2024a; Liu et al., 2024), have emerged. Our work builds upon these ideas, introducing an automated document collection process, a hybrid causal discovery method integrating statistical and relation extraction techniques, and a hybrid approach for new variable proposal.

Relation Extraction Relation extraction aims to transform unstructured textual relations into structured relation tuples of the form $< e_1, r, e_2 >$, where e_1 and e_2 represent entities and r denotes the relation between them (Yang et al., 2022; Dasgupta et al., 2018). While relation extraction can identify cause-effect relationships from documents, it fundamentally differs from causal discovery in that it relies on explicitly stated relations in texts, whereas causal discovery can uncover novel causal relationships from observational data even in the absence of explicit textual mentions. Nevertheless, relation extraction can serve as a complementary method for identifying commonly known causal relations in textual data. Several studies have focused on extracting causal relations from natural language texts (Balashankar et al., 2019; Bui et al., 2010; Chang and Choi, 2006; Feng et al., 2025). The methods for causality extraction can be divided into pattern-based and deep learning-based approaches.

Pattern-based methods utilize predefined linguistic patterns to extract relevant text segments, which are then converted into tuples using hand-crafted algorithms (Garcia, 1997; Khoo et al., 2000). However, these methods often suffer from limited coverage of causal relations and require significant effort in pattern design. Deep learning-based methods employ neural networks to learn high-level abstract features and representations from sentences, framing relation extraction as a sequence-to-sequence task (Zhao et al., 2023, 2024). While these approaches offer improved performance, they typically require large fine-tuning datasets and may not consistently produce structurally correct output tuples.

A notable limitation of many relation extraction systems is the lack of verification for extracted relations, potentially leading to the extraction of false or unreliable relations from untrustworthy sources (Si et al., 2024; Wadhwa et al., 2023). Our work addresses this issue by adopting a novel approach: instead of directly extracting causal relations from documents, we pre-create textual mentions of causal relations (e.g., "smoking causes lung cancer") and employ LLMs to verify the veracity of these relations based on relevant documents. We consider a causal relation to hold if the majority of documents support its veracity, thereby enhancing the reliability of our extracted causal relations.

Claim Verification Claim verification aims to assess the veracity of claims based on relevant documents (Bekoulis et al., 2021). This process typically encompasses several key components: claim detection, document retrieval, veracity prediction, and explanation generation. Research in this field often focuses on specific aspects of the verification pipeline. For instance, Panchendrarajan and Zubiaga (2024) and Li et al. (2024) concentrate on identifying check-worthy statements from large text corpora. Others, such as Wadden et al. (2022) and Mohr et al. (2022), prioritize veracity prediction, while Wang and Shu (2023) emphasize the importance of generating explanations for verification outcomes. The emergence of LLMs has significantly influenced the field, with numerous studies leveraging LLMs for claim verification through carefully crafted prompts (Kim et al., 2024; Bazaga et al., 2024; Asai et al., 2024). Building on these advancements, one branch of our hybrid causal discovery approach reframes causal discovery as a causal relation verification task. We employ LLMs to assess the veracity of causal relations based on retrieved documents, subsequently incorporating verified relations into a causal graph. This methodology bridges the gap between traditional claim verification techniques and causal discovery, offering a novel approach to uncovering and validating causal relations.

A.2 Reproducibility Statement

We release our code and scripts at `https://github.com/WilliamsToTo/iris`. Detailed descriptions of the algorithms used in each component of our framework can be found in Appendix A.3. We provide all prompts utilized throughout our framework in Appendix A.4. The ground-truth causal graphs employed in our evaluation experiments are outlined in Appendix A.7. Additionally, Appendix A.5 presents human annotation instruction and interface for the human annotation tasks involved in evaluating the expanded causal graphs. The annotated expanded causal graphs, alongside the predicted causal graphs, are documented in Appendix A.6.

A.3 Algorithms

In this section, we provide detailed descriptions of the algorithms for each component of our method. The data collection and value extraction process is outlined in Algorithm 1. The hybrid causal discovery algorithm can be found in Algorithm 2. Finally, the algorithm for proposing missing variables is detailed in Algorithm 3.

A.4 Prompt Engineering Details

Prompts were designed using different strategies and ultimately adopted the chain-of-thought (CoT) (Wei et al., 2022) prompting approach, as shown in Table 4. These prompts contain retrieved document, task descriptions and stepwise instructions to complete tasks. Then we require LLMs to output final answer with specific format to easily extract answers. We also tried zero-shot prompts, but it demonstrated poor performance for value extraction, causal relation extraction, and missing variable abstraction. Few-shot prompts often exceeded the maximum input length for LLMs, as they required incorporating multiple long documents into the prompt. In contrast, CoT prompting provided better performance in all components.

To demonstrate that IRIS is LLM-agnostic, we use the same prompt across all LLMs during the evaluation. In our study, we use a separate validation set (not overlapping with test data) to compare

prompts. This validation set is built using high-confident causal relations from CauseNet (Heindorf et al., 2020) with randomly paired non-causal relations. For prompt selection, we first manually write a pool of manually written prompts from different researchers, then use GPT-4 to refine these prompts. We evaluate all human and LLM-refined prompts on the validation set and select the prompt that has the best performance (highest F1).

For the evaluation of the whole framework, the prompt used in the 0-shot, CoT, and RAG baselines is shown in Table 5. For the evaluation of the missing variable proposal, the prompt used the 0-shot, CoT, and RAG baselines is shown in Table 6.

A.5 Causal Relation Annotation Task for Expanded Variables

The detailed instructions for the causal relation annotation task for expanded variables are presented in Table 7. This table provides comprehensive guidance to annotators on how to identify and annotate causal relations among the given variables.

A.6 Human-annotated Causal Graphs for Expanded Variables

The human-annotated causal graphs for expanded variables are demonstrated in Figure 3, 4, 5, 6, 7, 8. The statistics of human-annotated causal graphs is presented in Table 8.

A.7 Ground-Truth Causal Graphs of Initial Variables

The ground-truth causal graphs of initial variables for evaluating the causal discovery component can be found in Figure 9. Table 9 demonstrates the statistics of initial ground-truth causal graphs with initial variables.

A.8 Evaluation of Causal Discovery Component

The detailed evaluation results of the causal discovery component are presented in Table 10, 11, 12, 13, 14, and 15.

Algorithm 1 Document Collection and Value Extraction

Require: Initial Variables $\mathbb{Z}$, LLM $\boldsymbol{M}$, threshold T, prompt l

Document Collection

$\mathbb{D} \leftarrow \emptyset$ ▷ Initialize an empty set for collected documents

while $|\mathbb{D}| < T$ **do**

$queries \leftarrow [(\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_n), (\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_{n-1}), \ldots, (\mathrm{z}_i)]$ ▷ queries considering all variables and their synonyms

for each q in $queries$ **do**

$n \leftarrow 20 \times \text{len}(q)$ ▷ Determine the number of URLs to collect

$urls \leftarrow \text{google\_search}(q, n)$ ▷ Search with query q and retrieve top-n URLs

for each url in $urls$ **do**

$D \leftarrow$ extract text from url

$\mathbb{D} \leftarrow \mathbb{D} \cup \{D\}$ ▷ Add extracted text to the document set

end for

end for

end while

Value Extraction

$\boldsymbol{V} \leftarrow$ Matrix of dimensions $T \times N$ ▷ Initialize a matrix with T rows and N columns

for each d_i in $\mathbb{D}$ **do**

for each z_j in $\mathbb{Z}$ **do**

$o_{ij} \leftarrow \boldsymbol{M}(l(d_i, \mathrm{z}_j))$ ▷ Determine value of z_j in d_i by LLM

$v_{ij} \leftarrow \text{extract}(o_{ij})$ ▷ Extract value from LLM output

$\boldsymbol{V}[i][j] \leftarrow v_{ij}$ ▷ Store the value v_{ij} in matrix $\boldsymbol{V}$ at position (i, j)

end for

end for

Output: $\mathbb{D}, \boldsymbol{V}$

Value Extraction

Given a document: {doc}

Please complete the below task.
We have a variable named '{var}'. The value of variable '{var}' is True or False.
True indicates that the existence of '{var}' can be inferred from the document, whereas False suggests that the existence of '{var}' cannot be inferred from this document.
Based on the document provided, what is the most appropriate value for '{var}' that can be inferred?
Please form the answer using the following format.
First, provide an introductory sentence that explains what information will be discussed.
Next, list generated answer in detail, ensuring clarity and precision.
Finally, conclude the final answer of the inferred value for '{var}' using the following template:
The value of '{var}' is ____.

Causal Relation Verification

Given a document: {doc}

Please complete the below task.
We have a claim: '{claim}'. We need to check the veracity of this claim. The value of veracity is True or False or Unknown.
True indicates that the given document supports this claim,
False indicates that the given document refutes the claim.
Unknown indicates that the given document has no relation to the claim.
Please form the answer with a logical reasoning chain according to the following format.
First, provide an introductory sentence that explains what information will be discussed.
Next, list the logical reasoning chain in detail, ensuring clarity and precision.
Finally, conclude the veracity of claim '{claim}' using the following template:
The veracity of claim '{claim}' is ___.

Missing Variable Abstraction

Given a document: {doc}

Please complete the below task.
We have some given variables: '{initial_variables}'.
What are the high-level variables in the provided document that have causal relations to variables in the given variable set?
Please form the answer using the following format.
First, propose as many variables as possible that have causal relationships with the given variables, based on your understanding of the document.
Please ensure these proposed variables are different from the ones already provided.
Next, refine your list of candidate variables by reducing semantic overlap among them and shortening their names for clarity.
Finally, determine the most reliable variable candidate as the final answer using the template provided below:
The new abstracted variable is <var>____</var>.

Table 4: The prompts used in IRIS, where doc indicates the content of a document, claim refers to a causal relation (*e.g.,* smoking causes lung cancer).

0-shot

The task is to determine the cause-effect relation between two variables.
The variables are: variable1 and variable2.
Your answer should be one of the following:
variable1 $\rightarrow$ variable2 (if variable1 causes variable2)
variable1 $\leftarrow$ variable2 (if variable2 causes variable1)
No causal relation (if there is no clear cause-effect relationship)

Let's provide a step-by-step process to analyze the relation between them,
then provide your final answer using the following format:
The final answer is: variable1 $\rightarrow$ variable2 or variable1 $\leftarrow$ variable2 or No causal relation

CoT

The task is to determine the cause-effect relation between two variables.
The variables are: variable1 and variable2.
Your answer should be one of the following:
variable1 $\rightarrow$ variable2 (if variable1 causes variable2)
variable1 $\leftarrow$ variable2 (if variable2 causes variable1)
No causal relation (if there is no clear cause-effect relationship)

Let's analyze the relation through the following steps:
First, briefly describe each variable and its typical behavior.
Second, does one variable naturally precede the other in time or logic?
Third, are there common confounders or external factors that could explain the relationship?
Finally, provide your final answer in the following format:
The final answer is: variable1 $\rightarrow$ variable2 or variable1 $\leftarrow$ variable2 or No causal relation

RAG

Analyze the relevant information from the retrieved document:
doc
The task is to determine the cause-effect relation between two variables.
The variables are: variable1 and variable2.
Your answer should be one of the following:
variable1 $\rightarrow$ variable2 (if variable1 causes variable2)
variable1 $\leftarrow$ variable2 (if variable2 causes variable1)
No causal relation (if there is no clear cause-effect relationship)

Let's analyze the relation through the following steps:
First, briefly describe each variable and its typical behavior.
Second, does one variable naturally precede the other in time or logic?
Third, does the retrieved document provide any information that explains the relationship?
Finally, provide your final answer in the following format:
The final answer is: variable1 $\rightarrow$ variable2 or variable1 $\leftarrow$ variable2 or No causal relation

Table 5: The prompt used in the baselines for evaluation of expanded causal graphs.

0-shot
The task is to identify new variables that are causally related to the given variables. Given variables: variables Follow a step-by-step approach to analyze the given variables and determine relevant causal relationships. Then, present your final answer in the following format: Proposed variables: [variable1, variable2, ...]
CoT
The task is to identify new variables that are causally related to the given variables. Given variables: variables Let's break this down step by step to systematically analyze the given variables and determine relevant causal relationships: First, understand the given variables. Second, propose potential direct causes and effects associated with the given variables. Third, verify that the proposed variables align with real-world causal structures. Finally, present your final answer in the following format: Proposed variables: [variable1, variable2, ...]
RAG
Given a document: doc The task is to identify new variables that are causally related to the given variables. Given variables: variables Let's break this down step by step to systematically analyze the given variables and determine relevant causal relationships: First, understand the given variables. Second, propose potential direct causes and effects associated with the given variables based on the information in the given document. Third, verify that the proposed variables align with real-world causal structures. Finally, present the most reliable variable candidates as the final proposed variables in the following format: Proposed variables: [variable1, variable2, ...]

Table 6: The prompt used in the baselines to evaluate the missing variable proposal.

Causal Relation Annotation Task

Task overview:
Your task is to identify and annotate causal relations among a set of variables. A causal relation exists when one variable directly influences another.

Instructions:
1. Consider each pair of variables and determine if there is a direct causal relationship between them.
2. If you believe variable A causes variable B, indicate this as: A $\rightarrow$ B
3. Be cautious of confusing correlation with causation. Only mark a relationship if you believe there is a direct causal link.
4. Consider the direction of causality carefully. For example, "Obesity $\rightarrow$ Heart Failure" suggests obesity causes heart failure, not the other way around.
5. It's okay to have multiple causes for a single effect, or multiple effects from a single cause.
6. Not all variables will necessarily have causal relationships with others.
7. Use your best judgment based on available knowledge and logical reasoning.

Examples:
lifestyle -> obesity
heart defect -> cardiac output
genetic disorder -> heart defect

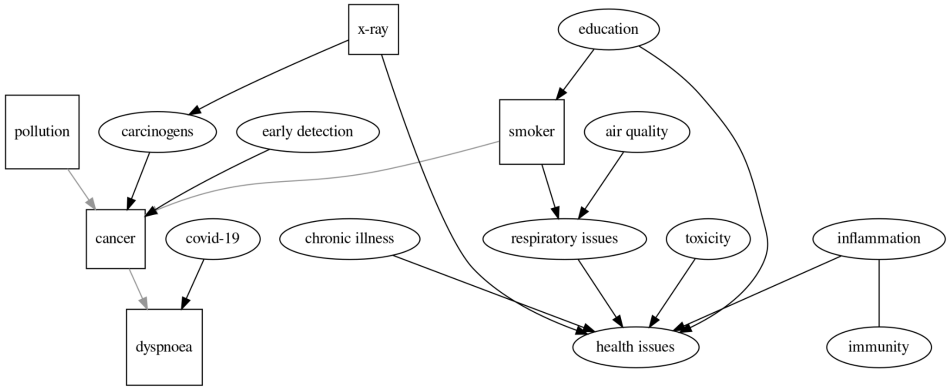
Submission:
Please submit your annotations as a list of causal relations in the format: Variable A -> Variable B
Thank you for your careful consideration of this task!

Task 1: Cancer

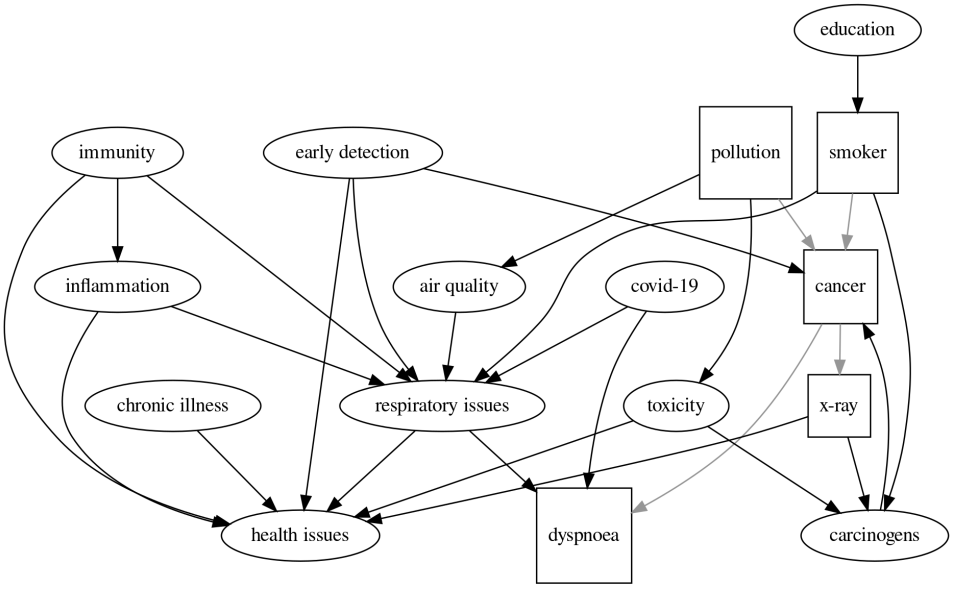
Variables:
pollution
smoker
cancer
x-ray
dyspnoea
air quality
education
health issues
toxicity
chronic illness
covid-19
inflammation
respiratory issues
immunity
carcinogens
early detection

Causal Relations:
...

Table 7: Instructions and interface of causal relation annotation task.

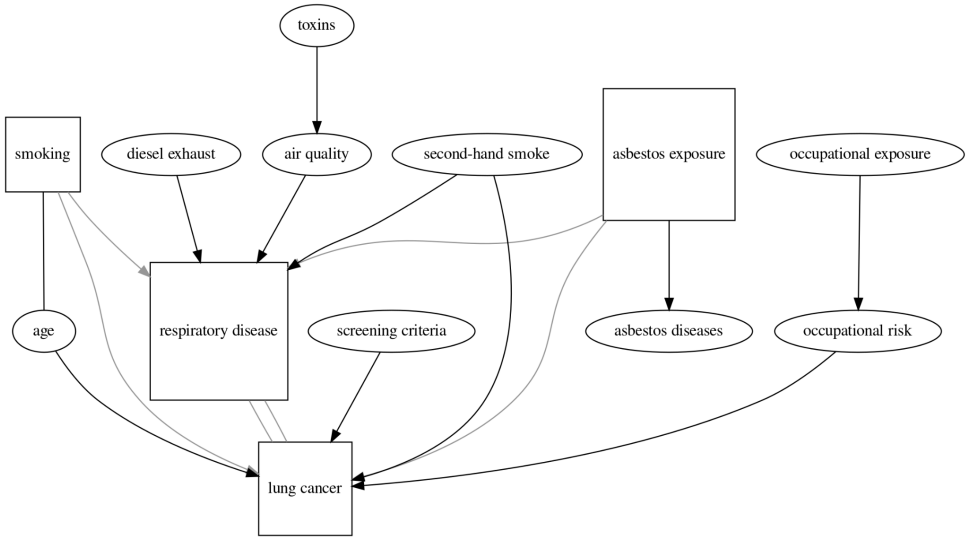


(a) IRIS

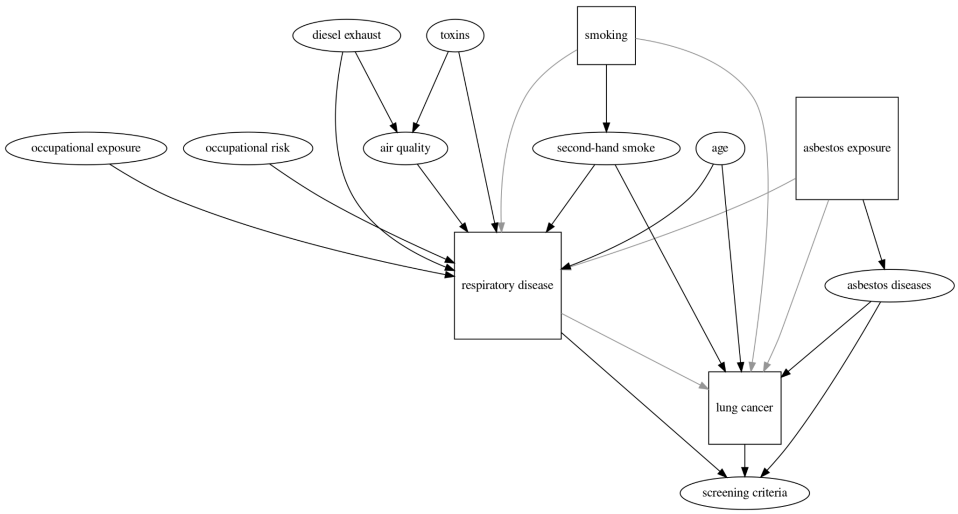


(b) Human

Figure 3: Illustration of expanded causal graphs for Cancer. Squared nodes represent initial variables, while round nodes denote new proposed variables.

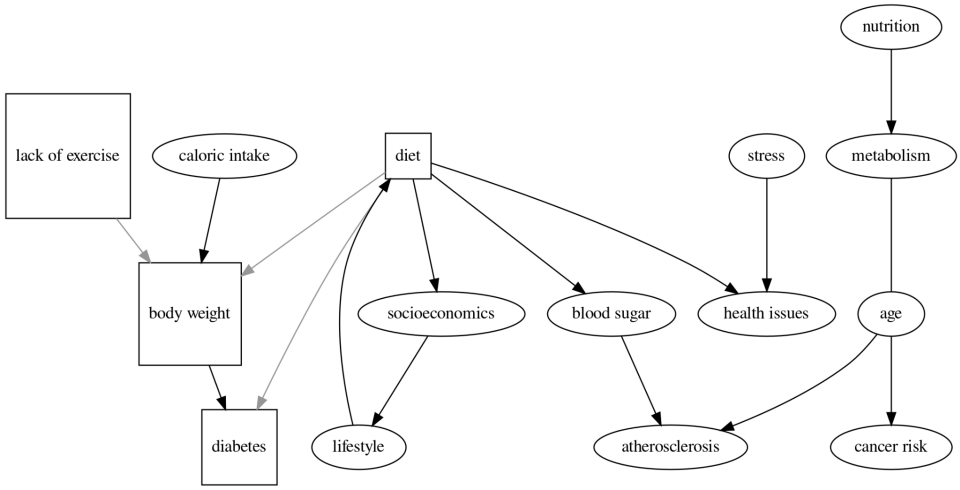


(a) IRIS

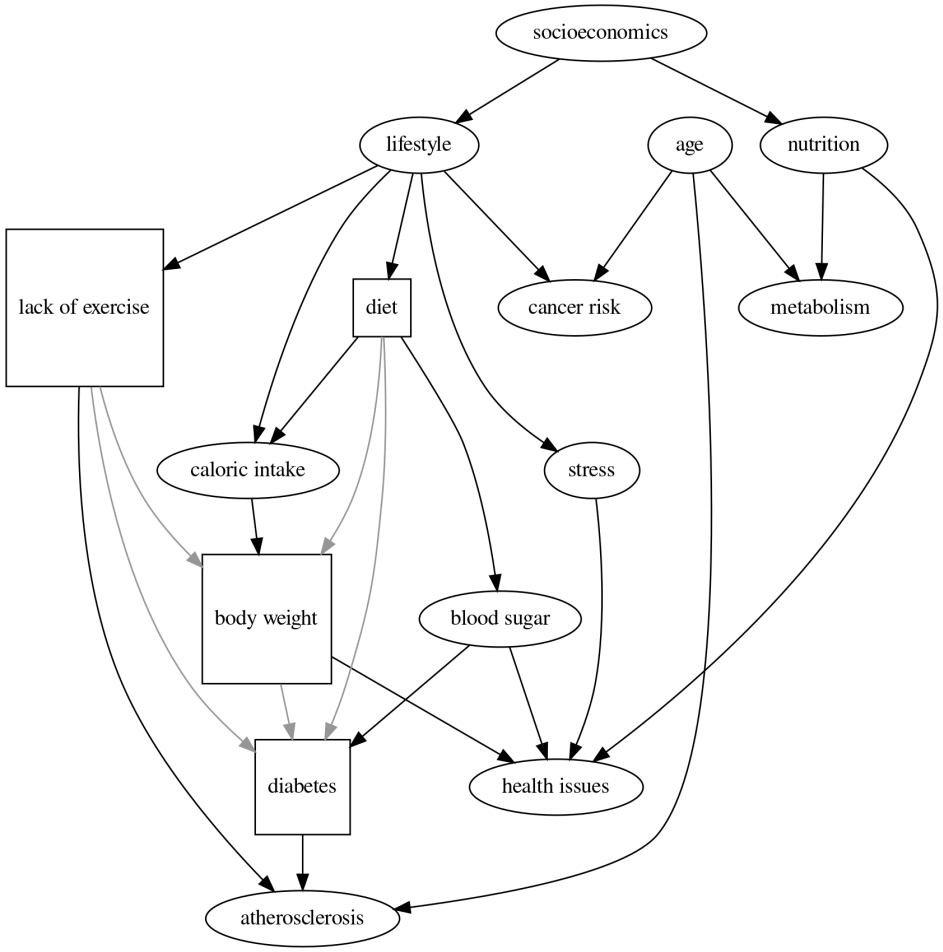


(b) Human

Figure 4: Illustration of expanded causal graphs for Respiratory Disease. Squared nodes represent initial variables, while round nodes denote new proposed variables.

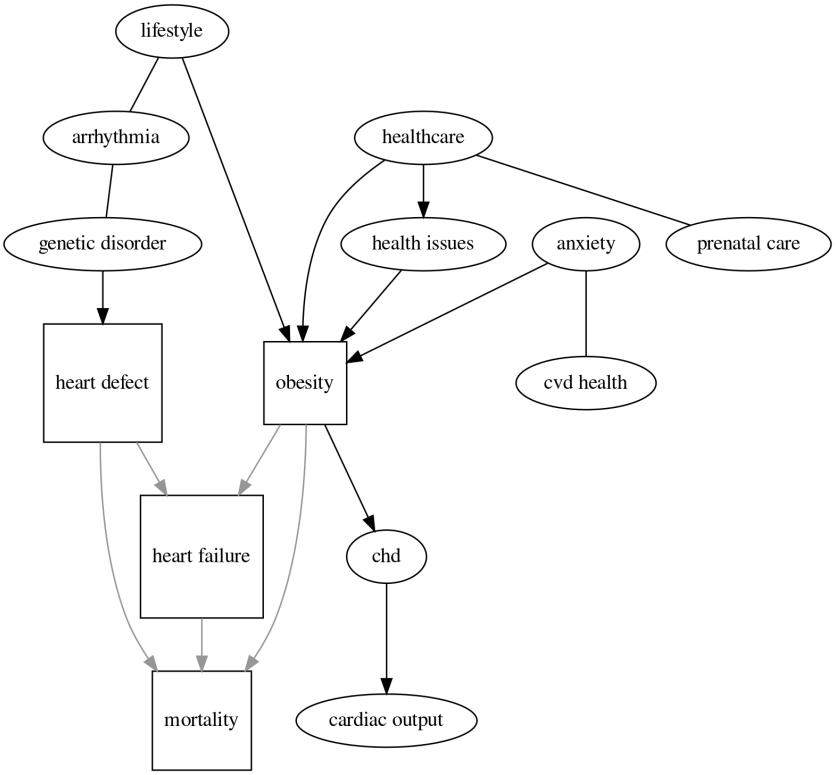


(a) IRIS

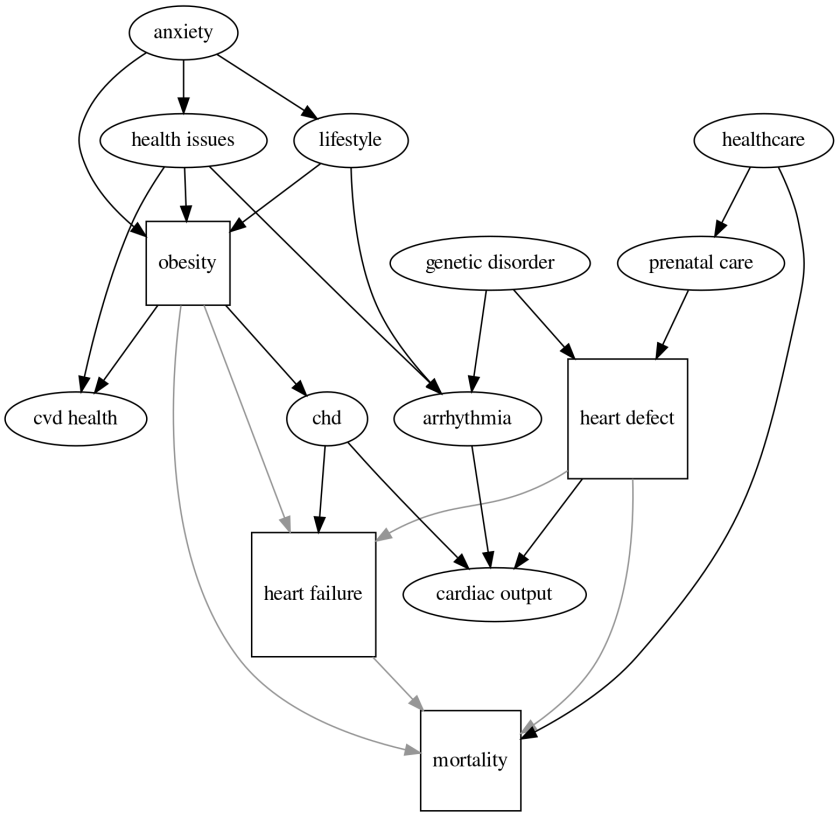


(b) Human

Figure 5: Illustration of expanded causal graphs for Diabetes. Squared nodes represent initial variables, while round nodes denote new proposed variables.

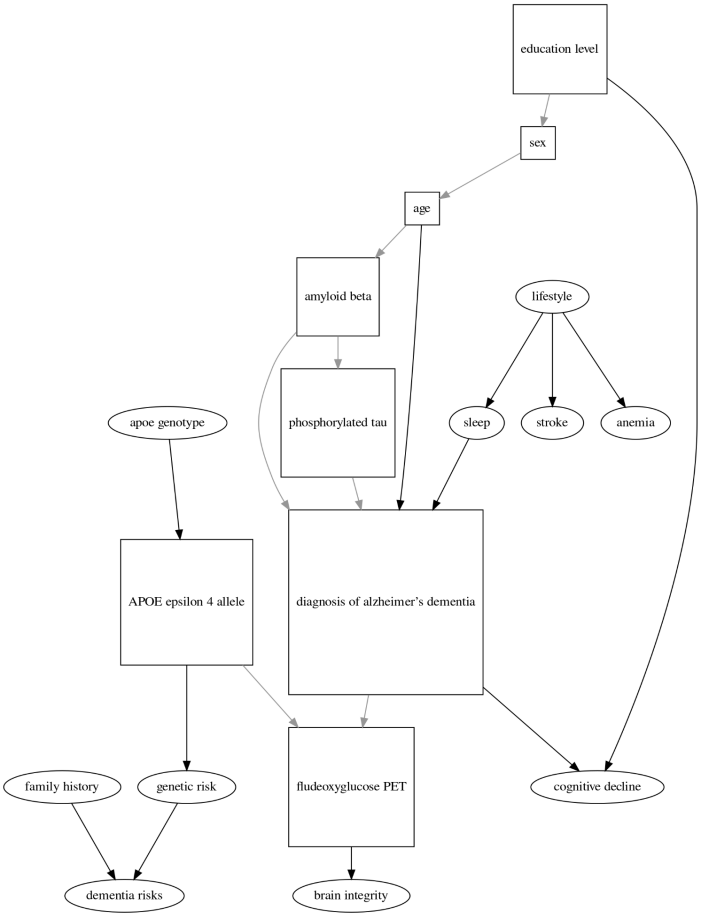


(a) IRIS

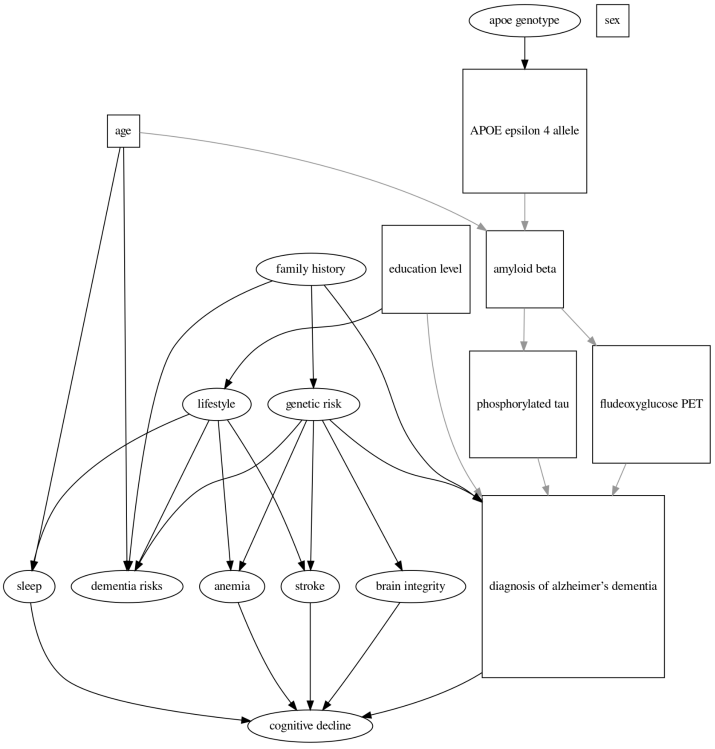


(b) Human

Figure 6: Illustration of expanded causal graphs for Obesity. Squared nodes represent initial variables, while round nodes denote new proposed variables.

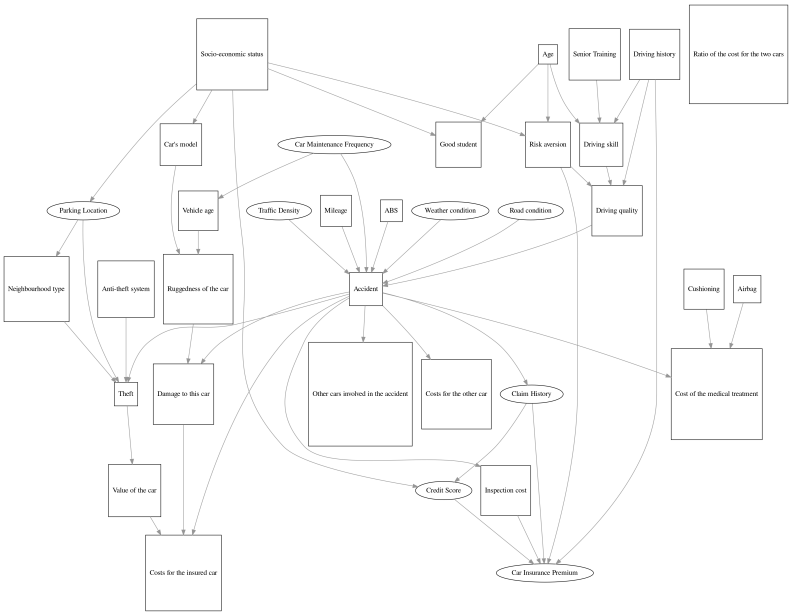


(a) IRIS

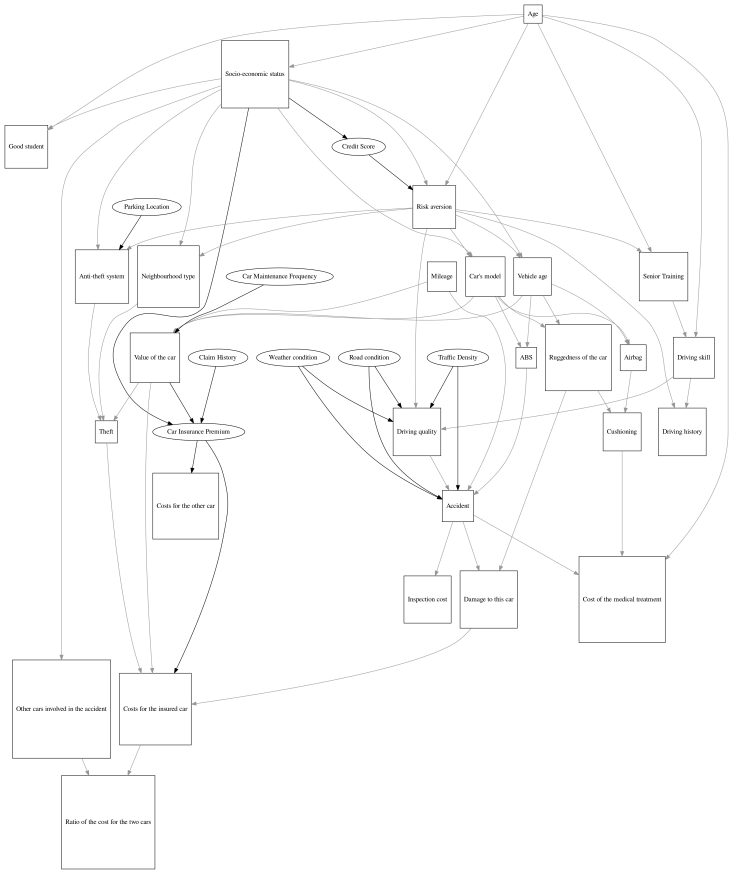


(b) Human

Figure 7: Illustration of expanded causal graphs for ADNI. Squared nodes represent initial variables, while round nodes denote new proposed variables.

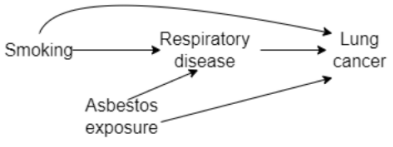


(a) IRIS

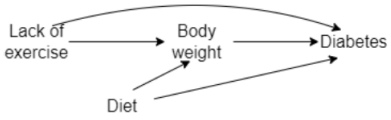


(b) Human

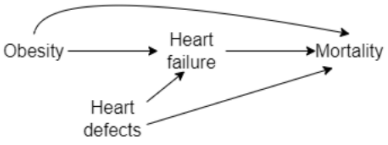
Figure 8: Illustration of expanded causal graphs for Insurance.



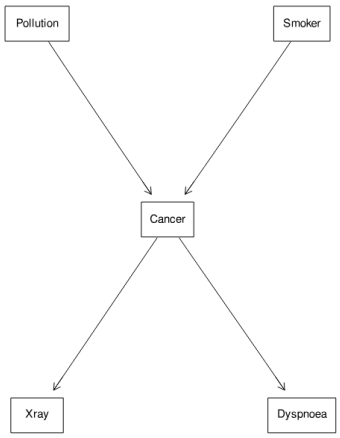
(a) Respiratory Disease



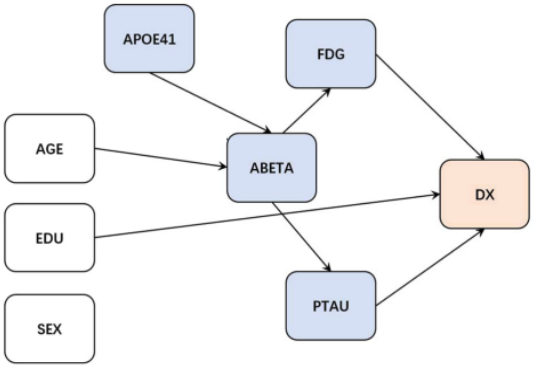
(b) Diabetes



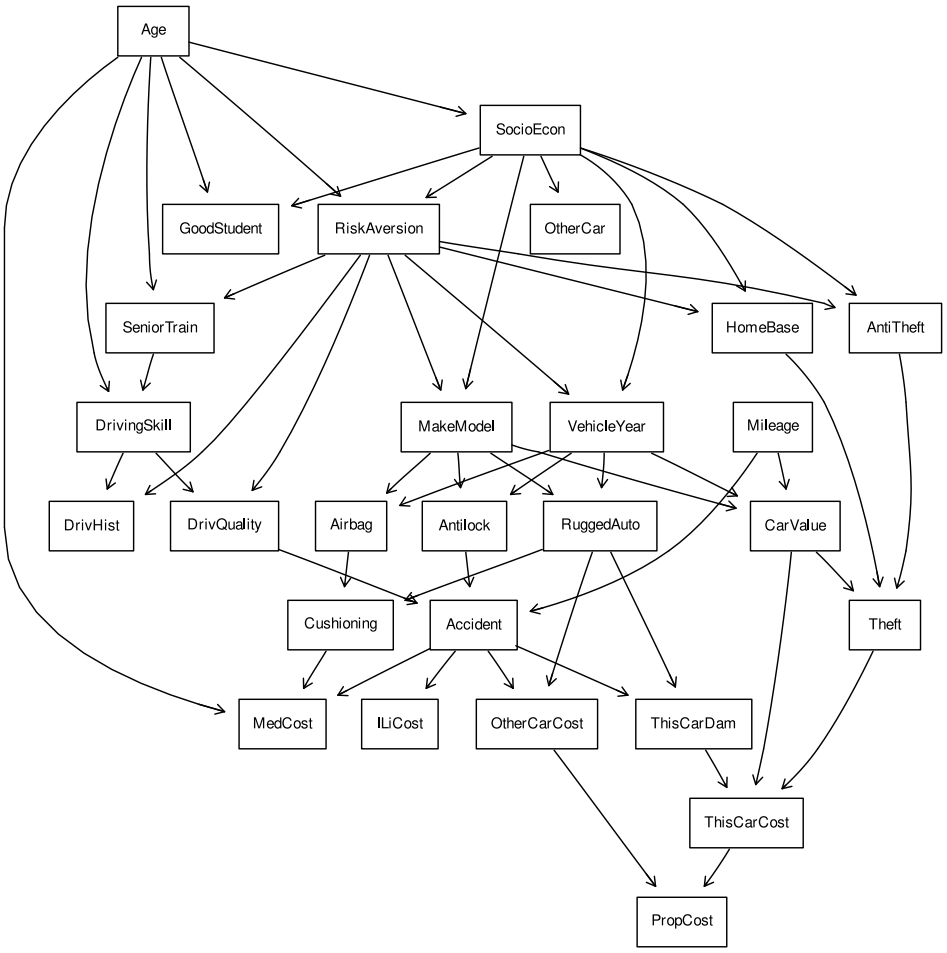
(c) Obesity



(d) Cancer



(e) Alzheimer's Disease Neuroimaging Initiative (ADNI)



(f) Insurance

Figure 9: The ground-truth causal graphs from original sources (Hernán et al., 2004; Long et al., 2022; Shen et al., 2020; Korb and Nicholson, 2010; Binder et al., 1997).

Algorithm 2 Hybrid Causal Discovery

Require: Initial variables $\mathbb{Z}$, LLM $\boldsymbol{M}$, structured data $\mathbb{X}$, prompt l, hyperparameters α, β

Statistical Causal Discovery

$\hat{\mathcal{G}}_s \leftarrow$ causal_discovery_alg($\mathbb{X}$) ▷ Apply causal discovery algorithms (e.g., PC algorithm)

Causal Relation Verification

$\hat{\mathcal{G}}_v \leftarrow$ causal graph with no edges
remove_edges $\leftarrow \emptyset$
for each z_i in $\mathbb{Z}$ **do**
  for each z_j in $\mathbb{Z}$ **do**
    if $\mathrm{z}_i \neq \mathrm{z}_j$ **then**
      $r \leftarrow$ "z_i causes z_j"
      $veracity_r \leftarrow \emptyset$ ▷ Initialize the veracity list for relation r
      for each d in $\mathbb{D}_{\mathrm{z}_i,\mathrm{z}_j}$ **do** ▷ $\mathbb{D}_{\mathrm{z}_i,\mathrm{z}_j}$ denotes documents containing both z_i and z_j
        $ver_d \leftarrow \boldsymbol{M}(l(r, d))$ ▷ Determine the veracity of r based on document d
        $veracity_r \leftarrow veracity_r \cup \{ver_d\}$
      end for
      if $veracity_r.count(True) > \alpha \times len(veracity_r)$ **then**
        $\hat{\mathcal{G}}_v \leftarrow \hat{\mathcal{G}}_v \cup \{r\}$ ▷ Add relation r to the causal graph $\hat{\mathcal{G}}_v$
      else if $veracity_r.count(False) > \beta \times len(veracity_r)$ **then**
        remove_edges $\leftarrow$ remove_edges $\cup \{r\}$
      end if
    end if
  end for
end for

Merge $\hat{\mathcal{G}}_s$ and $\hat{\mathcal{G}}_v$
for each edge r in $\hat{\mathcal{G}}_v$ **do**
  $\hat{\mathcal{G}}_s \leftarrow \hat{\mathcal{G}}_s \cup \{r\}$ ▷ Add relation r to $\hat{\mathcal{G}}_s$
end for
for each edge r in remove_edges **do**
  $\hat{\mathcal{G}}_s \leftarrow \hat{\mathcal{G}}_s \setminus \{r\}$ ▷ Remove relation r from $\hat{\mathcal{G}}_s$ if it exists
end for
$\hat{\mathcal{G}} \leftarrow \hat{\mathcal{G}}_s$ ▷ The final merged causal graph

Output: $\hat{\mathcal{G}}$

Algorithm 3 Missing Variable Proposal

Require: Initial variables $\mathbb{Z}$, LLM $\boldsymbol{M}$, collected documents $\mathbb{D}$, prompt l, hyperparameter α

Step 1: Abstract Missing Variable Candidates

$\mathbb{Z}_c \leftarrow \emptyset$ ▷ Initialize the set of candidates
for each document d in $\mathbb{D}$ **do**
  z $\leftarrow \boldsymbol{M}(l(\mathbb{Z}, d))$ ▷ Abstract a candidate variable from document d
  $\mathbb{Z}_c \leftarrow \mathbb{Z}_c \cup \{\mathrm{z}\}$
end for

Step 2: Missing Variable Proposal Based on Verified Causal Relations

$\mathbb{Z}_m \leftarrow \emptyset$ ▷ Initialize the set of missing variables
for each variable z_i in $\mathbb{Z}_c$ **do**
  for each given variable z_j in $\mathbb{Z}$ **do**
    $r_1 \leftarrow$ "z_i causes z_j"
    $veracity_{r_1} \leftarrow \emptyset$ ▷ Initialize the veracity list for relation r_1
    for each document d in $\mathbb{D}_{\mathrm{z}_i,\mathrm{z}_j}$ **do** ▷ $\mathbb{D}_{\mathrm{z}_i,\mathrm{z}_j}$ denotes documents containing both z_i and z_j
      $ver_d \leftarrow \boldsymbol{M}(l(r_1, d))$ ▷ Determine the veracity of $r1$ based on document d
      $veracity_{r_1} \leftarrow veracity_{r_1} \cup \{ver_d\}$
    end for
    if veracity$_{r_1}$.$count$(True) $>$ $\alpha \times$ veracity$_{r_1}$.$count$(False) **then**
      $\mathbb{Z}_m \leftarrow \mathbb{Z}_m \cup \{\mathrm{z}_i\}$ ▷ Add z_i to the set of proposed variables
    end if
    $r_2 \leftarrow$ "z_j causes z_i" ▷ Repeat the process for the reverse causal relation
  end for
end for

Step 3: Missing Variable Proposal Based on Statistical Methods

$\mathbb{S} \leftarrow \emptyset$ ▷ Initialize a set for PMI scores
for each variable z_i in $\mathbb{Z}_c$ **do**
  $s_i \leftarrow \emptyset$
  for each given variable z_j in $\mathbb{Z}$ **do**
    $s_{ij} \leftarrow \mathrm{PMI}(\mathrm{z}_i, \mathrm{z}_j)$ ▷ Compute PMI of $(\mathrm{z}_i, \mathrm{z}_j)$ by Equation 1
    $s_i \leftarrow s_i \cup \{s_{ij}\}$
  end for
  $\mathbb{S} \leftarrow \mathbb{S} \cup \{\sum(s_i)\}$ ▷ Aggregate the PMI scores for z_i
end for
$\mathbb{Z}_m \leftarrow \mathbb{Z}_m \cup$ top-k($\mathbb{S}, \mathbb{Z}_c$) ▷ PMI scores

Output: $\mathbb{Z}_m$

Causal Graph	Node	Edge
Cancer	16	28
Respiratory Disease	13	22
Diabetes	15	26
Obesity	14	25
ADNI	18	27
Insurance	35	67

Table 8: Statistics of human-annotated causal graph for expanded variables.

Causal Graph	Node	Edge
Cancer	5	4
Respiratory Disease	4	5
Diabetes	4	5
Obesity	4	5
ADNI	8	7
Insurance	27	52

Table 9: Statistics of ground-truth causal graph for initial variables.

Cancer (5 nodes, 4 edges)					
Method	Precision	Recall	**F1↑**	# of predicted edges	**NHD Ratio↓**
Pairwise-LLM	0.75	0.75	0.75	4	0.25
BFS-LLM	0.6	0.75	0.67	5	0.33
COAT	0.13	0.25	0.17	8	0.83
IRIS- GES	0.25	0.5	0.33	8	0.67
IRIS- NOTEARS	1.0	0.25	0.4	1	0.6
IRIS- PC	0.14	0.25	0.18	7	0.82
IRIS- VCR	1.0	0.75	**0.86**	3	**0.14**
IRIS (Llama) - NOTEARS+VCR	0.375	0.75	0.5	8	0.5
IRIS- NOTEARS+VCR	1.0	0.75	**0.86**	3	**0.14**

Table 10: Evaluation results of causal discovery on cancer graph. VCR refers to verified causal relations that are extracted from and validated by relevant academic documents. "Llama" refers to the use of the Llama-3.1-8b-instruct model as a substitute for GPT-4o in our method.

Respiratory Disease (4 nodes, 5 edges)					
Method	Precision	Recall	**F1↑**	# of predicted edges	**NHD Ratio↓**
Pairwise-LLM	1.0	0.6	0.75	3	0.25
BFS-LLM	0.67	0.4	0.5	3	0.5
COAT	1.0	0.8	0.89	4	0.11
IRIS- GES	1.0	0.8	0.89	4	0.11
IRIS- NOTEARS	1.0	0.2	0.33	1	0.67
IRIS- PC	0.83	1.0	0.91	6	0.09
IRIS- VCR	1.0	0.8	0.89	4	0.11
IRIS (Llama) - PC+VCR	1.0	0.8	0.89	4	0.11
IRIS- PC+VCR	0.83	1.0	**0.91**	6	**0.09**

Table 11: Evaluation results of causal discovery on respiratory disease graph.

Diabetes (4 nodes, 5 edges)					
Method	Precision	Recall	**F1↑**	# of predicted edges	**NHD Ratio↓**
Pairwise-LLM	0.67	0.4	0.5	3	0.5
BFS-LLM	0.67	0.4	0.5	3	0.5
COAT	0.25	0.2	0.22	4	0.78
IRIS- GES	0.5	0.6	0.55	6	0.45
IRIS- NOTEARS	0	0	0	0	1.0
IRIS- PC	0.25	0.2	0.22	4	0.78
IRIS- VCR	1.0	0.2	0.33	1	0.67
IRIS (Llama) - GES+VCR	0.67	0.4	0.5	3	0.5
IRIS- GES+VCR	1.0	0.6	**0.75**	3	**0.25**

Table 12: Evaluation results of causal discovery on diabetes graph.

Obesity (4 nodes, 5 edges)					
	Precision	Recall	**F1↑**	# of predicted edges	**NHD Ratio↓**
Pairwise-LLM	0.83	1.0	0.91	6	0.09
BFS-LLM	0.6	0.6	0.6	5	0.4
COAT	0.25	0.2	0.22	4	0.78
IRIS-GES	0.25	0.2	0.22	4	0.78
IRIS- NOTEARS	0	0	0	2	1.0
IRIS- PC	0.25	0.2	0.22	4	0.78
IRIS- VCR	1.0	1.0	**1.0**	5	**0**
IRIS (Llama) - PC+VCR	0.83	1.0	0.91	6	0.09
IRIS- PC+VCR	1.0	1.0	**1.0**	5	**0**

Table 13: Evaluation results of causal discovery on obesity graph.

ADNI (8 nodes, 7 edges)					
Method	Precision	Recall	**F1↑**	# of predicted edges	**NHD Ratio↓**
Pairwise-LLM	0.5	0.14	0.22	2	0.78
BFS-LLM	0.33	0.14	0.2	3	0.8
COAT	0.11	0.14	0.13	9	0.87
IRIS- GES	0.08	0.14	0.11	12	0.89
IRIS- NOTEARS	0.33	0.14	0.2	3	0.8
IRIS- PC	0.11	0.14	0.13	9	0.87
IRIS- VCR	0.4	0.29	0.33	5	0.67
IRIS (Llama) - NOTEARS+VCR	0.08	0.14	0.11	12	0.89
IRIS- NOTEARS+VCR	0.38	0.43	**0.4**	8	**0.6**

Table 14: Evaluation results of causal discovery on ADNI graph.

Insurance (27 nodes, 52 edges)					
Method	Precision	Recall	**F1↑**	# of predicted edges	**NHD Ratio↓**
Pairwise-LLM	0.37	0.34	0.35	58	0.68
BFS-LLM	0.33	0.24	0.28	41	0.76
COAT	0.38	0.37	0.37	53	0.65
IRIS- GES	0.41	0.47	0.44	37	0.58
IRIS- NOTEARS	0.28	0.40	0.32	33	0.71
IRIS- PC	0.31	0.40	0.35	31	0.69
IRIS- VCR	0.4	0.29	0.33	35	0.67
IRIS (Llama) - GES+VCR	0.43	0.49	0.46	43	0.55
IRIS- GES+VCR	0.58	0.53	**0.55**	47	**0.45**

Table 15: Evaluation results of causal discovery on Insurance graph.