Context-Aware Reasoning On Parametric Knowledge for Inferring Causal Variables

Ivaxi Sheth[1], Sahar Abdelnabi[2], Mario Fritz[1]
[1]CISPA Helmholtz Center for Information Security, [2] Microsoft
{ivaxi.sheth,fritz}@cispa.de, saabdelnabi@microsoft.com

Abstract

Scientific discovery catalyzes human intellectual advances, driven by the cycle of hypothesis generation, experimental design, evaluation, and assumption refinement. Central to this process is causal inference, uncovering the mechanisms behind observed phenomena. While randomized experiments provide strong inferences, they are often infeasible due to ethical or practical constraints. However, observational studies are prone to confounding or mediating biases. While crucial, identifying such backdoor paths is expensive and heavily depends on scientists' domain knowledge to generate hypotheses. We introduce a novel benchmark where the objective is to complete a partial causal graph. We design a benchmark with varying difficulty levels with over 4000 queries. We show the strong ability of LLMs to hypothesize the backdoor variables between a cause and its effect. Unlike simple knowledge memorization of fixed associations, our task requires the LLM to reason according to the context of the entire graph[1].

1 Introduction

Scientific discovery has been key to humankind's advances. It is a dynamic process revolving around inquiry and refinement. Scientists adhere to a process that involves formulating a hypothesis and then collecting pertinent data (Wang et al., 2023). They then draw inferences from these experiments, modify the hypothesis, formulate sub-questions, and repeat the process until the research question is answered (Kıcıman et al., 2023).

Central to scientific discovery is formulating hypotheses and identifying relevant variables that drive the underlying causal mechanisms of observed phenomena (Bunge, 2017). Randomized controlled trials are the gold standard for establishing causal relationships, but they are often in-
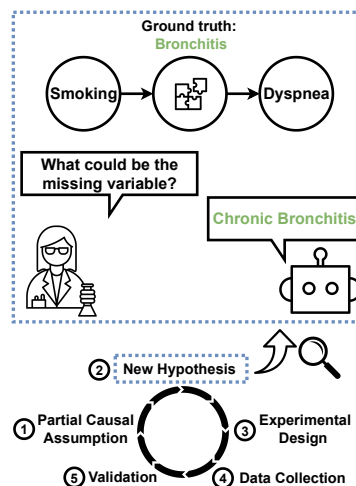

Figure 1: Scientific discovery iteratively generates hypotheses from assumptions using human expertise. We use LLMs as proxy experts to propose new hypotheses about missing variables in causal DAGs.

feasible due to ethical, financial, or logistical constraints (Nichol et al., 2010). In such cases, researchers rely on observational data, where a key challenge lies not only in analyzing relationships but in determining which variables should be observed and included in the analysis, particularly confounders or mediators that influence causal mechanisms underlying the outcomes (Ananth and Schisterman, 2017; Gupta et al., 2021).

With the recent advancement of Large Language Models (LLMs), there has been a growing interest in using them for scientific discovery (AI4Science and Quantum, 2023; Lu et al., 2024; Cory-Wright et al., 2024). LLMs have demonstrated strong performance in internalizing knowledge (Sun et al., 2024; Yu et al., 2024) and reasoning-based tasks (Valmeekam et al., 2023; Guo et al., 2025), including causal discovery, where they infer pairwise causal relationships based on variable semantics (Kıcıman et al., 2023; Long et al., 2023; Ban et al., 2023; Vashishtha et al., 2023; Darvariu et al., 2024; Binkyte et al., 2025).

Scientific reasoning is fundamentally context-

---

[1]Code available at https://github.com/ivaxi0s/inferring-causal-variables

driven; unlike simple factual retrieval, it requires adapting hypotheses based on new evidence and integrating knowledge across varying subpopulations. While recent work has explored the use of LLMs for causal discovery (Kıcıman et al., 2023; Long et al., 2023; Darvariu et al., 2024; Ban et al., 2023; Vashishtha et al., 2023), much of it assumes a fixed set of variables and focuses on identifying relationships among them. However, a critical and underexplored aspect is determining *which* variables should be considered in the first place. This demands flexible, context-sensitive reasoning to identify missing causal factors.

In our paper, we use the term reasoning operationally to describe the model's ability to generate hypotheses or identify variables that complete partial causal graphs. Our usage follows previous work by Kıcıman et al. (2023) in causal discovery and LLM research, where "reasoning" often refers to generating plausible hypotheses or prioritizing potential candidates given partial structural information, rather than strict deductive logic.

To address this gap, we propose a novel task: given a partial causal graph with missing variables, the LLM is prompted to hypothesize what those variables might be, using the structure and known nodes as context. By systematically omitting different variables, we generate diverse test cases to evaluate the robustness of model reasoning. We further decompose the benchmark into subtasks, starting from baseline variable identification to more realistic, open-ended settings where multiple unobserved mediators exist between known treatments and outcomes.

Our task mirrors real-world scientific workflows, where identifying missing variables—especially confounders and mediators is essential for valid causal inference. This typically demands costly, interdisciplinary effort. LLMs, trained on diverse knowledge sources, offer a scalable alternative. For example, in a stroke drug study, an LLM might suggest socioeconomic status as an unmeasured confounder. While recent works advocate using LLMs as co-pilots for causal tasks (Petersen et al., 2024; Alaa et al., 2024), systematic evaluations are lacking. Our benchmark addresses this gap by assessing LLMs' ability to infer missing causal variables across domains.

Our main **contributions** are: 1) We propose and formalize the novel task of LLM-assisted causal variable inference. 2) We propose a benchmark for inferring missing variables across diverse domains of causal graphs. 3) We design experimental tasks with different difficulty levels and knowledge assumptions, such as open-world and closed-world settings, the number of missing variables, etc. 4) Our benchmark allows for both grounded evaluations and a reproducible framework to benchmark LLMs' capabilities in hypothesis generation.

2 Related Work

LLMs and Causality. Our work builds on the foundational framework of causality by Pearl (2009). Prior studies have explored extracting causal relationships from text (Girju et al., 2002; Hassanzadeh et al., 2020; Tan et al., 2023; Dhawan et al., 2024) and using LLMs for causal reasoning (Kıcıman et al., 2023), including commonsense (Frohberg and Binder, 2021; Singh et al., 2021) and temporal causality (Zhang et al., 2020, 2022). Recent efforts prompt LLMs with variable names to discover causal structures (Kıcıman et al., 2023; Long et al., 2023; Darvariu et al., 2024; Ban et al., 2023; Vashishtha et al., 2023). Others integrate LLMs with deep structural causal models (Abdulaal et al., 2024; Yu et al., 2019), or focus on graph formatting (Sheth et al., 2024), query design (Jiralerspong et al., 2024), and causal inference (Jin et al., 2023). In contrast to prior work, we use LLMs to infer missing variables before data collection and evaluation, leveraging their pre-trained knowledge for this novel hypothesizing task.

LLMs and Hypothesis Generation. Existing work tested hypothesis generation with LLMs in reasoning tasks or free-form scientific hypotheses from background knowledge provided in the context (Gendron et al., 2023; Qi et al., 2023; Xu et al., 2023a,b; Qiu et al., 2024; Lu et al., 2024). In contrast, we consider the structured task of causal hypothesis generation, where the ground-truth variables are known and can be used for evaluation.

Context-aware reasoning has been explored through prompt engineering (Dutta et al., 2024; Zhou et al., 2023; Ranaldi and Zanzotto, 2023), premise ordering manipulation (Chen et al., 2024), diagnostic analyses (Prabhakar et al., 2024), and compositional reasoning evaluations (Press et al., 2022; Saparov et al., 2024). Unlike premise-based or linguistic evaluations, our setup requires reasoning over causal graph topology, using contextual cues by varying assumptions.

3 Preliminaries: Causal Graph

A causal relationship can be modeled via a Directed Acyclic Graph (DAG). A causal DAG represents relationships between a set of N variables defined by $\mathbf{V} = \{v_1, ..., v_N\}$. The variables are encoded in a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ where E is a set of directed edges between the nodes $\in$ V such that no cycle is formed. Mathematically, it can be expressed as:

$$\mathcal{G} = (\mathbf{V}, \mathbf{E}),$$

$$\mathbf{E} = \{e_{i,j} \mid v_i, v_j \in \mathbf{V}, i \neq j \text{ and } v_i \rightarrow v_j\}$$

Each edge $e_{i,j}$ denotes causal relationship and the influence from v_i to v_j, $v_i \xrightarrow{e_{i,j}} v_j$.

We define $\mathbf{d}(v)$ as the degree of a node v, representing the total number of edges connected to v. $\mathbf{d}_{\text{in}}(v)$ is the in-degree, representing the number of incoming edges to v. $\mathbf{d}_{\text{out}}(v)$ is the out-degree, representing the number of outgoing edges from v.
Source has no incoming edges; $d_{\text{in}}(v) = 0$.
Sink has no outgoing edges. Sinks are $d_{\text{out}}(v) = 0$.
Treatment is characterized by nodes that are being intervened upon.
Outcome is characterized by nodes that are observed for interventions from the treatments.
Mediator has both incoming and outgoing edges $(d_{\text{in}}(v) > 0$ and $d_{\text{out}}(v) > 0)$ as intermediaries in the pathways between treatment and outcome.
Confounder influences both treatment and outcome, exhibiting edges directed towards the treatment and outcome nodes $(d_{\text{out}}(v) \geq 2)$. Hence v is a confounder if it is a parent of both v_i and v_j.
Collider has two edges meeting, and $d_{\text{in}}(v) > 1$. I.e., v is a collider if it is a child of both v_i and v_j.

4 Inferring Causal Variables

Motivated by the challenge of discovering variables that block backdoor paths to ensure unbiased causal inference (Glymour et al., 2019), in this work, we leverage language models to infer missing variables in a causal DAG. We assume that a part of the graph is already known, and the aim is to find additional variables that can be incorporated into the existing DAG to enhance the underlying causal mechanism.

Formally, we assume a partially known causal DAG, $\mathcal{G}^* = (\mathbf{V}^*, \mathbf{E})$, where $\mathbf{V}^* \subseteq \mathbf{V}$. The objective is to identify the set of missing variables $V^* = \mathbf{V} \setminus \mathbf{V}_{\text{missing}}$ thereby expanding $\mathcal{G}^*$ to $\mathcal{G}$. This implies that all causal relationships (edges) among variables in V^* are known and correctly represented in $\mathcal{G}^*$; i.e., E is fully specified. Here,

"missing" variables are not latent or hidden by measurement error but known unknowns within the causal graph reflective of the LLM's perspective.

To systematically assess LLMs' ability to infer missing causal variables, we construct a multi-stage benchmark with increasing levels of complexity. We begin with a controlled setting, where the model is provided with a partial causal DAG and a set of multiple-choice options to identify missing variables. Then, the task becomes open-ended, where LLMs hypothesize missing variables, simulating an open-world paradigm. Additionally, as the task escalates, we introduce more complexity by omitting additional nodes, challenging the model to hypothesize multiple missing variables.

We evaluate the reasoning capability of LLMs through prompting. We represent the graph $\mathcal{G}^*$ using a prompt template $P_{\text{LLM}}(\cdot)$ which enables LLMs to parse causal relationships in the DAG.

4.1 Task 1: Out-of-Context Identification

Motivation. To assess whether LLMs can infer missing variables in causal graphs, we begin with a controlled multiple-choice setting that serves as a baseline. This task isolates the core challenge: identifying a single missing variable from a causal DAG. By restricting the search space to a fixed set of options, including the correct variable and out-of-context distractors, we evaluate whether the model can distinguish the variable that meaningfully completes the causal structure.

The partial DAG $\mathcal{G}^*$ is created by removing one variable, denoted as v_x, from the original DAG $\mathcal{G}$. The role of the LLM is to select a variable from the multiple choices, MCQ_{v_x}, that can be used to complete the graph. The out-of-context distractors are unrelated to the causal domain of the given DAG, chosen to minimize any contextual overlap with the true missing variable. Let v_x^* represent the variable selected by the LLM to complete $\mathcal{G}^*$.

$$v_x^* = P_{\text{LLM}}(\mathcal{G}^*, \text{MCQ}_{v_x}) \quad \forall v_x \in \mathbf{V}$$

4.2 Task 2: In-Context Identification

Motivation. In real-world domains like healthcare and finance, missing or unobserved variables often challenge causal inference (Hughes et al., 2019; Tian and Pearl, 2012). This task simulates such ambiguity by requiring LLMs to identify a relevant missing variable when presented with multiple plausible options, going beyond the baseline.
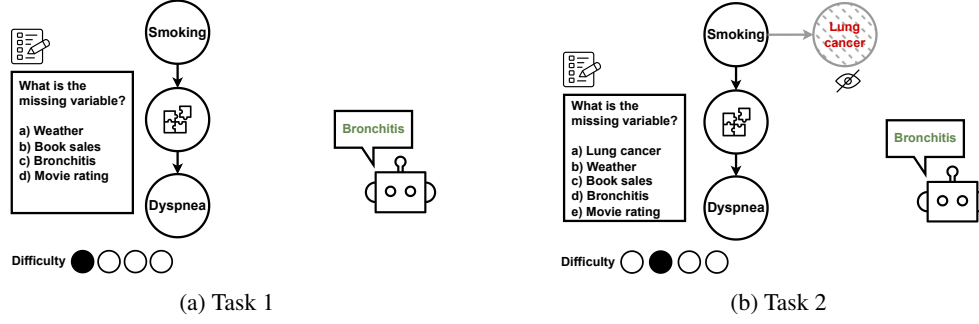
(a) Task 1

(b) Task 2

Figure 2: Leveraging LLM to identify the missing variable for a causal DAG in the presence of out-of-context distractors (a), an in-context distractor along with out-of-context distractors (b).

Here, instead of removing one node from the ground truth DAG $\mathcal{G}$, two nodes, v_{x_1} and v_{x_2}, are now removed to create the partial graph, $\mathcal{G}^*$.

$$\mathcal{G}^* = \mathcal{G} \setminus \{v_{x_1}, v_{x_2}\} \quad \text{for} \quad v_{x_1}, v_{x_2} \in \mathbf{V}$$

The MCQA paradigm provides multiple choices, including the missing variables v_{x_1} and v_{x_2}. The task for the LLM here is to select the correct variable v_{x_1} only, given an in-context choice v_{x_2} and out-of-context choices. The in-context variables are plausible within the same causal graph, allowing the LLM to use DAG-defined context inference to distinguish the relevant from the irrelevant options. We ensure v_{x_1} and v_{x_2} are not directly connected i.e., neither is a parent of the other.

$$v_{x_1}^* = P_{\text{LLM}}(\mathcal{G}^*, \text{MCQ}_{v_{x_1}, v_{x_2}}) \quad \forall \, v_{x_1}, v_{x_2} \in \mathbf{V}$$

$$\text{and} \quad v_{x_1} \not\to v_{x_2}, \quad v_{x_2} \not\to v_{x_1}$$

4.3 Task 3: Hypothesizing in Open World

Motivation. Previous tasks constrained the model to select from predefined options. However, real-world reasoning rarely offers such scaffolding. This task increases complexity by removing the multiple-choice format entirely.

Given a partial DAG $\mathcal{G}^*$, formed by removing a node v_x, the model must generate potential missing variables without any provided candidates (see Figure 3a). The output is a ranked list of hypotheses $\{v_{x,1}^*, ..., v_{x,k}^*\}$ for k suggestions, simulating open-ended discovery.

$$\{v_{x,1}^*, v_{x,2}^*, ..., v_{x,k}^*\} = P_{\text{LLM}}(\mathcal{G}^*) \, \forall \, v_x \in \mathbf{V}$$

4.4 Task 4: Iteratively Hypothesizing in Open World

Motivation. Building on the open-world setting, we further increase task difficulty by removing multiple nodes from the causal graph. The goal is no longer to recover a single missing variable but to iteratively hypothesize a set of mediators that link a treatment to an outcome.

Given a partial DAG $\mathcal{G}^* = \mathcal{G} \setminus v_{x_1}, ..., v_{x_M}$, the task (illustrated in Figure 3b) involves generating a sequence of missing mediators $M = v_{m_1}, v_{m_2}, ..., v_{m_H}$ that plausibly connect a treatment variable v_t to an outcome variable v_y.

At each iteration i, , the LLM is prompted with the current partial graph and returns a hypothesis for the next mediator. This process continues until all of the mediators are inferred.

$$v_{m_i}^* = P_{\text{LLM}}(\mathcal{G}^* \cup \{v_{m_1}^*, ..., v_{m_{i-1}}^*\}),$$

for $i = 1, ..., H$. The sequence of mediators $M = \{v_{m_1}, v_{m_2}, ..., v_{m_H}\}$ is chosen at random.

To assess how mediator order affects performance, we draw on mediation analysis concepts (Pearl, 2014), specifically the Natural Direct Effect (NDE)—the treatment's effect not mediated by a variable—and the Natural Indirect Effect (NIE)—the portion mediated by it (see Appendix A.4). We propose the Mediation Influence Score (MIS) to quantify each mediator's impact between a treatment and outcome. Defined as the ratio of NIE to NDE, MIS is a scale-free, positive measure of a mediator's relative contribution:

$$\text{MIS}(v_{m_i}) = \left| \frac{\text{NIE}(v_{m_i})}{\text{NDE}(v_{m_i})} \right| \quad \text{for} \quad i = 1, ..., H.$$

This metric quantifies the relative importance of the indirect effect (through the mediator) compared to the direct impact. Mediators are then ranked and prioritized based on their MIS scores, with higher scores indicating a stronger mediation effect.

5 Evaluation and Results

Graphs. We evaluate a variety of causal graphs spanning diverse domains. We use the semi-
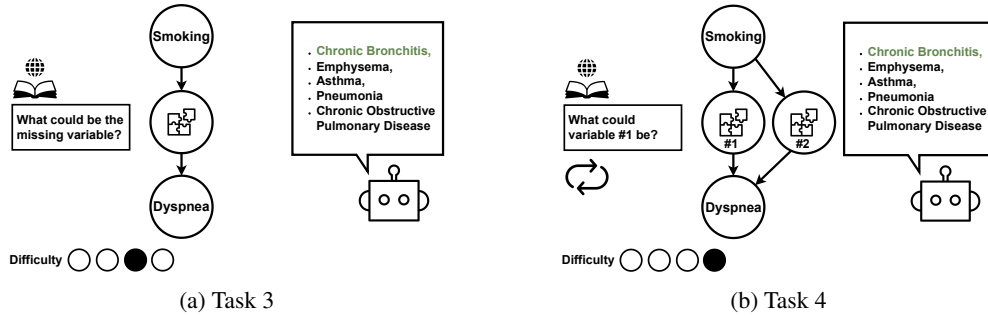
(a) Task 3   (b) Task 4

Figure 3: Leveraging LLM to hypothesize the missing variable in a causal DAG in an open-world setting for one variable (a), in an iterative fashion for multiple missing mediators (b).

synthetic DAGs from BNLearn repository - Cancer (Korb and Nicholson, 2010), Survey (Scutari and Denis, 2021), Asia (Lauritzen and Spiegelhalter, 1988), Child (Spiegelhalter, 1992), Insurance (Binder et al., 1997), and Alarm (Beinlich et al., 1989). We also evaluate our approach on a realistic Alzheimer's Disease graph (Abdulaal et al., 2024), developed by five domain experts and Law (VanderWeele and Staudt, 2011). See Appendix A.1 for further details.

Graph	V	E	Description
Cancer	5	4	Factors around lung cancer
Survey	6	6	Factors for choosing transportation
Asia	8	8	Factors affecting dysponea
Law	8	20	factors around legal system
Alzheimer	9	16	Factors around Alzheimer's Disease
Child	20	25	Lung related illness for a child
Insurance	27	52	Factors affecting car accident insurance
Alarm	37	46	Patient monitoring system

Table 1: Datasets used in the benchmark.

Models. We evaluate our setups across different open-source and closed models. The models we use are GPT-4o (Hurst et al., 2024), GPT-4 (OpenAI, 2023), LLama3-chat-8b (Touvron et al., 2023), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Mixtral-7B-Instruct-v0.1 (Jiang et al., 2024), Zephyr-7b-Beta (Tunstall et al., 2023) and Neural-chat-7b-v3-1 (Intel, 2023).

Prompt. We used the textual prompting strategy from Sheth et al. (2024) after performing experiments on some of the proposed encoding methods (see Appendix B.10). Implementation details are in Appendix A and prompts in Appendix F. Our code will be available after anonymity period.

5.1 Task 1

Setup. The input to the LLM consists of a partial DAG $\mathcal{G}^*$, and multiple choices including the correct missing variable v_x and several out-of-context distractors. This task includes 120 queries. We define accuracy to assess the LLM's v_x prediction.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(v_x^* = v_x^i)$$

Results. In Figure 4a, we report the accuracy of different LLMs in identifying the missing variable. GPT-4, followed closely by Mixtral and GPT-4o, consistently performs well, achieving perfect accuracy on most of the graphs. Other models, including Mistral-7b, Llama-7b, Neural, and Zephyr-7b, have varying degrees of success. Insurance remains the most challenging graph, potentially due to the high number of edges present in the DAG. All models significantly outperform the random baseline. However, we conjecture that the high performance could be partially attributed to the simplicity of the task. The models might be using the context of the graph domain to exclude unrelated distractors rather than engaging in deeper causal reasoning among multiple plausible choices. To investigate this, we introduce an in-domain choice among the multiple choices in the next experiment.

5.2 Task 2

Setup. This is a more challenging task where the partial graph has two missing nodes. In addition to out-of-context distractors and the ground-truth variable, v_{x_1}, the multiple-choice set includes the second missing variable v_{x_2} as an in-context distractor. This setup tests the model's ability to reason over indirect causal relations contextually to identify the correct variable. This task results in over 3800 queries. To evaluate performance, we use two metrics: Accuracy and False Node Accu-
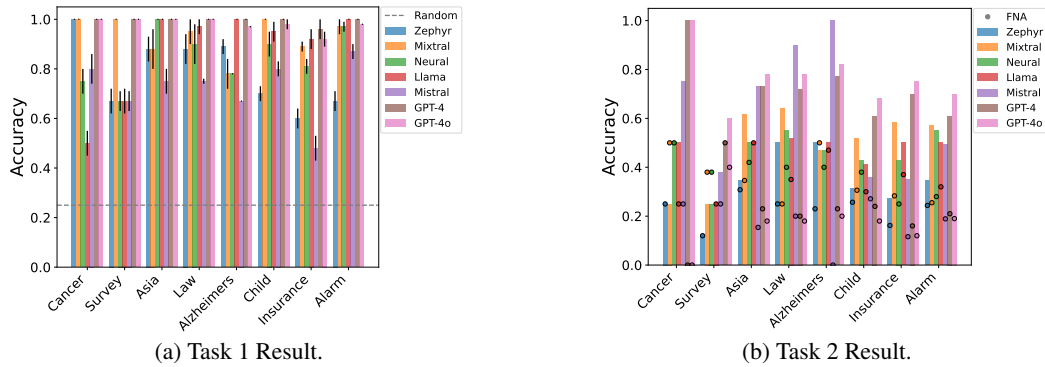
(a) Task 1 Result.

(b) Task 2 Result.

Figure 4: Accuracy of LLMs in identifying the missing causal variable from multiple choices with out-of-context distractors (a), and from both out-of-context and in-context distractors (b).

racy (FNA). FNA captures how often the model incorrectly selects the in-context distractor instead:

$$\text{FNA} \downarrow = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(v_{x_1}^* = v_{x_2})$$

Results. In Figure 4b, we report Accuracy and False Node Accuracy (FNA) across graphs. Accuracy reflects how often the correct missing variable is chosen, while FNA measures how often the model incorrectly selects the in-context distractor—another missing variable included to test deeper causal reasoning. Since there are 5 options, random accuracy is 0.2, and FNA under random guessing would be around 0.2 as well. GPT-4 and GPT-4o achieve high accuracy and low FNA, showing that they reliably distinguish the true missing node from both distractors and the in-context variable. GPT-4o slightly outperforms GPT-4 on several graphs. Open models like Mistral, Zephyr, and Mixtral show more variability, performing well on simpler graphs like Cancer but struggling on complex ones like Alarm. While most models exceed random chance, higher FNA in some cases highlights a tendency to confuse plausible but incorrect variables, emphasizing the difficulty of reasoning over multiple missing nodes.

5.3 Task 3

Setup. In real-world settings, partial causal graphs provided by domain experts often lack ground truth and multiple choices. Hypotheses may vary depending on context, data, or domain knowledge. To simulate this, we prompt the LLM to generate. The LLM generates $k = 5$ suggestions for the missing node v_x. This task has 120 queries. We compare suggestions to the ground truth, recognizing that real-world cases often lack a single correct answer.

Since traditional metrics may miss contextual nuances, we use two evaluations: semantic similarity and LLM-as-Judge (see Appendix B.4).

1. **Semantic Similarity.** We compute the cosine similarity between the embeddings of the predictions, $v_{x_{1:5}}^*$, and the ground truth v_x, averaging the highest similarity scores across all nodes $v_x \in \mathbf{V}$ (see Appendix A.5 for details).

2. **LLM-Judge.** Inspired by Zheng et al. (2023), this two-step metric assesses contextual semantic similarity beyond exact matches. First, LLM ranks suggestions $v_{x_{1:5}}^*$ based on how well they fit the partial graph. Second, it rates the best match on a 1–10 scale. Scores are averaged across nodes for an overall measure (see Appendix A.6).

Results. We report models' performances using both semantic similarity and LLM-Judge metrics in Table 2. For brevity, we provided the variances in Appendix B.1. We provide a detailed analysis of each metric across different types of node variables (defined in Section 3). We evaluate sources, sinks, colliders, and mediators for each of the partial causal graphs. The results, fine-grained by node type, are given in Figure 5, which shows each model's average performance across graphs with a detailed performance per graph in Figure 8. GPT-4, GPT-4o and Mixtral generally achieve higher semantic similarity and LLM-as-Judge scores across most graphs (Figure 8). We observe that semantic similarity is a stricter metric than LLM-as-judge since it cannot encode contextual information about the causal DAG (see example in Table 10). Despite different scales, both metrics seem to be fairly correlated. Figure 5, shows that models display stronger performance for colliders and mediators

	Cancer		Survey		Asia		Law		Alzheimers		Child		Insurance		Alarm		Avg	
	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J
Zephyr	0.36	0.61	0.34	0.60	0.45	0.66	0.41	0.70	0.35	0.75	0.51	0.70	0.45	0.44	0.46	0.69	0.42	0.63
Mixtral	0.41	0.66	0.39	0.66	**0.66**	0.75	0.38	0.69	0.31	0.77	**0.53**	**0.77**	0.46	0.56	**0.50**	0.72	0.46	0.70
Neural	0.38	0.77	0.43	0.55	0.53	0.55	0.47	0.72	0.44	0.71	0.48	0.70	0.47	0.43	0.47	0.67	0.45	0.63
Llama	0.40	0.48	0.40	0.54	0.53	0.58	**0.67**	0.65	0.45	0.61	0.48	0.63	0.42	0.34	0.46	0.65	0.45	0.55
Mistral	0.33	0.67	0.44	0.65	0.60	0.73	0.49	0.67	0.34	0.76	0.48	0.68	0.46	0.47	0.47	0.71	0.44	0.67
GPT-4	**0.49**	**0.90**	**0.51**	0.67	**0.66**	0.76	0.55	0.78	0.47	**0.98**	0.36	0.53	0.52	0.56	0.49	0.75	0.50	0.73
GPT-4o	0.52	0.89	0.50	**0.71**	**0.66**	**0.78**	0.58	**0.80**	**0.50**	0.91	0.40	0.60	**0.54**	**0.58**	0.44	**0.76**	**0.54**	**0.76**

Table 2: Task 3 Results. Average semantic similarity and LLM-as-Judge metrics to evaluate LLMs in hypothesizing the missing variable in a causal DAG.

on average. This suggests that these models are better at reasoning about common causes and indirect causal relationships. Sinks are typically the nodes that represent the outcomes or effects of interventions (treatments) applied to other nodes. Source nodes represent the causes in a causal graph. Lower performance on these nodes indicates to reason about the potential causes and outcomes of the causal graphs is difficult.

In Figure 6a, model performance improves with more suggestions (k). Figure 6b shows that accuracy also correlates with node degree ($d_{in} + d_{out}$), indicating that more context aids prediction. Overall, LLMs perform well on many nodes, especially mediators and colliders, making them promising tools for real-world causal discovery where treatments and outcomes are known.
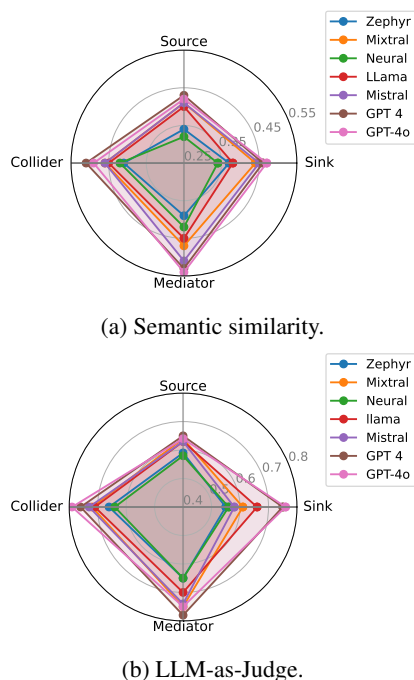

(a) Semantic similarity.

(b) LLM-as-Judge.

Figure 5: Task 3 Results. Visualizing each model's performances, averaged across the different graphs, for Sink, Source, Mediator, and Collider nodes.

5.3.1 Hypothesizing Confounder

Backdoor paths are alternative causal pathways that confound the estimation of causal effects and introduce bias if not accounted for. Hence, hypothesizing and controlling for confounders is an important task in causal inference (Pourhoseingholi et al., 2012). We extract confounder subgraphs from (Sachs et al., 2005), Alarm, and Insurance graphs. From Table 3 and Appendix D, we find

	Sachs	Alarm	Insurance
Zephyr	0.10	0.45	0.53
Mixtral	**0.95**	**0.85**	0.63
Neural	0.30	0.45	0.61
LLama	0.20	0.47	0.63
Mistral	0.20	**0.85**	0.61
GPT-4	**0.95**	0.73	**0.78**
GPT-4o	**0.95**	0.70	0.73

Table 3: Hypothesizing Confounders in Task 3.

that while LLMs accurately hypothesize some confounders, models struggle with domain-specific graphs like SACHS. Larger models like GPT-4o don't necessarily always perform best, underscoring the need for diverse benchmarks.

5.4 Task 4

Setup. We adopt an iterative approach for hypothesizing mediators, allowing the model to refine predictions step-by-step—unlike global prediction, which yields lower performance (Appendix B.6). This aligns with Chain-of-Thought (Wei et al., 2022) reasoning and improves accuracy. There are more than 140 queries for this task, ranging from 1-10 missing mediators. For **unordered evaluation**, mediators are given in random order and scored via average semantic similarity. For **ordered evaluation**, we rank mediators using the Mediation Influence Score (MIS) and compare model performance when prompted in ascending vs. descending MIS order. We define a metric, Δ, to capture this

difference.

	Asia		Child		Insurance		Alarm	
	Sim	Δ	Sim	Δ	Sim	Δ	Sim	Δ
Zephyr	0.61	−0.02	**0.54**	0.17	0.47	0.19	0.51	0.20
Mixtral	**0.87**	0.01	0.50	0.18	0.48	0.15	0.52	0.13
Neural	0.65	0.04	0.48	0.21	0.42	0.16	0.46	0.12
Llama	0.80	0.07	0.49	−0.05	0.44	0.21	0.51	0.07
Mistral	0.33	0.02	0.50	0.12	0.48	0.13	0.47	0.11
GPT-4	0.49	0.04	0.39	0.16	**0.52**	0.14	0.60	−0.07
GPT-4o	0.55	0.00	0.48	0.10	0.51	0.08	**0.62**	0.01

Table 4: Task 4 Results. Accuracy of iterative mediator prediction when prompted in random order. Δ reflects the change in performance when mediators are ordered by their Mediation Influence Score (MIS).

Results. The results of this experiment are in Table 4. Results with variances are provided in Appendix B.1. In this highly complex environment with more than one node missing and with open-world search space, LLMs can still maintain their performance. Unlike the overall consistent performance of GPT-4 across all graphs, other models showed superior performance in Insurance and Alarm graphs only. As the complexity of the graph increases, we observe larger differences in hypothesizing the mediators according to the MIS order. Positive Δ values suggest that prompting the LLM based on the MIS metric leads to higher semantic similarity between the mediator hypotheses and the ground truth variables. In summary, we observe that LLMs can be effective in iteratively hypothesizing multiple mediators in a DAG, and if present, some domain knowledge about the significance of the mediator can boost the performance.

5.5 Memorization

A concern in evaluating pretrained LLMs on knowledge-intensive tasks is contamination i.e., memorization of evaluation data from training. This is especially relevant for public datasets like those in the BNLearn repository, which may have appeared in training corpora.

To assess this, we tested whether models could recall the number and names of variables from each of the eight datasets in our benchmark. This included well-known BNLearn graphs (e.g., Asia, Child, Insurance, Alarm) and less common ones (e.g., Law, Alzheimer's). We prompted each model to report node counts and variable names, including explicit references to BNLearn for relevant datasets, to detect signs of memorization.

In Table 5, except GPT family models, which exhibited partial recall for some widely known BNLearn datasets, we observe that full reconstruction

Model	Cancer	Survey	Asia	Law	Alz	Child	Insurance	Alarm
Zephyr	✗	✗	✗	✗	✗	✗	✗	✗
Mixtral	✗	✗	0.71	✗	✗	✗	✗	0.13
Neural	✗	✗	✗	✗	✗	✗	✗	✗
LLama	✗	✗	✓	✗	✗	✗	✗	✗
Mistral	✓	✗	✗	✗	✗	✗	✗	✗
GPT-4	✓	✓	✓	✗	0.55	✓	✓	✓
GPT-4o	✓	✓	✓	✗	0.45	✓	✓	✓

Table 5: Memorization analysis: Whether the model could correctly recall node information from the dataset (✓), failed to recall (✗), or proportion of nodes recalled.

of the graphs' details was rare. This recall was consistently absent for lesser-known datasets such as Law and Alzheimer's, which are less likely to have appeared during pretraining. While these findings cannot eliminate memorization with certainty, they suggest that it is not predominant for most models.

To further test GPT-4, we explicitly mentioned the graph provenance (e.g., "This graph is from BNLearn") during "Task 3", shown in Table 17. GPT-4's performance improved across most graphs. This suggests that its initial responses were not purely reciting these graphs but potentially based on broader parametric knowledge.

5.6 Discussion

The results show that LLMs effectively hypothesize missing variables, especially mediators, though performance varies with task complexity. Simple tasks, like identifying missing variables from controlled options, had high success rates.

Performance differences across domains may stem from biases in LLM training data, affecting parametric memory. For instance, confounder hypothesis quality varied across graphs, with domain-specific gaps lowering accuracy, like in the Sachs graph (Appendix D).

We explored fine-tuning and few-shot prompting to enhance performance, but small DAG sizes limited the graph size, yielding mixed results (Appendix C.1). While fine-tuning may help specialization, it can also reduce reliance on general parametric knowledge (Yang et al., 2024). Future work could explore domain-specific fine-tuning.

Though model training data is undisclosed, we used a recently released graph (Abdulaal et al., 2024) that postdates cut-off dates (at the time of performing experiments). Our novel task and verbalization approach further reduce the risks of memorization. Table 2 confirms LLMs generate novel hypotheses rather than retrieving memorized patterns, with no evidence of direct graph reconstruction. Our work relies on reasoning via parametric

knowledge rather than explicit memorization.

Our setup assumes known edges among missing variables for controlled evaluation, which future work can extend. We envision this as a human-LLM collaboration under expert supervision, as LLMs cannot self-assess plausibility or confidence (Zhou et al., 2024). Future work could also refine filtering mechanisms and improve performance on source and sink nodes.

5.7 Human Evaluation

To complement our automatic evaluation metrics, we conducted a small-scale human evaluation on three representative graphs (Cancer, Survey, and Asia). Two independent annotators (a CS PhD student and a CS PhD graduate) rated the quality of LLM-suggested variables. We then measured the agreement between human judgments, semantic similarity, and our LLM-judge using Spearman correlation.

	Correlation	p-value
Sim – LLM-judge	0.430	0.2475
Sim – R1	0.781	0.0130
Sim – R2	0.623	0.0732
LLM-judge – R1	0.622	0.0738
LLM-judge – R2	0.831	0.0055
R1 – R2	0.823	0.0065

Table 6: Spearman correlations between human annotators (R1, R2), semantic similarity (Sim), and LLM-judge scores on Cancer, Survey, and Asia graphs.

Table 6 indicates strong correlations among human annotators and between human judgments and the automatic metrics. In particular, the LLM-judge shows high alignment with both annotators, suggesting that it serves as a reliable proxy for human evaluation. This supports the use of our automatic evaluation framework as a scalable approach for benchmarking causal reasoning tasks.

6 Conclusion

Most causality research focuses on identifying relationships from observed data, while hypothesizing which variables to observe remains largely reliant on expert knowledge. We propose using LLMs as proxies for this step and introduce a novel task: hypothesizing missing variables in causal graphs. We formalize this with a benchmark that spans varying levels of difficulty and ground-truth knowledge. Our results highlight LLMs' strengths in inferring backdoor paths, including colliders, confounders,

and mediators, which often lead to biased causal inference when unaccounted for. Our work LLMs can serve as useful tools for early-stage hypothesis generation, supporting scientists in formulating plausible causal variables before data collection. By evaluating models across different graph completeness, open- and closed-world settings, we highlight their potential and limitations.

7 Limitations

While this work presents promising advancements in leveraging LLMs for hypothesizing missing variables in causal graphs, there are some limitations to consider. Our evaluation relies on established DAGs and comparisons with known ground truth, limiting assessment in scenarios without a defined baseline. Future work can include validation using human in loop evaluation. Future work can also integrate our work into the full causal discovery pipeline with statistical data.

8 Ethics and Risk

Our work leverages LLMs for hypothesis generation in causal discovery but comes with ethical risks. Biases from training data may lead to skewed hypotheses, and over-reliance on AI without expert validation could result in misleading conclusions. While we design our task to minimize memorization, risks of data leakage remain. Additionally, LLM performance varies across domains, making errors in high-stakes fields like healthcare particularly concerning. To mitigate these risks, we emphasize human-AI collaboration, transparency in model limitations, and improved evaluation frameworks for reliability.

9 Acknowledgements

References

Ahmed Abdulaal, adamos hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. 2024. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *ICLR*.

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv*.

Ahmed Alaa, Rachael V Phillips, Emre Kıcıman, Laura B Balzer, Mark van der Laan, and Maya Petersen. 2024. Large language models as co-pilots for causal inference in medical studies. *arXiv*.

Cande V Ananth and Enrique F Schisterman. 2017. Confounding, causality, and confusion: the role of intermediate variables in interpreting observational studies in obstetrics. *American journal of obstetrics and gynecology*, 217(2):167–175.

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. *arXiv*.

Ingo A Beinlich, Henri Jacques Suermondt, R Martin Chavez, and Gregory F Cooper. 1989. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89: Second European Conference on Artificial Intelligence in Medicine, London, August 29th–31st 1989. Proceedings*, pages 247–256. Springer.

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

Ruta Binkyte, Ivaxi Sheth, Zhijing Jin, Mohammad Havaei, Bernhard Schölkopf, and Mario Fritz. 2025. Causality is key to understand and balance multiple goals in trustworthy ml and foundation models. *arXiv*.

Mario Bunge. 2017. *Causality and modern science*. Routledge.

Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. *arXiv*.

Ryan Cory-Wright, Cristina Cornelio, Sanjeeb Dash, Bachir El Khadir, and Lior Horesh. 2024. Evolving scientific discovery by unifying data and background knowledge with ai hilbert. *Nature Communications*, 15(1):5922.

Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. 2024. Large language models are effective priors for causal graph discovery. *arXiv*.

Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul G Krishnan, and Chris J Maddison. 2024. End-to-end causal effect estimation from unstructured natural language data. *arXiv*.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning.

Jörg Frohberg and Frank Binder. 2021. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. *arXiv*.

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2023. Large language models are not abstract reasoners. *arXiv*.

Roxana Girju, Dan I Moldovan, et al. 2002. Text mining for causal relations. In *FLAIRS conference*, pages 360–364.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*.

Akash Gupta, Ivaxi Sheth, Vyas Raina, Mark Gales, and Mario Fritz. 2024. Llm task interference: An initial study on the impact of task-switch in conversational history. *arXiv*.

Shantanu Gupta, Zachary C Lipton, and David Childers. 2021. Estimating treatment effects with observed confounders and mediators. In *Uncertainty in Artificial Intelligence*, pages 982–991. PMLR.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2020. Causal knowledge extraction through large-scale text mining. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 13610–13611.

Rachael A Hughes, Jon Heron, Jonathan AC Sterne, and Kate Tilling. 2019. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International journal of epidemiology*, 48(4):1294–1304.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv*.

Intel. 2023. Intel neural-chat-7b model achieves top ranking on llm leaderboard!

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv*.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023. Can large language models infer causation from correlation? *arXiv*.

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

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *NeurIPS*.

Kevin B Korb and Ann E Nicholson. 2010. *Bayesian artificial intelligence*. CRC press.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *EMNLP*.

Steffen L Lauritzen and David J Spiegelhalter. 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.

Stephanie Long, Tibor Schuster, Alexandre Piché, ServiceNow Research, et al. 2023. Can large language models build causal graphs? *arXiv*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv*.

AD Nichol, M Bailey, DJ Cooper, On behalf of the PO-LAR, et al. 2010. Challenging issues in randomised controlled trials. *Injury*, 41:S20–S23.

OpenAI. 2023. Gpt-4 technical report. *arXiv*.

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

Judea Pearl. 2014. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459.

Maya Petersen, Ahmed Alaa, Emre Kıcıman, Chris Holmes, and Mark van der Laan. 2024. Artificial intelligence–based copilots to generate causal evidence.

Mohamad Amin Pourhoseingholi, Ahmad Reza Baghestani, and Mohsen Vahedi. 2012. How to control confounding effects by statistical analysis. *Gastroenterology and hepatology from bed to bench*, 5(2):79.

Akshara Prabhakar, Thomas L Griffiths, and R Thomas McCoy. 2024. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. *EMNLP Findings*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *EMNLP Findings*.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al. 2024. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. In *ICLR*.

Leonardo Ranaldi and Fabio Massimo Zanzotto. 2023. Hans, are you clever? clever hans effect analysis of neural systems. *SEM@ACL*.

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.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. 2024. Testing the general deductive reasoning capacity of large language models using ood examples. *NeurIPS*, 36.

Marco Scutari and Jean-Baptiste Denis. 2021. *Bayesian networks: with examples in R*. CRC press.

Ivaxi Sheth, Bahare Fatemi, and Mario Fritz. 2024. Causalgraph2llm: Evaluating llms for causal queries. *arXiv preprint arXiv:2410.15939*.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. Com2sense: A commonsense reasoning benchmark with complementary sentences. *arXiv*.

David J Spiegelhalter. 1992. Learning in probabilistic expert systems. *Bayesian statistics*, 4:447–465.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (llms)? aka will llms replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng. 2023. Unicausal: Unified benchmark and repository for causal text mining. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 248–262. Springer.

Jin Tian and Judea Pearl. 2012. On the testable implications of causal models with hidden variables. *arXiv*.

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

Ruibo Tu, Kun Zhang, Bo Bertilson, Hedvig Kjellstrom, and Cheng Zhang. 2019. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. *Advances in Neural Information Processing Systems*, 32.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv*.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *NeurIPS*, 36:38975–38987.

Tyler J VanderWeele and Nancy Staudt. 2011. Causal diagrams for empirical legal research: a methodology for identifying causation, avoiding bias and interpreting results. *Law, Probability & Risk*, 10(4):329–354.

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

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023a. Are large language models really good logical reasoners? a comprehensive evaluation from deductive, inductive and abductive views. *arXiv*.

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

Haoran Yang, Yumeng Zhang, Jiaqi Xu, Hongyuan Lu, Pheng-Ann Heng, and Wai Lam. 2024. Unveiling the generalization power of fine-tuned large language models. In *NAACL:HLT*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2024. Kola: Carefully benchmarking world knowledge of large language models. In *ICLR*.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. Dag-gnn: Dag structure learning with graph neural networks. In *ICML*, pages 7154–7163. PMLR.

Jiayao Zhang, Hongming Zhang, Weijie Su, and Dan Roth. 2022. Rock: Causal inference principles for reasoning about commonsense causality. In *ICML*, pages 26750–26771. PMLR.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with wikihow. *arXiv*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2023. Least-to-most prompting enables complex reasoning in large language models. *ICLR*.

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *arXiv*.

A Implementation

A.1 Datasets

We use 7 real-world based graphs. These graphs span different domain knowledge topics. These graphs have ground truth graphs along with their observational data. The simplest graph used is the cancer graph with 4 edges and 5 node variables. In addition to the semi-synthetic graphs from the BNLearn library, we also evaluate our approach on a realistic Alzheimer's Disease graph (Abdulaal et al., 2024), which was developed by five domain experts. Given that each expert created a different causal graph, the final causal DAG comprises only those edges that were agreed upon by consensus.

graph	V	E	Description
Cancer	5	4	Factors around lung cancer
Survey	6	6	Factors for choosing transportation
Asia	8	8	Factors affecting dysponea
Law	8	20	factors around legal system
Alzheimer	9	16	Factors around Alzheimer's Disease
Child	20	25	Lung related illness for a child
Insurance	27	52	Factors affecting car accident insurance
Alarm	37	46	Patient monitoring system

Table 7: graph description.

A.2 Reproducibility

For reproducibility, we used temperature 0 and top-p value as 1 across all of the models. We also mentioned the snapshot of the model used. We have also included the prompts and examples below. Our code will be released upon acceptance. The graphs are under CC BY-SA 3.0, which allows us to freely modify the graphs for benchmarking. Our benchmark will be released under the CC BY-SA License.

   GPT-4o, GPT-4 was accessed via API. The rest of the models were run on 1 A100 GPU. Since we used an off-the-shelf LLM, there was no training to be performed. Since many of the models were run by API, it is difficult to calculate the entire computation, however, all of the experiments for each model took ≈ 6 hours.

A.3 Controlled Variable Identification

For variable identification, we generate multiple choices that remain consistent across all missing nodes and all of the graphs. The words were randomly chosen to be far enough from the nodes. The options chosen were weather, book sales, and movie ratings. We wanted to make sure that the options were not from one specific domain, such that the LLM could do the process of elimination.

A.4 Causal effect

Average Treatment Effect.   Average Treatment Effect (ATE) quantifies the expected change in the outcome v_y caused by the unit change of the treatment v_t. ATE is a part of the causal do-calculus introduced by (Pearl, 2009). We consider binary causal DAGs, i.e., each variable can either take 0 or 1 as values.

$$\text{ATE} = \mathbb{E}[v_y|\text{do}(v_t = 1)] - \mathbb{E}[v_y|\text{do}(v_t = 0)]$$

where the $\text{do}(\cdot)$ operator, represents an intervention. The $E[v_y|\text{do}(v_t = 1)]$ represents the expected value of the outcome variable v_y when we intervene to set the treatment variable v_t to 1 (i.e., apply the treatment), and $E[v_y|\text{do}(v_t = 0)]$ represents the expected value of v_y when we set v_t to 0 (i.e., do not apply the treatment).

Mediation Analysis.   Mediation analysis is implemented to quantify the effect of a treatment on the outcome via a third variable, the mediator. The total mediation effect can be decomposed into the Natural Direct Effect (NDE) and the Natural Indirect Effect (NIE). The Natural Direct Effect (NDE) is the effect of the treatment on the outcome variable when not mediated by the mediator variable. The Natural Indirect

Effect (NIE) is the effect of the treatment variable on the outcome variable when mediated by the mediator variable.

$$\text{NDE} = \mathbb{E}[v_{t=1}, v_{m=0} - v_{t=0}, v_{m=0}]$$

Here, NDE is calculated by comparing the expected outcome when the treatment variable is set to 1 and the mediator is fixed at the level it would take under the control treatment $v_t = 0$, with the expected outcome when both the treatment and the mediator are set to the control level.

$$\text{NIE} = \mathbb{E}[v_{t=0}, v_{m=1} - v_{t=0}, v_{m=0}]$$

Here, NIE is calculated by comparing the expected outcome when the treatment variable is set to 1 and the mediator is allowed to change as it would under the treatment, with the expected outcome when the treatment variable is set to 1 but the mediator is fixed at the control level.

A.5 Semantic Similarity

Given the task of hypothesizing missing nodes in a partial graph $\mathcal{G}^*$ in the absence of multiple-choices, we evaluate the semantic similarity between the model's predictions and the ground truth node variable. We leverage an open model namely 'all-mpnet-base-v2' to transform the textual representations of the model's predictions and the ground truth into high-dimensional vector space embeddings. Post transforming textual representations into embeddings and normalizing them, we calculate the cosine similarity. Scores closer to 1 indicate a high semantic similarity, suggesting the model's predictions align well with the ground truth. This metric gives a score of similarity without the contextual knowledge of the causal graph. We perform our experiments to consider every node of the ground truth as a missing node iteratively. For all the suggestions for a node variable, we calculate the semantic similarity. The average similarity reported is the highest semantic similarity for each of the variable suggestions.

Algorithm 1 Evaluating Semantic Similarity for Hypothesized Missing Nodes

---

1: **Input:** Partial graph $\mathcal{G}^*$, Ground truth node variables V_{GT}, Language model $LM =$ 'all-mpnet-base-v2'
2: **Output:** Average highest semantic similarity score
3: **procedure** SEMANTICSIMILARITY($\mathcal{G}^*, V_{\text{GT}}, LM$)
4:     Initialize $similarityScores$ as an empty list
5:     **for** each node v_{GT} in **v do**
6:         $predictions \leftarrow$ GeneratePredictions($\mathcal{G}^*, LM$)
7:         Initialize $nodeScores$ as an empty list
8:         **for** each prediction p in $predictions$ **do**
9:             $embedding_{\text{GT}} \leftarrow$ Embed(v_{GT}, LM)
10:            $embedding_p \leftarrow$ Embed(p, LM)
11:            Normalize $embedding_{\text{GT}}$ and $embedding_p$
12:            $score \leftarrow$ CosineSimilarity($embedding_{\text{GT}}, embedding_p$)
13:            Append $score$ to $nodeScores$
14:         **end for**
15:         $maxScore \leftarrow$ Max($nodeScores$)
16:         Append $maxScore$ to $similarityScores$
17:     **end for**
18:     $averageScore \leftarrow$ Average($similarityScores$)
19:     **return** $averageScore$
20: **end procedure**

Ground Truth:	Smoking status				
LLM Suggestions:	Smoking	Alcohol Consumption	Exposure to Radiation	Poor Diet	Genetic Predisposition
Semantic similarity :	0.72	0.38	0.22	0.22	0.17
Ground Truth:	Employee or self-employed				
LLM Suggestions:	Income Level	Job Location	Environmental Awareness	Lifestyle Preferences	Health Consciousness
Semantic similarity :	0.30	0.25	0.17	0.15	0.10
Ground Truth:	Dyspnea laboured breathing				
LLM Suggestions:	Shortness of breath	Chest Pain	Coughing	Fatigue	Weight Loss
Semantic similarity :	0.57	0.41	0.36	0.29	0.11
Ground Truth:	Montreal Cognitive Assessment score				
LLM Suggestions:	Cognitive Function	Neurological Function	Mental Health Status	Risk of Alzheimer's Disease	Memory Performance
Semantic similarity :	0.60	0.47	0.38	0.36	0.16
Ground Truth:	Grunting in infants				
LLM Suggestions:	Respiratory distress	Asthma	Pneumonia	Pulmonary infection	Bronchopulmonary dysplasia (BPD)
Semantic similarity :	0.22	0.18	0.17	0.11	0.01
Ground Truth:	Driving history				
LLM Suggestions:	Previous accidents	Distance driven daily	Type of car insurance	Frequency of car maintenance	Location of parking
Semantic similarity :	0.55	0.42	0.27	0.26	0.18
Ground Truth:	Heart rate blood pressure				
LLM Suggestions:	Pulse Rate	Blood Pressure	Respiratory Rate	EKG Reading	Blood Oxygen Level
Semantic similarity :	0.78	0.78	0.57	0.49	0.42

Table 8: Examples of model suggestions from and the corresponding semantic similarity score for a missing node variable from each of the graphs.

A.6 LLM-as-Judge

To capture the domain knowledge of the expert that selects the most relevant causal variable, we use LLM-as-Judge as a proxy expert. This also allows for evaluation based on contextual DAG knowledge as well. Given the impressive results of GPT-4 in (Zheng et al., 2023), we use GPT-4 as a judge for all of the experiments.

Algorithm 2 Evaluating Model Suggestions with LLM as Judge

1: **Input:** Partial graph $\mathcal{G}^*$, Ground truth node variables V_{GT}, Predictions P, Language model LLM = GPT-4
2: **Output:** Average quality rating of model's suggestions
3: **procedure** LLMASJUDGE($\mathcal{G}^*, V_{\text{GT}}, P$, LLM)
4:     Initialize $qualityRatings$ as an empty list
5:     **for** each node v_{GT} in **V do**
6:         $suggestions \leftarrow$ GenerateSuggestions($\mathcal{G}^*, P$, LLM)
7:         $bestSuggestion \leftarrow$ SelectBestSuggestion($suggestions, v_{\text{GT}}$, LLM)
8:         $rating \leftarrow$ RateSuggestion($bestSuggestion$, LLM)
9:         Append $rating$ to $qualityRatings$
10:     **end for**
11:     $averageRating \leftarrow$ Average($qualityRatings$)
12:     **return** $averageRating$
13: **end procedure**
14: **function** GENERATESUGGESTIONS($\mathcal{G}^*, P$, LLM)
15:     **return** A set of suggestions for missing nodes based on P
16: **end function**
17: **function** SELECTBESTSUGGESTION($suggestions, v_{\text{GT}}$,LLM)
18:     Prompt LLM with $\mathcal{G}^*$, v_{GT}, and $suggestions$
19:     **return** LLM's choice of the best fitting suggestion
20: **end function**
21: **function** RATESUGGESTION($suggestion, LM$)
22:     Prompt LLM to rate $suggestion$ on a scale of 1 to 10
23:     **return** LLM's rating
24: **end function**

Ground Truth:	Education up to high school or university degree
Top ranked suggestion:	Education level
Rating :	9.5
Ground Truth:	Pollution
Top ranked suggestion:	Smoking history
Rating :	2.0
Ground Truth:	Bonchitis
Top ranked suggestion:	smoking behavior
Rating :	2.0
Ground Truth:	Lung XRay report
Top ranked suggestion:	Lung Damage
Rating :	8.0
Ground Truth:	Socioeconomic status
Top ranked suggestion:	Driver's lifestyle
Rating :	7.0

Table 9: Examples of model suggestions from and the corresponding LLM-as-judge score for a missing node variable.

Ground Truth: Dyspnea laboured breathing
LLM Suggestion: Shortness of breath
Semantic similarity to GT: 0.57
LLM-as-Judge score: 9.5

Table 10: Example comparing the semantic similarity and LLM-as-Judge metrics. Dyspnea is a medical term for shortness of breath. In this example, the contextual information, beyond exact matching, is better captured by LLM-as-Judge.

Shortcomings of LLM-as-judge. LLM-as-judge uses GPT-4 as a judge model which could be biased towards some data. Since the training graphs are not public for this model, it would be hard to judge how these biases might affect the final score. Hence for robust evaluation we also evaluate using the semantic similarity.

A.7 Iteratively Hypothesizing in Open World

For each order, the algorithm prompts the LLM to generate mediator suggestions, selects the suggestion with the highest semantic similarity to the context, and iteratively updates the partial graph with these mediators. Δ, quantifies the impact of mediator ordering by comparing the average highest semantic similarity scores obtained from both descending and ascending orders. This methodical evaluation sheds light on how the sequence in which mediators are considered might affect the LLM's ability to generate contextually relevant and accurate predictions.

Algorithm 3 Random Order Mediator Hypothesis

1: **Input:** Partial graph $\mathcal{G}^*$ (where $\mathcal{G}^* = \mathcal{G} - H$), Treatment v_t, Outcome v_y, Number of mediators H, Number of suggestions k
2: **Output:** Updated graph $\mathcal{G}^*$ with selected mediators
3: **procedure** GENERATEMEDIATORSRANDOM($\mathcal{G}^*, v_t, v_y, H, k$)
4:     **for** $i \leftarrow 1$ to H **do**
5:         $suggestions \leftarrow$ Generate k suggestions for v_{m_i} using $P_{\text{LLM}}(\mathcal{G}^*)$
6:         Initialize $highestSimilarity \leftarrow 0$
7:         Initialize $selectedMediator \leftarrow$ null
8:         **for** each $suggestion$ in $suggestions$ **do**
9:             $similarityScore \leftarrow$ Calculate semantic similarity for $suggestion$
10:             **if** $similarityScore > highestSimilarity$ **then**
11:                 $highestSimilarity \leftarrow similarityScore$
12:                 $selectedMediator \leftarrow suggestion$
13:             **end if**
14:         **end for**
15:         Update $\mathcal{G}^* \leftarrow \mathcal{G}^* \cup \{selectedMediator\}$
16:     **end for**
17:     **return** $\mathcal{G}^*$
18: **end procedure**

Algorithm 4 Ordered Mediator Generation and Evaluation Based on MIS

1: **Input:** Partial graph $\mathcal{G}^*$, Treatment v_t, Outcome v_y, Set of potential mediators M, Number of suggestions k

2: **Output:** Δ - measure of the influence of mediator ordering

3: **procedure** CALCULATEMIS(v_t, v_y, M)

4:     Initialize MISList as an empty list

5:     **for** each mediator v_{m_i} in M **do**

6:         Calculate NIE(v_{m_i}) and NDE(v_{m_i})

7:         MIS(v_{m_i}) $\leftarrow \frac{\text{NIE}(v_{m_i})}{\text{NDE}(v_{m_i})}$

8:         Append MIS(v_{m_i}) to MISList

9:     **end for**

10:     **return** MISList

11: **end procedure**

12: **procedure** GENERATEMEDIATORSORDERED($\mathcal{G}^*, v_t, v_y, M, k$)

13:     MISList $\leftarrow$ CALCULATEMIS(v_t, v_y, M)

14:     Sort M in descending order of MISList to get M_{desc}

15:     Sort M in ascending order of MISList to get M_{asc}

16:     $averageDesc \leftarrow$ GENERATEANDEVALUATE($\mathcal{G}^*, M_{\text{desc}}, k$)

17:     $averageAsc \leftarrow$ GENERATEANDEVALUATE($\mathcal{G}^*, M_{\text{asc}}, k$)

18:     $\Delta \leftarrow \frac{|averageDesc - averageAsc|}{averageDesc}$

19:     **return** Δ

20: **end procedure**

21: **function** GENERATEANDEVALUATE($\mathcal{G}^*, M_{\text{order}}, k$)

22:     Initialize similarityScores as an empty list

23:     **for** each mediator v_{m_i} in M_{order} **do**

24:         Perform the same steps as in the refined random order mediator generation

25:         (Generate k suggestions, select the most similar, update $\mathcal{G}^*$)

26:         Append the highest similarity score to similarityScores

27:     **end for**

28:     **return** Average of similarityScores

29: **end function**

B Further results

B.1 Variances

For brevity we didnt add variance in the main text, the following results have variances:

	Cancer		Survey		Asia		Alzheimers		Child		Insurance		Alarm		Avg	
	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J
Zephyr	0.36 ±0.04	0.61 ±0.06	0.34 ±0.07	0.60 ±0.05	0.45 ±0.05	0.66 ±0.04	0.35 ±0.03	0.75 ±0.03	0.51 ±0.02	0.70 ±0.04	0.45 ±0.04	0.44 ±0.05	0.46 ±0.03	0.69 ±0.02	0.42 ±0.04	0.63 ±0.04
Mixtral	0.41 ±0.03	0.66 ±0.04	0.39 ±0.05	0.66 ±0.06	**0.66** ±0.02	0.75 ±0.03	0.31 ±0.04	0.77 ±0.02	**0.53** ±0.03	**0.77** ±0.02	0.46 ±0.03	**0.56** ±0.04	**0.50** ±0.03	0.72 ±0.06	0.46 ±0.03	0.70 ±0.05
Neural	0.38 ±0.02	0.77 ±0.05	0.43 ±0.02	0.55 ±0.03	0.53 ±0.03	0.55 ±0.04	0.44 ±0.05	0.71 ±0.03	0.48 ±0.04	0.70 ±0.03	0.47 ±0.04	0.43 ±0.05	0.47 ±0.02	0.67 ±0.03	0.45 ±0.03	0.63 ±0.04
Llama	0.40 ±0.05	0.48 ±0.04	0.40 ±0.05	0.54 ±0.05	0.53 ±0.03	0.58 ±0.06	0.45 ±0.04	0.61 ±0.04	0.48 ±0.04	0.63 ±0.03	0.42 ±0.01	0.34 ±0.05	0.46 ±0.02	0.65 ±0.03	0.45 ±0.03	0.55 ±0.04
Mistral	0.33 ±0.01	0.67 ±0.05	0.44 ±0.05	0.65 ±0.04	0.60 ±0.03	0.73 ±0.04	0.34 ±0.04	0.76 ±0.02	0.48 ±0.04	0.68 ±0.03	0.46 ±0.03	0.47 ±0.01	0.47 ±0.03	0.71 ±0.03	0.44 ±0.03	0.67 ±0.03
GPT-4	**0.49** ±0.02	**0.90** ±0.03	**0.51** ±0.06	0.67 ±0.04	**0.66** ±0.02	**0.76** ±0.03	**0.47** ±0.02	0.98 ±0.02	0.36 ±0.05	0.53 ±0.04	**0.52** ±0.03	**0.56** ±0.03	0.49 ±0.06	**0.75** ±0.02	**0.50** ±0.04	**0.73** ±0.03

Table 11: Average semantic similarity and LLM-as-Judge metrics to evaluate LLMs in hypothesizing the missing variable in a causal DAG.

	Asia		Child		Insurance		Alarm	
	Sim	Δ	Sim	Δ	Sim	Δ	Sim	Δ
Zephyr	0.61 ±0.03	−0.02 ±0.01	**0.54** ±0.04	0.17 ±0.02	0.47 ±0.05	0.19 ±0.02	0.51 ±0.05	0.20 ±0.02
Mixtral	**0.87** ±0.02	0.01 ±0.01	0.50 ±0.05	0.18 ±0.02	0.48 ±0.05	0.15 ±0.02	0.52 ±0.05	0.13 ±0.01
Neural	0.65 ±0.06	0.04 ±0.02	0.48 ±0.04	0.21 ±0.02	0.42 ±0.04	0.16 ±0.02	0.46 ±0.04	0.12 ±0.01
Llama	0.80 ±0.08	0.07 ±0.02	0.49 ±0.05	−0.05 ±0.01	0.44 ±0.06	0.21 ±0.02	0.51 ±0.05	0.07 ±0.01
Mistral	0.33 ±0.03	0.02 ±0.01	0.50 ±0.05	0.12 ±0.02	0.48 ±0.04	0.13 ±0.02	0.47 ±0.04	0.11 ±0.01
GPT-4	0.49 ±0.07	0.04 ±0.01	0.39 ±0.05	0.16 ±0.02	**0.52** ±0.05	0.14 ±0.02	**0.60** ±0.06	−0.07 ±0.01

Table 12: Sim: semantic similarity for iteratively hypothesizing the mediator nodes when prompted with random order. Δ measures the change in the prediction of each model according to the MIS.

B.2 Breaking down the performance

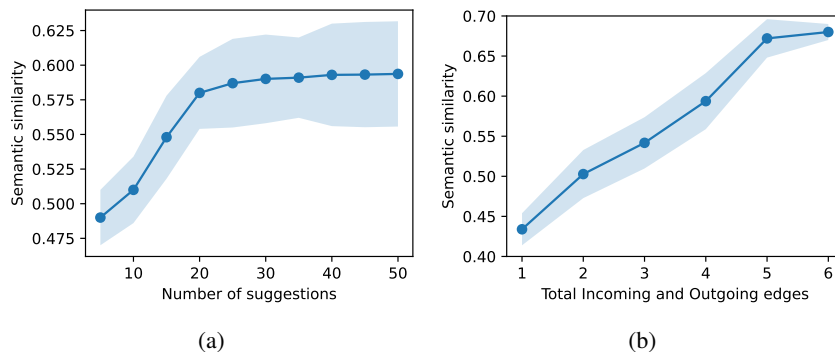(a)   (b)

Figure 6: L: Plot of semantic similarity with an increasing number of suggestions for GPT-4 on the Alarm graph. R: Plot of semantic similarity against the total number of incoming and outgoing edges for GPT-4 on the Alarm graph.

B.3 Effect of context

We observed notable differences in the accuracy of LLM predictions for missing nodes within causal graphs when context was provided versus when it was absent. Specifically, the inclusion of contextual information about the causal graph significantly enhanced the LMs' ability to generate accurate and relevant predictions. In realistic settings, when this setup is being used by a scientist, they would provide

the context of the task along with the partial graph. When context was not provided, the models often struggled to identify the most appropriate variables, leading to a decrease in prediction accuracy, especially for smaller models. Unsurprisingly, providing context was more important for smaller graphs than larger graphs. LLMs were able to understand the context of the graph via multiple other nodes in the graph for larger graphs.

	Cancer		Survey		Asia		Insurance		Alarm	
	X	✓	X	✓	X	✓	X	✓	X	✓
In-Context	0.75	1.00	0.67	1.00	0.68	0.88	0.85	0.90	0.96	0.96
Out-of-Context	0.00	0.25	0.33	0.33	0.53	0.61	0.58	0.58	0.60	0.57
Open world Hypothesis	0.39	0.41	0.40	0.39	0.63	0.66	0.49	0.50	0.44	0.46

Table 13: Model-Mixtral to evaluate the effect of context given in the prompt.

B.4 Using explanations

While using LLMs for hypothesizing the missing nodes within the causal graph for the open world setting, an additional question is for the model to provide explanations for each of its predictions. This was motivated by the fact that incorporating a rationale behind each prediction might enhance the model's semantic similarity. We present the results in the Table below. We observe that evaluating semantic similarity with explanations leads to a decrease in performance as compared to the earlier setting where the language model returned phrases. This is because semantic similarity, as a metric, evaluates the closeness of the model's predictions to the ground truth in a high-dimensional vector space, focusing on the semantic content encapsulated within the embeddings. It is a metric that leaves little room for interpretative flexibility, focusing strictly on the degree of semantic congruence between the predicted and actual variables. The introduction of explanations, while enriching the model's outputs with contextual insights, did not translate into improved semantic alignment with the ground truth.

	Cancer		Survey		Asia		Insurance		Alarm	
	X	✓	X	✓	X	✓	X	✓	X	✓
Sim	0.49 ±0.02	0.38 ±0.07	0.51 ±0.06	0.44 ±0.10	0.66 ±0.02	0.57 ±0.09	0.52 ±0.03	0.40 ±0.07	0.49 ±0.06	0.40 ±0.06
LLM-Judge	0.90 ±0.03	0.91 ±0.02	0.67 ±0.04	0.69 ±0.02	0.76 ±0.03	0.76 ±0.04	0.56 ±0.03	0.55 ±0.03	0.75 ±0.02	0.75 ±0.02

Table 14: Model-GPT 4. Evaluating the effect of explanations on different metrics from Task 3.

Ambiguous predictions which semantically represent the same variable. An important linguistic concern that could be missed by semantic similarity is an ambiguous hypothesis by the LLM that may have the same semantics, which again breaks the semantic similarity metric. This further motivates the LLM-judge metric, whose input is the context of the causal graph, the partial causal graph, the ground truth variable, and the model predictions. Given the rich context of the LLM-judge metric, we suspect it would be able to overcome the ambiguity. We prompted the model to justify its hypothesis variables using explanations. We observe that evaluating semantic similarity with explanations leads to a decrease in performance as compared to the earlier setting where the language model returned just phrases. In Table 14, we observed a drop in performance for semantic similarity. In contrast, we observe a similar or slight improvement in the LLM-judge metric when the explanation of the model hypothesis is given.

B.5 Chain of thought

Chain-of-Thought prompting has gained popularity due to its impressive performance in proving the quality of LLMs' output (Kojima et al., 2022), also in metadata-based causal reasoning (Vashishtha et al., 2023). We also incorporated COT prompting for our prompts. We perform ablation studies in Table 15. We observe that COT particularly improves the performance of the identification experiments.

	Cancer		Survey		Asia		Insurance		Alarm	
	X	✓	X	✓	X	✓	X	✓	X	✓
In-Context	1.00	1.00	0.83	1.00	0.75	0.88	0.74	0.90	0.91	0.96
Out-of-Context	0.50	0.25	0.18	0.33	0.57	0.61	0.56	0.58	0.54	0.57

Table 15: Model-Mixtral to evaluate the effect of COT given in the prompt.

B.6 Iterative mediator search vs all at once

For Task 4, we iteratively hypothesize the missing variables (mediators). Our choice was primarily driven by the complexity of Task 4, which involves predicting multiple missing mediators, ranging from 1 to 10. For a Task with 10 missing mediators, the model would have to predict 50 suggestions at once. We initially hypothesized that LLMs might struggle with making multiple predictions across different variables simultaneously. This was indeed reflected in our results and GPT-4 outputs from Table X. The iterative approach allows the model's prediction to narrow the search space, which would not be possible in a non-iterative approach. This method is more aligned with the scientific discovery process, where hypotheses are often refined iteratively based on new findings. Furthermore, our approach simulates a human-in-the-loop scenario, where the most plausible answer is selected and used to guide the next prediction.

	Asia	Child	Insurance	Alarm
Non-iterative	0.42 +- 0.07	0.33 +- 0.06	0.45 +- 0.09	0.54 +- 0.05
Iterative	0.49 +- 0.05	0.39 +-0.03	0.52 +- 0.02	0.60 +- 0.04

B.7 Results on Neuropathic graph

We added a new graph, the neuropathic pain graph (Tu et al., 2019), which is not part of common LLM training corpora as one needs to use a python script to download it. The graph consists of 221 nodes and 770 edges, but for feasibility, we selected a subset of the graph for evaluation. We ran experiments for Task 1, Task 2, and Task 3.

Model	Task 1	Task 2 Result	Task 2 FNA	Task 3 Sim	Task 3 LLM-J
Mistral	0.64	0.51	0.32	0.38	0.53
Mixtral	0.83	0.55	0.34	0.45	0.69
Llama	0.78	0.49	0.27	0.44	0.63
GPT-4	0.94	0.68	0.24	0.51	0.76

Table 16: Comparison of model performances across tasks on Neuropathic graph.
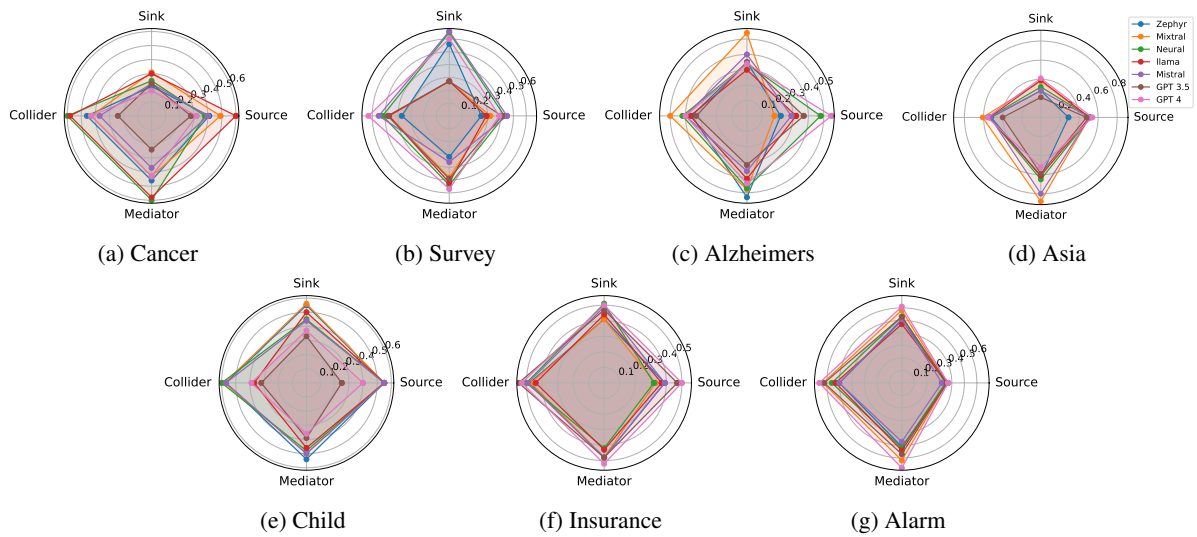
B.8 Fine grained model performance

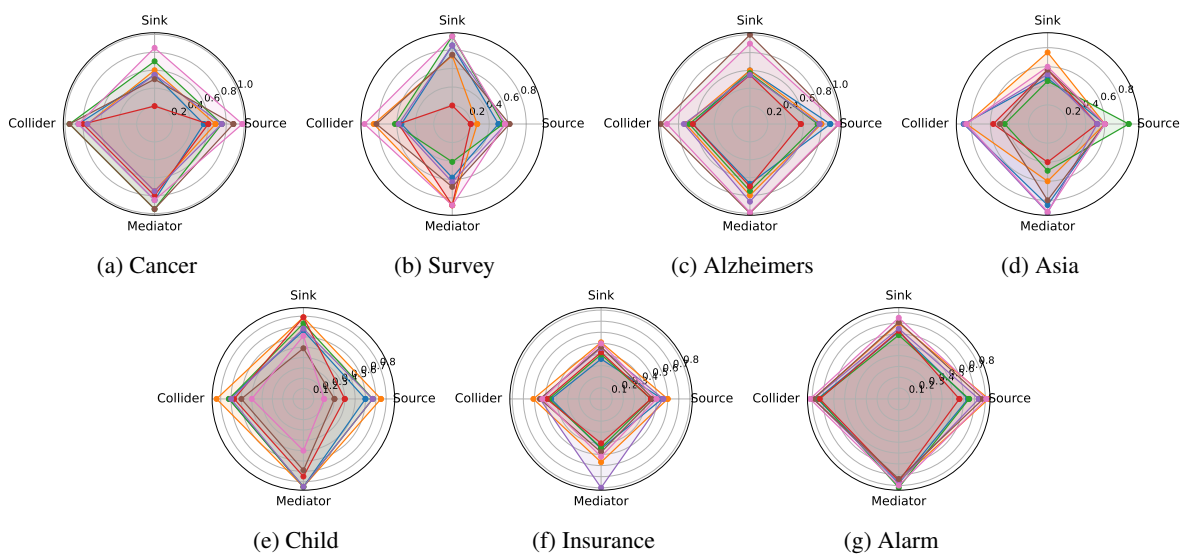Figure 7: Detailed spider plots for Semantic similarity

Figure 8: Detailed spider plots for LLM-as-judge metric

B.9 Testing whether LLM is using context or parametric graphs for GPT*

Dataset	Current Sim	Memorization	Est. Sim w/ BNLearn Context
Cancer	0.49	✓	0.60
Survey	0.51	✓	0.62
Asia	0.66	✓	0.78
Law	0.55	✗	0.56
Alzheimers	0.47	0.55	0.52
Child	0.36	✓	0.52
Insurance	0.52	✓	0.64
Alarm	0.49	✓	0.62

Table 17: Estimated similarity improvement for GPT-4 when informed that graphs are from the BNLearn repository. The memorization column shows whether GPT-4 recalled structural details.

To test whether GPT-4's original performance was driven by the retrieval of memorized content, we reran the variable inference task with explicit prompts stating that the graphs were from the BNLearn repository. We observed modest gains in similarity for well-known graphs (e.g., Asia, Alarm), indicating that GPT-4 can retrieve additional details when cued. However, the performance in the original setting, without such cues, was already strong, suggesting that the model was not merely retrieving memorized structures. Instead, its responses appear to reflect contextual reasoning and generalization beyond rote recall.

B.10 Converting causal graph to prompt

We observe that different graph representations yield similar performance across tasks, with the most variation for Task 2 where we have 2 missing variables on Mistral and Mixtral models.

Model	Asia	Child	Insurance	Alarm
JSON	0.80	0.79	0.50	0.85
GraphML	0.80	0.78	0.47	0.85
Textual (ours)	0.78	0.80	0.49	0.85

Table 18: Different encoding strategies for Task 1

Model	Asia	Child	Insurance	Alarm
JSON	0.73/0.16	0.45/0.30	0.37/0.17	0.50/0.21
GraphML	0.70/0.15	0.41/0.29	0.37/0.12	0.53/0.22
Textual (ours)	0.73/0.17	0.42/0.31	0.40/0.12	0.51/0.22

Table 19: Different encoding strategies for Task 2 (Acc/FNA)

Model	Asia	Child	Insurance	Alarm
JSON	0.75	0.67	0.46	0.69
GraphML	0.71	0.67	0.50	0.72
Textual (ours)	0.73	0.68	0.47	0.71

Table 20: Different encoding strategies for Task 3 (LLM-J)

C Finetuning and Few-shot prompting

C.1 Finetuning

We aim to assess the LLM's causal reasoning via prompting. The following are the reasons why fine-tuning is not the most practical solution:

- Pretrained models come with a wealth of general knowledge, which we aim to leverage. Fine-tuning these models could potentially limit their ability to draw on this broad knowledge base. We aim to understand the utility of pretrained models, as fine-tuning large models like GPT-4 is not always feasible.

- The training graph is too small for fine-tuning. Despite considering a large 52-edged graph: Insurance, we would have just 27 datapoints or Alarm with 37 datapoint. Additionally:

1. Using the same graph as part of train and test would unfortunately lead to training data leakage.

2. If we consider different graphs for train and test, there would exist a domain shift in the two graphs and the model may be overfitted to the domain of the train graph.

However, to illustrate our hypothesis and alleviate the reviewer's concern, we performed Supervised Fine-Tuning using QLoRA on the Mistral-7b-Instruct model for hypothesizing in the open world task. The train set here is all of the graphs minus the respective graph it was tested on. We tested on Survey, Insurance and Alzheimers graphs. The model was trained to give one best-fit suggestion for the missing variable.

	Insurance	Survey	Alzheimers
No fine-tuning	0.42 +- 0.03	0.44 +- 0.05	0.34 +- 0.04
Fine-tuned	0.39 +- 0.04	0.39 +- 0.03	0.36 +- 0.07

Table 21: Finetuning results.

From the above results, it is evident that finetuning does not significantly improve over the prompting results. This is because during training the LLM gets biased towards the domains of training graphs which are contextually distant from the test domain, given the diversity of graphs chosen. One may think that training might help the LLM to understand the task, but from prompt-based model output, it was evident that the LLM can instruction-follow. In summary, we were able to extract the LLM knowledge via prompting and domain-specific fine-tuning could be closely looked at in the future works.

C.2 Fewshot prompting

Similar to fine-tuning, few-shot learning's success depends on balancing domain specificity and generality. To avoid test examples becoming part of the shots, we have to use different domains as examples. Given the complexity of the Alarm graph, we decided to use them as a prior. We performed experiments with 1-shot and 5-shots for the Mixtral 8x7b model. We would like to remind you that Alarm was a medical graph which means that providing more examples in a different domain might hinder the model performance. Drop in performance when changing domain for in-context learning has been discussed in (Kwan et al., 2024) and (Gupta et al., 2024).

graph	0-shot	1-shot	5-shot
Cancer	0.41	0.43	0.46
Survey	0.39	0.38	0.36
Asia	0.66	0.70	0.72
Alzheimer's	0.31	0.33	0.34
Child	0.53	0.55	0.56
Insurance	0.46	0.42	0.45

Table 22: Fewshot prompting results.

D Confounders

	Sachs	Alarm1	Alarm2	Ins1	Ins2	Ins3	Ins4	Ins5	Ins6	Ins7
Zephyr	0.12	0.37	0.29	0.45	0.49	0.37	0.29	0.33	0.46	0.73
Mixtral	0.89	0.54	0.57	0.57	1.0	0.32	0.23	0.38	0.28	1.0
Neural	0.34	0.27	0.28	0.42	0.47	0.34	0.48	0.48	0.38	0.48
LLama	0.27	0.39	0.44	0.55	1.0	0.29	0.22	0.57	0.45	1.0
Mistral	0.23	0.62	0.46	0.58	1.0	0.28	0.28	0.28	0.28	1.0
GPT-4	0.91	0.49	0.44	0.62	0.39	0.58	0.44	0.58	0.52	1.0

Table 23: Semantic similarity

	Sachs	Alarm1	Alarm2	Ins1	Ins2	Ins3	Ins4	Ins5	Ins6	Ins7
Zephyr	0.10	0.40	0.30	0.45	0.60	0.40	0.40	0.30	0.70	0.80
Mixtral	0.95	0.70	1.0	0.75	1.0	0.80	0.20	0.20	0.20	1.0
Neural	0.30	0.60	0.30	1.0	0.60	0.30	0.80	0.30	0.40	0.60
LLama	0.20	0.50	0.44	0.40	1.0	0.50	0.20	0.70	0.45	1.0
Mistral	0.20	0.90	0.80	0.55	1.0	0.30	0.20	0.70	0.30	1.0
GPT-4	0.95	0.65	0.80	0.60	0.70	0.80	0.85	0.80	0.75	1.0

Table 24: LLM judge
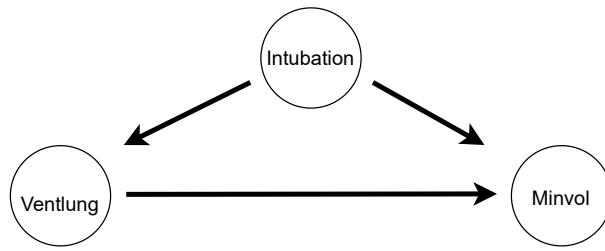
Figure 9: Alarm 1
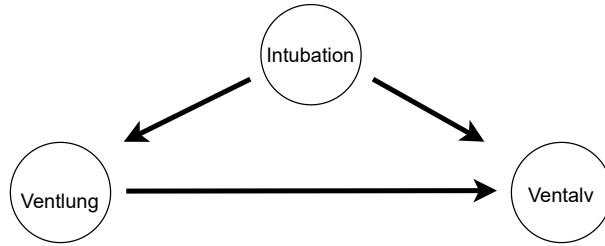
Figure 10: Alarm 2

E   Causal graphs
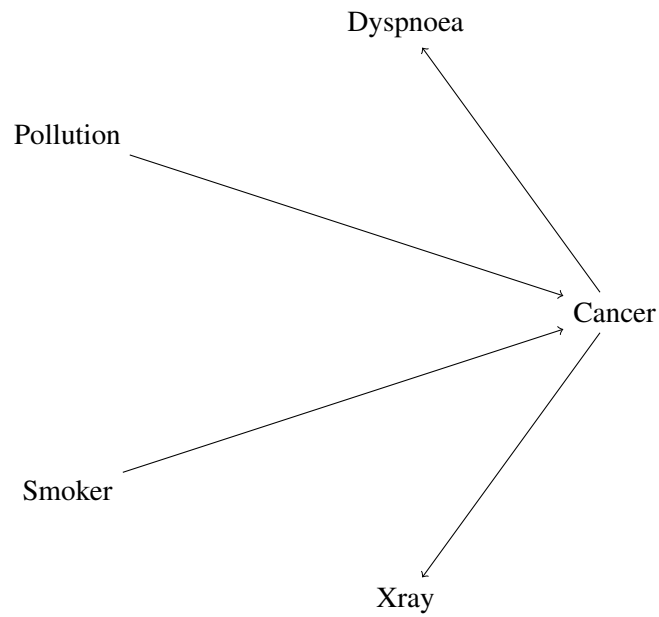
Figure 18: Cancer DAG
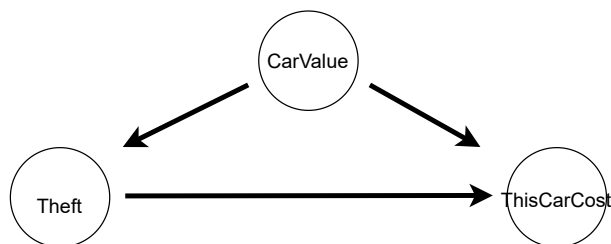
Figure 11: Insurance 1
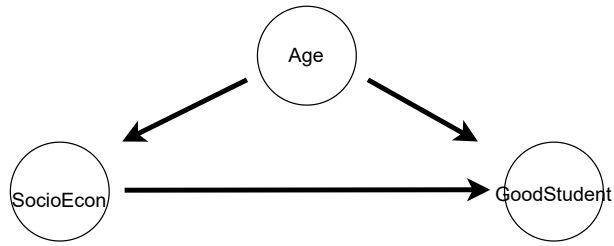
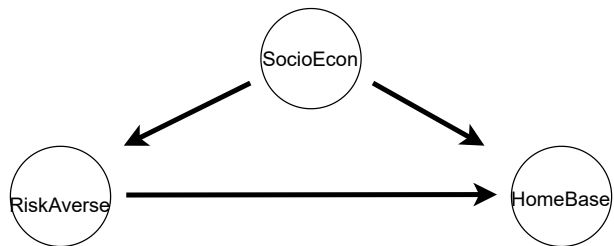Figure 12: Insurance 2
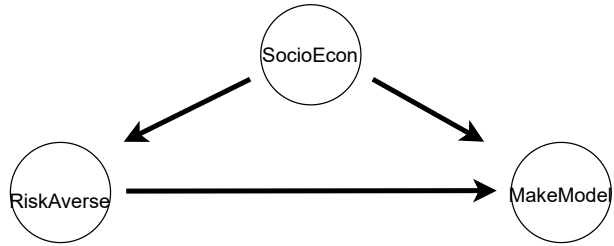
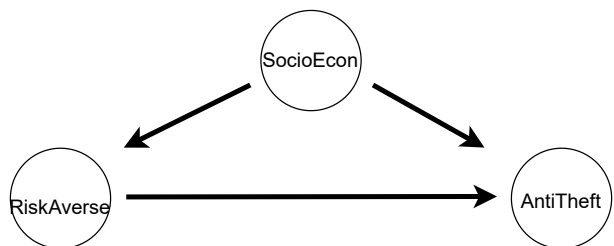Figure 13: Insurance 3
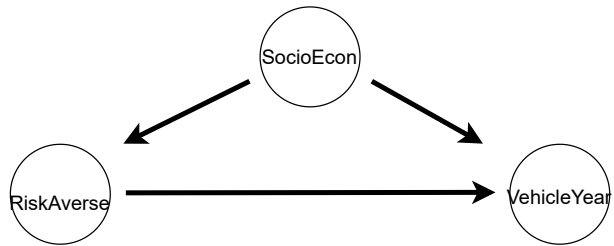
Figure 14: Insurance 4

Figure 15: Insurance 5

Figure 16: Insurance 6
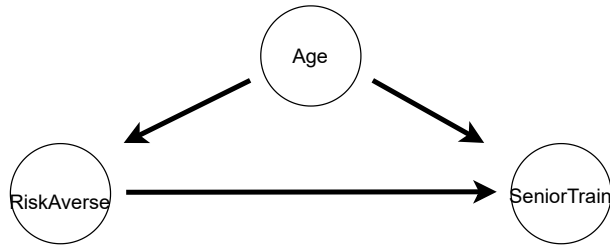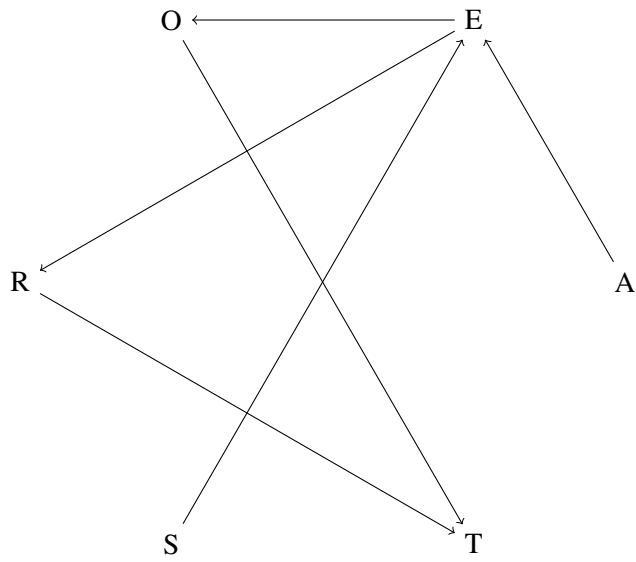
Figure 17: Insurance 7
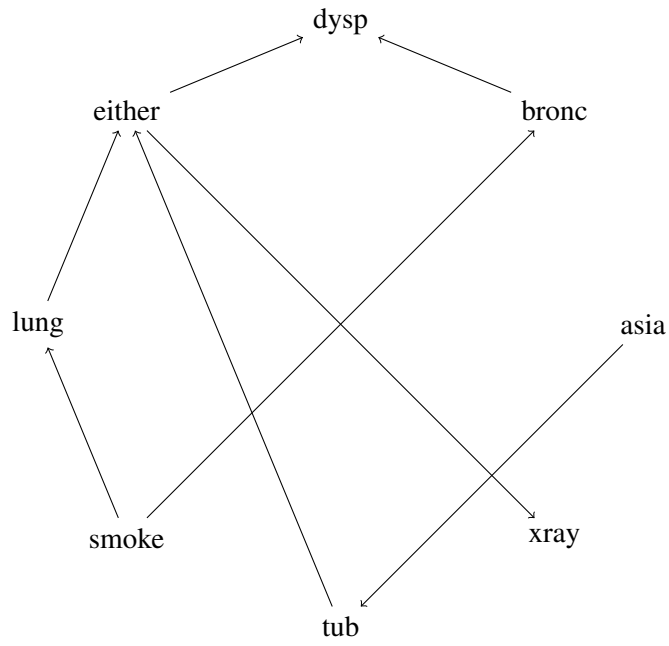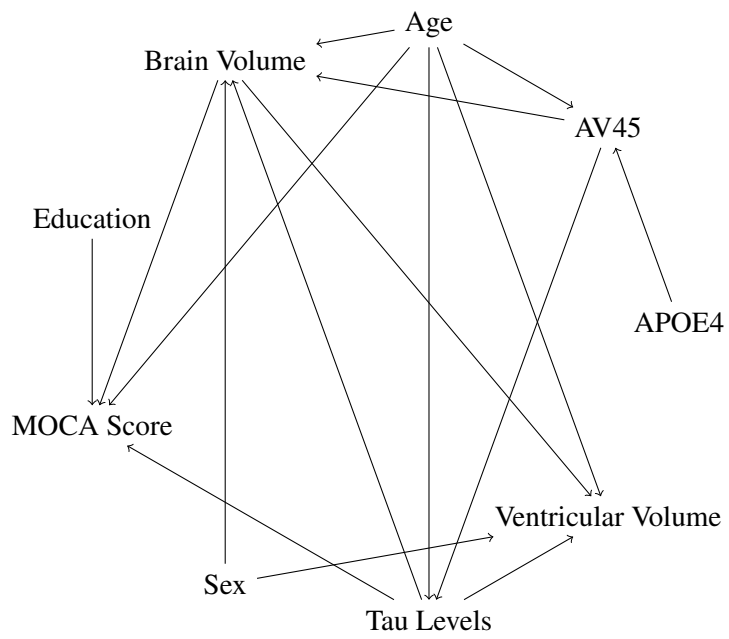

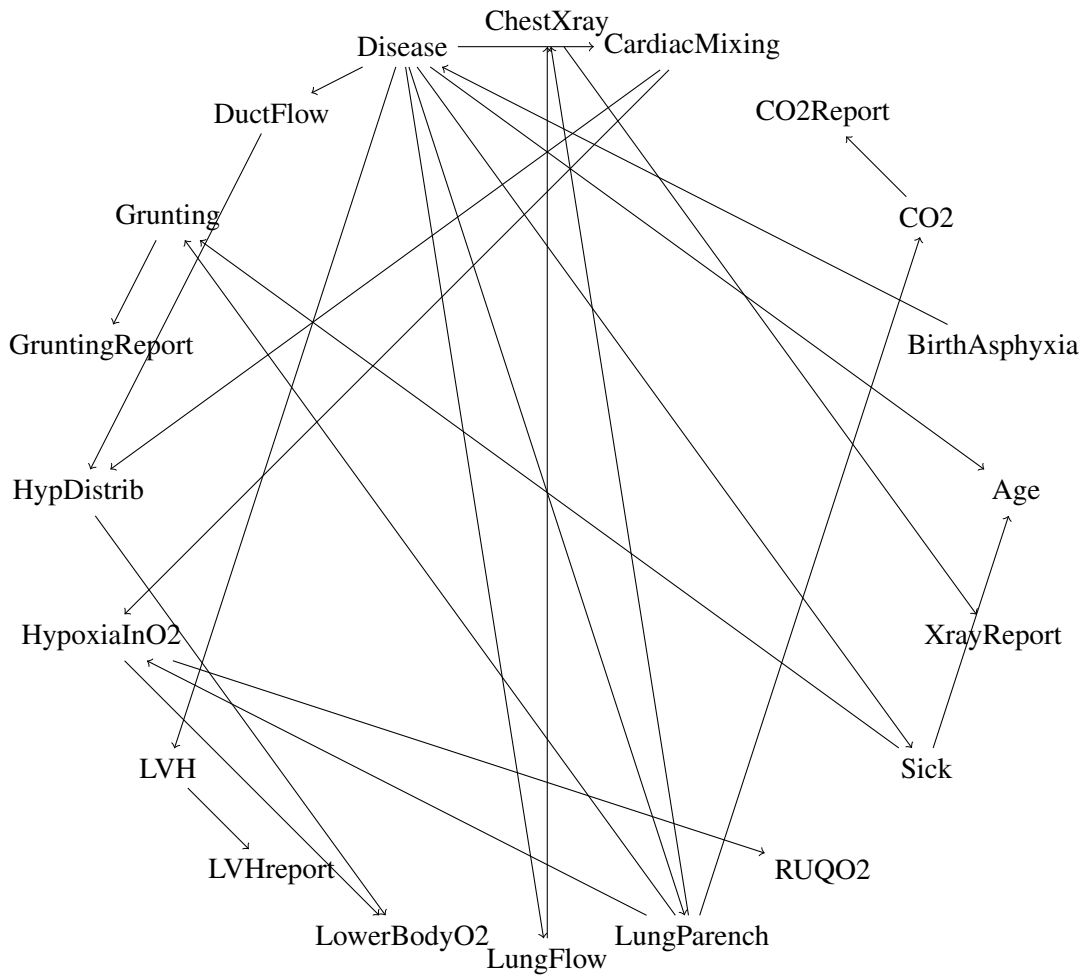Figure 19: Survey DAG

Figure 20: Asia DAG

Figure 21: Alzheimer's DAG

Figure 22: Child DAG

Figure 23: Insurance DAG
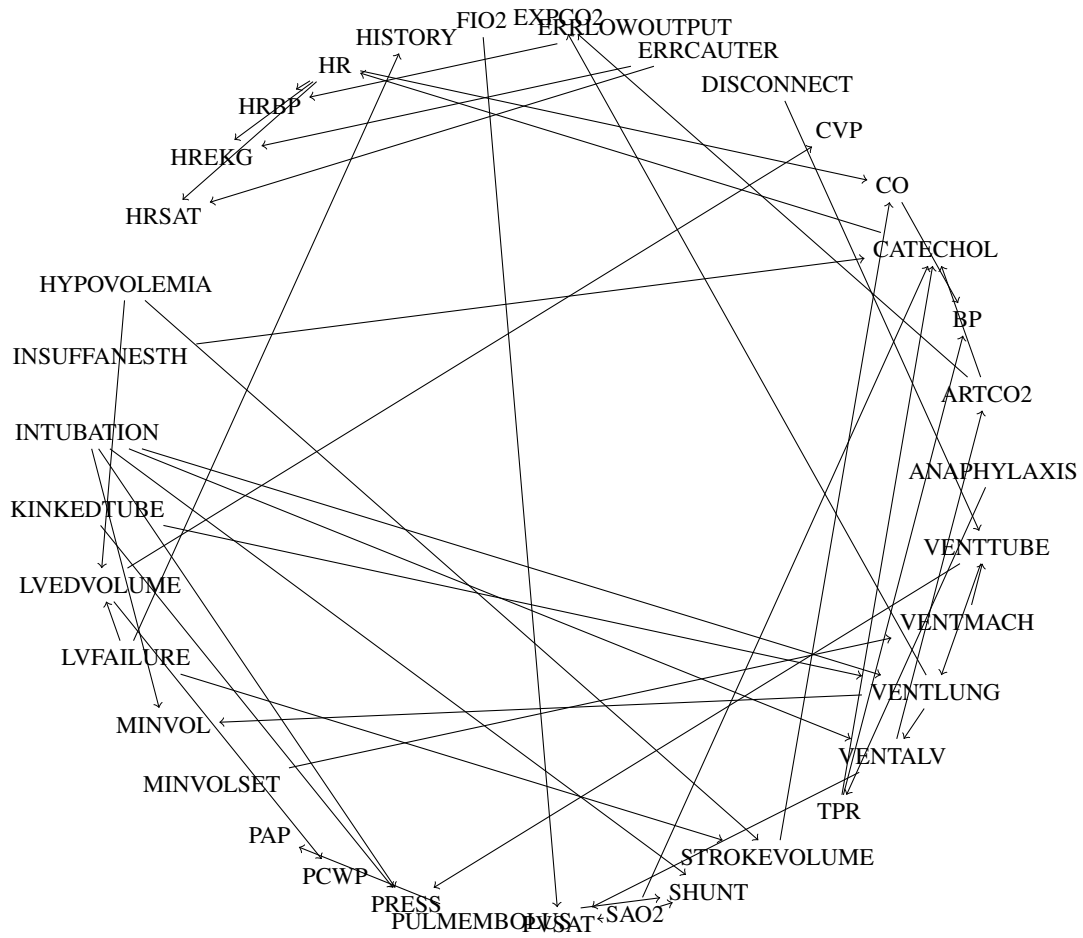
Figure 24: Alarm DAG

F Prompt template

Hello. You will be given a causal graph. The context of the graph [CONTEXT]. Please understand the causal relationships between the variables - [VERBALISED DAG].

Prompt 1: Base prompt to describe the causal graph

Hello. You will be given a causal graph. The context of the graph is hypothetical patient monitoring system in an intensive care unit (ICU). Please understand the causal relationships between the variables - < anaphylaxis > causes < total peripheral resistance >. < arterial co2 > causes < expelled co2 >. < arterial co2 > causes < catecholamine >. < catecholamine > causes < heart rate >. < cardiac output > causes < blood pressure >. < disconnection > causes < breathing tube >. < error cauter > causes < heart rate displayed on ekg monitor >. < error cauter > causes < oxygen saturation >. < error low output > causes < heart rate blood pressure >. < high concentration of oxygen in the gas mixture > causes < pulmonary artery oxygen saturation >. < heart rate > causes < heart rate blood pressure >. < heart rate > causes < heart rate displayed on ekg monitor >. < heart rate > causes < oxygen saturation >. < heart rate > causes < cardiac output >. < hypovolemia > causes < left ventricular end-diastolic volume >. < hypovolemia > causes < stroke volume >. < insufficient anesthesia > causes < catecholamine >. < intubation > causes < lung ventilation >. < intubation > causes < minute volume >. < intubation > causes < alveolar ventilation >. < intubation > causes < shunt - normal and high >. < intubation > causes < breathing pressure >. < kinked chest tube > causes < lung ventilation >. < kinked chest tube > causes < breathing pressure >. < left ventricular end-diastolic volume > causes < central venous pressure >. < left ventricular end-diastolic volume > causes < pulmonary capillary wedge pressure >. < left ventricular failure > causes < previous medical history >. < left ventricular failure > causes < left ventricular end-diastolic volume >. < left ventricular failure > causes < stroke volume >. < the amount of time using a breathing machine > causes < the intensity level of a breathing machine >. < sudden blockage in the pulmonary arteries > causes < shunt - normal and high >. < sudden blockage in the pulmonary arteries > causes < pulmonary artery pressure >. < pulmonary artery oxygen saturation > causes < oxygen saturation >. < oxygen saturation > causes < catecholamine >. < shunt - normal and high > causes < oxygen saturation >. < stroke volume > causes < cardiac output >. < total peripheral resistance > causes < catecholamine >. < total peripheral resistance > causes < blood pressure >. < alveolar ventilation > causes < arterial co2 >. < alveolar ventilation > causes < pulmonary artery oxygen saturation >. < lung ventilation > causes < expelled co2 >. < lung ventilation > causes < minute volume >. < lung ventilation > causes < alveolar ventilation >. < the intensity level of a breathing machine > causes < breathing tube >. < breathing tube > causes < lung ventilation >. < breathing tube > causes < breathing pressure >.

Prompt 2: An example of the base prompt for Alarm graph. Each relationship is enclosed in pointed brackets,<> followed by a full stop.

Prompt: Hello. You will be given a causal graph. The context of the graph is different aspects of a patient's health related to lung diseases. Please understand the causal relationships between the variables - < x > causes < tuberculosis >. < bronchitis > causes < dyspnoea-laboured breathing >. < individual has either tuberculosis or lung cancer > causes < positive xray >. < individual has either tuberculosis or lung cancer > causes < dyspnoea-laboured breathing >. < lung cancer > causes < individual has either tuberculosis or lung cancer >. < smoking cigarettes > causes < lung cancer >. < smoking cigarettes > causes < bronchitis >. < tuberculosis > causes < individual has either tuberculosis or lung cancer >. Based on your understanding and analytical skills, please identify what the variable X could most likely be from the given multiple choices. It's important to note that you MUST select an option from the provided choices, as one of them is the correct answer. The choices are: [visited Asia; weather; book sales; movie rating]. Please think step by step. Your answer should be formatted as follows - Answer: X = choice. The choice must be one of the provided options, and it should be formatted exactly as it appears in the choices. Generating your own options is not allowed in this task.

Answer: X = visited Asia.My thought process is as follows:The causal graph suggests that X causes tuberculosis, and tuberculosis is related to lung diseases. Visiting Asia is a possible explanation for why an individual may have been exposed to tuberculosis, as the disease is more prevalent in some Asian countries. Therefore, visited Asia is the most likely option for X.

Prompt 3: Out-of-context controlled variable identification, Ground truth variable: visited Asia

Prompt: Hello. You will be given a causal graph. The context of the graph is different aspects of a patient's health related to lung diseases Please understand the causal relationships between the variables - < x > causes < tuberculosis >. < bronchitis > causes < dyspnoea-laboured breathing >. < individual has either tuberculosis or lung cancer > causes < y >. < individual has either tuberculosis or lung cancer > causes < dyspnoea-laboured breathing >. < lung cancer > causes < individual has either tuberculosis or lung cancer >. < smoking cigarettes > causes < lung cancer >. < smoking cigarettes > causes < bronchitis >. < tuberculosis > causes < individual has either tuberculosis or lung cancer >. Based on your understanding and analytical skills, please identify what the variable X could most likely be from the given multiple choices. It's important to note that you MUST select an option from the provided choices, as one of them is the correct answer. The choices are: [weather; book sales; movie rating; positive xray; visited Asia]. Please think step by step. Your answer should be formatted as follows - Answer: X = choice. The choice must be one of the provided options, and it should be formatted exactly as it appears in the choices. Generating your own options is not allowed in this task.

Answer: Answer: X = visited Asia. My thought process is as follows:1. The graph shows that X causes tuberculosis, which suggests that X is a variable related to healthcare or medicine. 2. The graph also shows that bronchitis causes dyspnoea-laboured breathing, which further supports the idea that X is a variable related to healthcare or medicine. 3. The graph shows that individual has either tuberculosis or lung cancer causes y, which suggests that X could be a variable related to diagnosing or treating lung diseases. 4. The graph shows that smoking cigarettes causes lung cancer and bronchitis, which suggests that X could be a variable related to smoking or lung health. 5. Based on the above observations, it seems unlikely that X would be weather, book sales, or movie rating.6. Visited Asia is a variable that could potentially relate to healthcare or medicine, particularly if the patient has traveled to a region with a high prevalence of tuberculosis or other lung diseases.7. Therefore, based on the information provided in the graph, it seems most likely that X = visited Asia.

Prompt 4: In-context controlled variable identification, Ground truth variable: visited asia

Prompt: Hello. You will be given a causal graph. The context of the graph is different aspects of a patient's health related to lung diseases. Please understand the causal relationships between the variables - < x > causes < tuberculosis >. < bronchitis > causes < dyspnoea-laboured breathing >. < individual has either tuberculosis or lung cancer > causes < positive xray >. < individual has either tuberculosis or lung cancer > causes < dyspnoea-laboured breathing >. < lung cancer > causes < individual has either tuberculosis or lung cancer >. < smoking cigarettes > causes < lung cancer >. < smoking cigarettes > causes < bronchitis >. < tuberculosis > causes < individual has either tuberculosis or lung cancer >. Using your causal knowledge, can you suggest what variable X most likely? Give 5 suggestions for what variable X could be. Give reason for your suggestions. Finally, return your answer (without reasoning) in the following format: <Answer> [first suggestion, second suggestion, third suggestion .. and so on] </Answer> No explanations, no reasoning, just answers as a list as described earlier. For example, if the suggestions are A,B,C,D,E then - <Answer> [A, B, C, D, E] </Answer>

Answer: <Answer> [HIV/AIDS, Chronic respiratory conditions, **Pneumonia**, Exposure to asbestos, Silicosis] </Answer>
Reasons:
1. HIV/AIDS: It is known to weaken the immune system, making the individual more susceptible to infections like tuberculosis.
2. Chronic respiratory conditions: These can make an individual more susceptible to lung diseases like tuberculosis.
3. **Pneumonia**: It is a lung infection that can lead to serious conditions like tuberculosis if not treated properly.
4. Exposure to asbestos: Long-term exposure to asbestos can lead to lung diseases including lung cancer and tuberculosis.
5. Silicosis: It is a lung disease caused by inhalation of silica dust, leading to inflammation and scarring of the lung tissue, making the individual more susceptible to tuberculosis.

Prompt 5: Hypothesizing missing variable in open world, Ground truth variable: Visited Asia

Prompt: Hello. You will be given a causal graph. The context of the graph is different aspects of a patient's health related to lung diseases. Please understand the causal relationships between the variables - < visited asia > causes < tuberculosis >. < bronchitis > causes < dyspnoea-laboured breathing >. < individual has either tuberculosis or lung cancer > causes < positive xray >. < individual has either tuberculosis or lung cancer > causes < dyspnoea-laboured breathing >. < x > causes < individual has either tuberculosis or lung cancer >. < smoking cigarettes > causes < x >. < smoking cigarettes > causes < bronchitis >. < tuberculosis > causes < individual has either tuberculosis or lung cancer >. Using your causal knowledge, can you suggest what variable X most likely? Give 5 suggestions for what variable X could be. Give reason for your suggestions. Finally, return your answer (without reasoning) in the following format: <Answer> [first suggestion, second suggestion, third suggestion .. and so on] </Answer> No explanations, no reasoning, just answers as a list as described earlier. For example, if the suggestions are A,B,C,D,E then - <Answer> [A, B, C, D, E] </Answer>

Answer: <Answer> [**Lung Cancer**, Chronic Obstructive Pulmonary Disease, Emphysema, Pneumonia, Asthma] </Answer>

Reasons:

1. **Lung Cancer**: Smoking is a major cause of lung cancer and lung cancer can lead to tuberculosis or other lung diseases.

2. Chronic Obstructive Pulmonary Disease: Smoking is a leading cause of COPD and it can lead to other lung diseases including tuberculosis and lung cancer.

3. Emphysema: It is a type of COPD often caused by smoking and can lead to other lung diseases.

4. Pneumonia: Smoking increases the risk of pneumonia and pneumonia can lead to other lung diseases.

5. Asthma: Smoking can trigger asthma attacks and severe asthma can lead to other lung diseases.

Prompt 6: Hypothesizing missing variable in open world, Ground truth variable: Lung cancer

G Assumptions

The **causal sufficiency** of $\mathcal{G}$, by definition, implies that for every pair of variables within $\mathbf{V}$, all common causes are also included within $\mathbf{V}$. Extending this assumption to $\mathcal{G}^*$, we assume that the partial graph inherits causal sufficiency for its given that all edges among these variables are preserved as in $\mathcal{G}$. This preservation ensures that the observed relationships within V^* are not confounded by omitted common causes. Since the faithfulness of $\mathcal{G}$ ensures that the observed conditional independencies among variables in $\mathbf{V}$ are accurately reflected by the causal structure represented by $\mathbf{E}$. By maintaining the same set of edges $\mathbf{E}$ in $\mathcal{G}^*$ for the subset V^*, we uphold the faithfulness assumption within the partial graph.