Context-Aware Reasoning On Parametric Knowledge for Inferring Causal Variables

Anonymous ACL submission

Abstract

Scientific discovery catalyzes human intellectual advances, driven by the cycle of hypothesis generation, experimental design, evaluation, and assumption refinement. This process, while crucial, is expensive and heavily dependent on the domain knowledge of scientists to generate hypotheses. Recent work shows the potential of LLMs in assisting in scientific discovery and inferring causal relationships. Building on this, we introduce a novel task where the input is a partial causal graph with missing variables, and the output is a hypothesis about the missing variables. To evaluate this, we design a benchmark with varying difficulty levels. We show the strong ability of LLMs to hypothesize the mediation variables between a cause and its effect. In contrast, they underperform in hypothesizing the cause and effect variables themselves. Unlike simple knowledge memorization of fixed associations, this task requires the LLM to reason according to the context of the entire graph. This enables researchers to identify new variables of interest during the evolving scientific discovery process. By easily creating new examples with different missing variables, our benchmarks also test the robustness of models' parametric knowledge and their propositional reasoning between variables.

1 Introduction

Scientific discovery has been key to humankind's advances. It is a dynamic process revolving around inquiry and refinements. 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).

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., 2023b; Vashishtha et al., 2023; Darvariu et al., 2024).

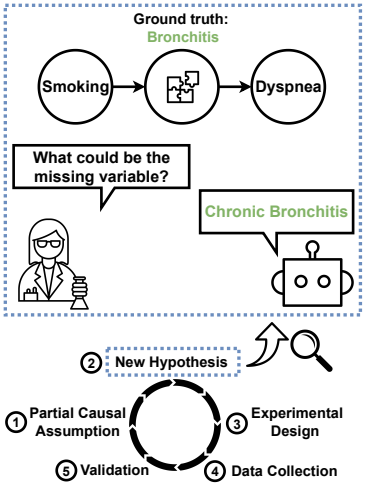


Figure 1: Scientific discovery iteratively generates hypotheses from assumptions using human expertise. We use LLMs as proxy experts to propose new hypotheses about missing variables in causal DAGs.

Unlike simple factual retrieval, scientific reasoning is inherently context-driven. Researchers dynamically adapt their hypotheses based on new observations, integrating knowledge across subpopulations. While prior work has explored LLMs for causal discovery (Kıcıman et al., 2023; Long et al., 2023; Darvariu et al., 2024; Ban et al., 2023b; Vashishtha et al., 2023), it has largely focused on identifying causal relationships from predefined variables. An essential yet underexplored aspect of reasoning is identifying *which* variables should be considered in the first place. This requires flexible reasoning, where context shapes the identification of missing causal factors.

Motivated by the need for flexible hypothesis generation, we propose a novel task where an LLM is given a partial causal graph with missing variables and the task is to generate hypotheses for these variables, using the known context of the graph as a primer. By changing which variable(s) to omit, we can instantiate new examples to test the robustness of the models' reasoning. We break down causal hypothesis generation into smaller tasks, starting with baseline experiments, and progressing to realistic scenarios where multiple nodes between treatments and outcomes are unknown.

In addition, this task aligns with real-world applications. Beyond capturing correlations, causal representations establish the underlying mechanisms that drive scientific phenomena, allowing researchers to test and refine hypotheses systematically. Establishing the variables of interest is a task that is heavily dependent on expert knowledge. We thus leverage LLMs to propose memorized or inferred variables based on their parametric knowledge. This assists researchers in identifying missing variables to guide data collection.

Contributions. 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 is based on the framework of causality as proposed by Pearl (2009). The intersection of language and causality is explored to extract causal relationships from a large corpus of text (Girju et al., 2002; Hassanzadeh et al., 2020; Tan et al., 2023; Dhawan et al., 2024). There has been an interest in using LLMs for causal reasoning (Kıcıman et al., 2023). Some works have focused on commonsense causality (Frohberg and Binder, 2021; Singh et al., 2021) and temporal causal reasoning (Zhang et al., 2020, 2022). More recently (Kıcıman et al., 2023; Long et al., 2023; Darvariu et al., 2024; Ban et al., 2023b; Vashishtha et al., 2023; Ban et al., 2023a) introduced methods to discover causal structures by prompting LLMs with variable names. Abdulaal et al. (2024) combined data-based deep structural causal models, such as (Yu et al., 2019), with LLMs generated causal structure. Jin et al. (2023) focused on causal inference using LLMs. Recent works have also attempted causality-informed transformer training (Vashishtha et al., 2024; Zhang et al., 2024). In contrast to prior work, we explore using 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).

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 $e_{i,j} \in \mathbf{E}$ denotes causal relationship between v_i and v_j, $v_i \xrightarrow{e_{i,j}} v_j$, emphasizing the influence from v_i to v_j. Beyond visualization, causal DAGs allow for the mathematical characterization of different node types for a causal model to understand the influences and dependencies.

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 with no incoming edges. Mathematically sources are $d_{\text{in}}(v) = 0$ where d_{in} is the in-degree of the graph.
Sink has no outgoing edges. Sinks are $d_{\text{out}}(v) = 0$ where d_{out} is the out-degree of the graph.
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 $d_{\text{out}}(v) > 0$), acting as intermediaries in the causal 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 are variables v that have two edges meeting, and have an in-degree greater than one $d_{\text{in}}(v) > 1$. Hence v is a collider if it is a child of both v_i and v_j.

4 Inferring Causal Variables

In this work, we aim to leverage language models to infer variables in a causal DAG. Motivated by the process of hypothesizing a causal graph from a partially known structure (Glymour et al., 2019). We proceed under the assumption that some elements of the graph are already known. The aim is to find additional variables that can be incorporated into the existing causal structure to enhance the underlying causal mechanism.

We assume a partially known causal DAG, defined as $\mathcal{G}^* = (\text{V}^*, \text{E})$, where $\text{V}^* \subseteq \mathbf{V}$. The objective is to identify the set of missing variables $\text{V}^* = \mathbf{V} \setminus \text{V}_{\text{missing}}$ 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.

Our methodology increases task complexity to assess LLMs' ability to infer missing causal variables. 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 must hypothesize missing variables simulating 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

This task (depicted in Figure 2a) evaluates LLMs' ability to identify missing variables in a causal graph from a list of multiple choices, thereby reconstructing the original graph. The partial DAG $\mathcal{G}^*$ is created by removing one variable from the original DAG $\mathcal{G}$. Let us denote the removed node as v_x. Along with the partial graphs, we operate in the multiple-choice question answering (MCQA) paradigm. 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 multiple choices include the missing variable v_x and out-of-context distractors. The out-of-context distractors are unrelated to the causal domain of the given DAG, chosen to minimize any contextual and overlap with the true missing variable. Let v_x^* represent the variable selected by the LLM to complete $\mathcal{G}^*$.

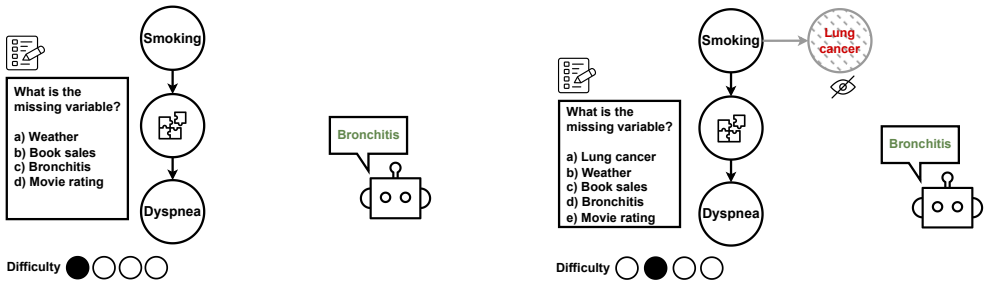
$$v_x^* = P_{\text{LLM}}(\mathcal{G}^*, \text{MCQ}_{v_x}) \quad \forall v_x \in \mathbf{V}$$

4.2 Task 2: In-Context Identification

In practical applications, such as healthcare (Robins, 1986) and finance (Hughes et al., 2019), dealing with missing data and unobserved latent variables is a major challenge (Tian and Pearl, 2012; Bentler, 1980). Therefore, it is important to identify the missing variables and their underlying mechanism. To simulate this, a more challenging task is introduced (see Figure 2b). 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, $\mathcal{G}^*$.

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

We use the MCQA paradigm to provide 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.



(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).

We introduce the non-parental constrain for v_{x_1} and v_{x_2}. This prevents the removal of both a parent node and its immediate child node in $\mathcal{G}^*$.

$$v^*_{x_1} = P_{\text{LLM}}(\mathcal{G}^*, \text{MCQ}_{v_{x_1}, v_{x_2}}) \quad \forall\, v_{x_1}, v_{x_2} \in \mathbf{V} \text{ and } v_{x_1} \not\to v_{x_2}, \quad v_{x_2} \not\to v_{x_1}$$

4.3 Task 3: Hypothesizing in Open World

So far, our testbeds required variable identification in a partial DAG given the controlled world knowledge in the form of distractors. This assumption allows for the evaluation of the language model's ability to select the correct answer from a set of options. However, in the open-world setting, we increase the complexity to provide no choices, as shown in Figure 3a. Hence the task is to predict the missing node v_x given the partial graph $\mathcal{G}^*$ to complete the ground truth graph $\mathcal{G}$. Here, the model returns a set of potential hypotheses, $\{v^*_{x,1}, ..., v^*_{x,k}\}$ where k is the number of hypotheses.

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

4.4 Task 4: Iteratively Hypothesizing in Open World

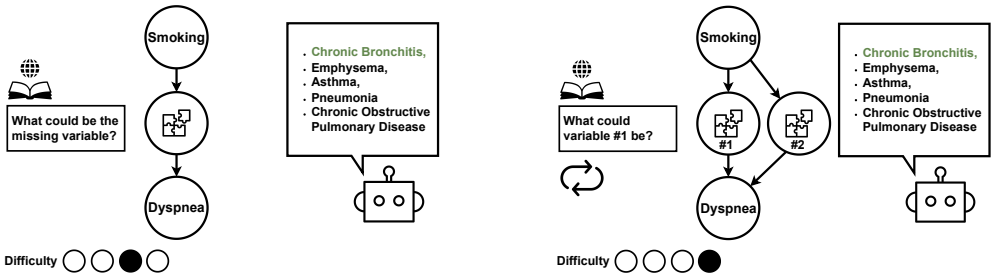
In addition to the search space relaxation, we further relax the number of missing variables. The partial DAG here, is obtained for one or more missing node variables. $\mathcal{G}^* = \mathcal{G} \setminus \{v_{x_1} ... v_{x_M}\}$. The fine-grained results from the open-world setting reveal that language models exhibit a particularly strong performance in identifying mediator variables. Thus, the LLM is used here to iteratively hypothesize mediator variables in a causal DAG given a treatment and an effect. The task (shown in Figure 3b) is set up as follows: given a partial graph $\mathcal{G}^*$, which includes observed treatment and outcome variables, we aim to hypothesize a set of mediators, denoted as $M = \{v_{m_1}, v_{m_2}, ..., v_{m_H}\}$, that mediates the treatment v_t to the outcome v_y. Here, H represents the number of direct, and indirect mediators. A pair of treatments and outcomes are considered iteratively across the causal DAG. In the first iteration, the LLM generates a hypothesis for the mediator v_{m_1}. The hypothesized mediator, v_{m_1} is then added to the graph, updating $\mathcal{G}^* \to \mathcal{G}^* \cup \{v_{m_1}\}$. The partial graph that now also includes $v^*_{m_1}$ can be used to identify the second mediator $v^*_{m_2}$ and so on. Therefore, in each subsequent iteration i, the LLM is tasked to generate a hypothesis for the next missing mediator v_{m_i} given the updated graph $\mathcal{G}^* \cup \{v^*_{m_1}, ..., v_{m^*_{i-1}}\}$.

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

for $i = 1, ..., H$. The sequence of mediators $M = \{v_{m_1}, v_{m_2}, ..., v_{m_H}\}$ is chosen at random. To formally investigate how the order of hypothesized mediators influences LLMs' performance, we borrow concepts from the mediation analysis literature (Pearl, 2014), specifically the Natural Direct Effect (NDE) and the Natural Indirect Effect (NIE). NDE measures the effect of the treatment on the outcome that is not mediated by a particular mediator, while NIE measures the effect of the treatment that is mediated by the mediator(see Appendix A.4). We introduce the Mediation Influence Score (MIS) that quantifies the influence of each mediator between a treatment and an effect. MIS defined as the ratio of NIE to NDE, provides a scale-free measure of a mediator's relative influence, enabling prioritization. MIS is always positive, reflecting the absolute contribution of mediators.

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

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.



(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).

5 Evaluation and Results

Experimental setup. We evaluate a variety of causal datasets spanning diverse domains. We use the semi-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 dataset (Abdulaal et al., 2024), developed by five domain experts and Law (VanderWeele and Staudt, 2011). See Appendix A.1 for further details.

We evaluate our setups across different open-source and closed models. The models we use are GPT-3.5 (Brown et al., 2020), GPT-4 (OpenAI, 2023), LLama2-chat-7b (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). Implementation details are in Appendix A and prompts in Appendix F.

5.1 Task 1

Here, the input to the LLM is the ground truth variable name in addition to out-of-context multiple choices for the missing variable v_x and the partial DAG $\mathcal{G}^*$. We then calculate the models' accuracy in correctly predicting v_x.

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

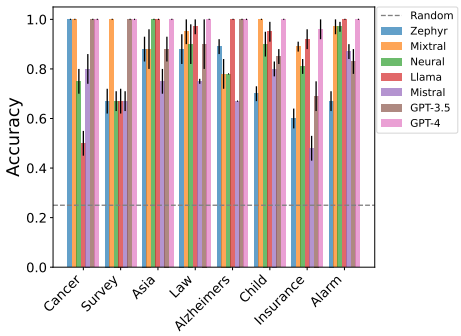
Results. In Figure 4a, we report the accuracy of different LLMs in identifying the missing variable. GPT-4, followed by Mixtral, consistently performs well, achieving perfect accuracy on most of the datasets. GPT-3.5 also shows overall strong performance, apart from the Insurance and Alarm datasets. Other models, including Mistral-7b, Llama-7b, and Zephyr-7b, demonstrate varying degrees of success. Insurance is the most challenging dataset, which could potentially be due to the high number of edges present in the DAG. All models significantly outperform the random baseline. However, we may conjecture that the high performance could be attributed to the simplicity of the task. The models might be using the context of the dataset domain to exclude obviously unrelated distractors rather than using the context to infer and reason from multiple plausible choices. To investigate this, we introduce an in-domain choice in the multiple choices in the next experiment.

5.2 Task 2

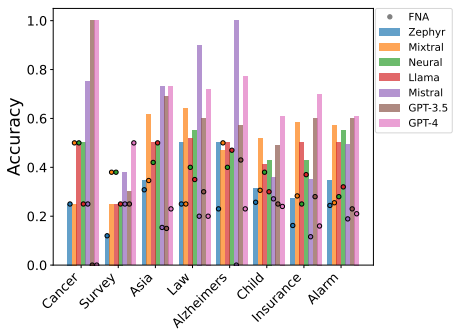
We introduce a more complex setting in which the partial graph has two missing nodes. Alongside the out-of-context choices and the ground truth variable, the multiple-choice options also include the second missing node from the partial graph as an in-context distractor. This requires the language model to reason about indirect causal relationships to identify the correct missing variable. To evaluate models' performance, we present two metrics: Accuracy and False Node Accuracy (FNA). FNA measures the confusion of LLMs in picking the in-context variable instead of the ground truth.

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

Results. In Figure 4b, we plot both Accuracy and FNA across different datasets. Ideally, accuracy should be 1.0, and the FNA should be 0.0. Since there were 5 multiple choices, the random chance is 0.2. We observe that most of the models for larger datasets achieve much higher accuracy than random chance. GPT-3.5 and GPT-4 consistently perform well across all datasets. On the other hand, open-source models like Mistral, Zephyr, and



(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).

	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-3.5	0.48	0.74	0.42	**0.79**	0.47	0.61	0.52	0.73	0.39	**1.00**	0.36	0.60	0.47	0.52	0.48	0.73	0.44	0.71
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**

Table 1: Task 3 Results. Average semantic similarity and LLM-as-Judge metrics to evaluate LLMs in hypothesizing the missing variable in a causal DAG.

Mixtral show varying performance across different datasets. For instance, Mistral performs well on the easy Cancer dataset but underperforms in the more complex Alarm dataset. In summary, we observe that most language models can highly outperform random chance for identifying causal variables in the presence of multiple missing nodes and an in-context distractor. However, the strength of their reasoning is dependent on the ability of the model to adapt to the specific causal context of the DAG.

5.3 Task 3

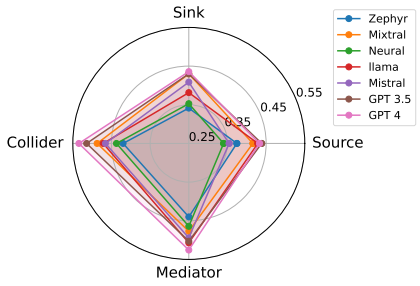
In realistic scenarios, where scientists provide incomplete graphs without pre-defined answers, there is often no single 'ground truth' for what the missing variable should be. The correct hypothesis may vary based on domain expertise or available data, making this task fundamentally open-ended. Hence in this task, we leverage LLMs to hypothesize the causal variables. The language model is prompted for $k = 5$ suggestions for the missing node v_x.

We compare the suggestions to the ground truth while acknowledging that real-world cases often lack a single correct answer. This complexity challenges traditional metrics, which may not fully capture models' performance, especially when the context within the causal graph is crucial (see Appendix C.5). Thus, we employ two evaluation metrics: semantic similarity and LLM-as-Judge.

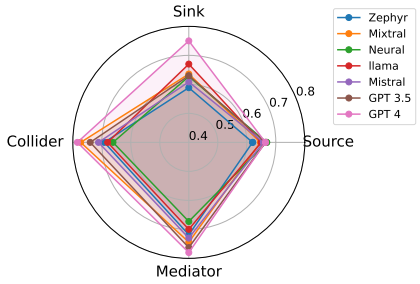
Semantic Similarity. We compute the cosine similarity between the embeddings of the predictions, $v^*_{x_{1:5}}$, and the ground truth v_x, averaging the highest similarity scores across all nodes $v_x \in \mathbf{V}$ (see Appendix A.5 for details).

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 1. For brevity, we provided the variances in Appendix C.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 datasets with a detailed performance per dataset in Figure 18. GPT-4 and Mistral generally achieve higher semantic similarity and LLM-as-Judge scores across most



(a) Semantic similarity.



(b) LLM-as-Judge.

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

datasets (Figure 18). GPT-3.5 also shows good average performance. 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 7). Despite different scales, both metrics seem to be fairly correlated. Figure 5, shows that models display stronger performance for colliders and mediators 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 16a, we observe that models' performance increases with k, i.e., with more suggestions. From Figure 16b, it is also evident that the performance is proportional to the number of total edges, $d_{in} + d_{out}$ (more context about the node). In summary, LLMs show impressive performance across some of the nodes and can be particularly useful to hypothesize mediators and colliders in a partial causal DAG. It is, hence, potentially beneficial to use LLMs in the real world because, in practice, treatment and outcomes are usually known.

5.3.1 Hypothesizing Confounder

In causal inference, backdoor paths are alternative causal pathways that confound the estimation of causal effects. They introduce bias when estimating causal effects if not appropriately addressed. 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 2, with detailed results in Appendix B, we observe that while some confounders were easily hypothesized by LLMs, achieving perfect accuracy, the genomic domain of the SACHS posed challenges for models with potentially less domain-specific knowledge. Similar to the mediator analysis, a large model: GPT-4, does not always perform best across all datasets. This highlights the need for a diverse set of benchmarks, like ours, to fully assess models' performance. Considering the importance of backdoor paths, we have benchmarked LLMs' performance for confounders in addition to colliders. LLMs typically perform well when hypothesizing a collider, however, the results for confounders are varied.

	Sachs	Alarm	Ins
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-3.5	0.40	0.49	0.67
GPT-4	**0.95**	0.73	**0.78**

Table 2: Hypothesizing Confounders in Task 3.

5.4 Task 4

For hypothesizing mediators, we adopted an iterative approach rather than a global (all-at-once) strategy. This interactive process allows the language model to progressively refine its predictions, reducing the search space for subsequent variables. As observed in our empirical results (see Appendix C.7), LLMs underperform when tasked with making multiple simultaneous predictions across different mediators. The iterative approach aligns more closely with human reasoning, as evidenced by Chain-of-Thought (CoT) (Wei et al., 2022) strategies, where sequential decision-making enhances accuracy.

For **unordered mediator evaluation**, the model is prompted iteratively with mediators presented in random order, and the final semantic similarity

is averaged across all predictions. In contrast, **ordered mediator evaluation** ranks the mediators using the Mediation Influence Score (MIS), prompting the model in both ascending and descending orders of significance. We introduce the metric Δ, presenting the difference in performance when mediators are iteratively presented to the LLM in ascending vs. descending orders of significance defined by the MIS. We selected the Asia, Child, Insurance, and Alarm datasets, as they offer a wider range of mediators, ranging from *1 to 10* mediators.

Results. The results of this experiment are in Table 3. Results with variances are provided in Appendix C.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 datasets, other models showed superior performance in Insurance and Alarm datasets only. As the complexity of the dataset 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.

	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-3.5	0.48	0.01	0.36	0.25	0.48	0.17	0.51	0.02
GPT-4	0.49	0.04	0.39	0.16	**0.52**	0.14	**0.60**	−0.07

Table 3: Task 4 Results. Iteratively hypothesizing the mediator nodes when prompted with random order. Δ measures the change in the prediction of each model when repeating the experiment with ordering according to the MIS metric, it measures the difference in performance between having ascending vs. descending order.

5.5 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. We evaluated the relative rankings across models (Appendix C.2) and found no model, including GPT-4, consistently outperformed others. Performance differences across domains may stem from biases in LLM training data, affecting parametric memory. For instance, confounder hypothesis quality varied across datasets, with domain-specific gaps lowering accuracy like in Sachs dataset (Appendix B).

We explored fine-tuning and few-shot prompting to enhance performance, but small DAG sizes limited the dataset size, yielding mixed results (Appendix D.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 dataset (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 1 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.

6 Conclusion

Most causality literature focuses on establishing causal relationships once data is collected, but generating hypotheses about which variables to observe is typically done by human experts. LLMs, trained on large datasets, can act as proxies for this task. We introduce the novel task of using LLMs to hypothesize missing variables in causal graphs, formalizing it with benchmarks that vary in difficulty and knowledge of the ground truth graph. We evaluate models on identifying missing variables from in-context and out-of-context distractors and hypothesizing variables in an open-world setting. We also explore an iterative approach for populating graphs with up to 10 missing mediator nodes. Our results show that LLMs are particularly effective at hypothesizing mediators, which are often less known than treatments and outcomes. This emphasizes the critical role of context-aware reasoning, where the model's ability to adapt its reasoning is based on the structure of the causal graph.

7 Limitation

While this work presents promising advancements in leveraging LLMs for hypothesizing missing variables in causal graphs, there are some limitations to consider. The potential for biases in the model's training data can influence the hypotheses generated which may affect downstream performance. Our evaluation relies on established DAGs and comparisons with known ground truth, limiting assessment in scenarios without a defined baseline. Future work can include human in loop evaluation.

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.

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*.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. 2023a. Causal structure learning supervised by large language model. *arXiv*.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023b. 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.

Peter M Bentler. 1980. Multivariate analysis with latent variables: Causal modeling. *Annual review of psychology*, 31(1):419–456.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

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 preprint arXiv:2402.18216*.

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.

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*.

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*.

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.

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*.

James Robins. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

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.

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, Abhinav Kumar, Abbavaram Gowtham Reddy, Vineeth N Balasubramanian, and Amit Sharma. 2024. Teaching transformers causal reasoning through axiomatic training. *arXiv*.

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.

Jiaqi Zhang, Joel Jennings, Agrin Hilmkil, Nick Pawlowski, Cheng Zhang, and Chao Ma. 2024. Towards causal foundation model: on duality between optimal balancing and attention. In *ICML*.

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 datasets. These datasets span different domain knowledge topics. These datasets have ground truth graphs along with their observational data. The simplest dataset used is the cancer dataset with 4 edges and 5 node variables. In addition to the semi-synthetic datasets from the BNLearn library, we also evaluate our approach on a realistic Alzheimer's Disease dataset (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.

Dataset	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 4: Dataset 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 datasets are under CC BY-SA 3.0 which allows us to freely modify the datasets for benchmarking. Our benchmark will be released under the CC BY-SA License.

GPT-3.5, GPT-4 was accessed via API. The rest of the models were run on 1 A100 GPU. Since we used 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 compute, 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 datasets. 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 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

Input: Partial graph $\mathcal{G}^*$, Ground truth node variables V_{GT}, Language model LM = 'all-mpnet-base-v2'
Output: Average highest semantic similarity score
procedure SEMANTICSIMILARITY($\mathcal{G}^*, V_{\text{GT}}, LM$)
    Initialize $similarityScores$ as an empty list
    for each node v_{GT} in **v do**
        $predictions \leftarrow$ GeneratePredictions($\mathcal{G}^*, LM$)
        Initialize $nodeScores$ as an empty list
        for each prediction p in $predictions$ **do**
            $embedding_{\text{GT}} \leftarrow$ Embed(v_{GT}, LM)
            $embedding_p \leftarrow$ Embed(p, LM)
            Normalize $embedding_{\text{GT}}$ and $embedding_p$
            $score \leftarrow$ CosineSimilarity($embedding_{\text{GT}}, embedding_p$)
            Append $score$ to $nodeScores$
        end for
        $maxScore \leftarrow$ Max($nodeScores$)
        Append $maxScore$ to $similarityScores$
    end for
    $averageScore \leftarrow$ Average($similarityScores$)
    return $averageScore$
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 5: Examples of model suggestions from and the corresponding semantic similarity score for a missing node variable from each of the datasets.

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

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

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
Output: Updated graph $\mathcal{G}^*$ with selected mediators
procedure GENERATEMEDIATORSRANDOM($\mathcal{G}^*, v_t, v_y, H, k$)
    for $i \leftarrow 1$ to H **do**
        $suggestions \leftarrow$ Generate k suggestions for v_{m_i} using $P_{\text{LLM}}(\mathcal{G}^*)$
        Initialize $highestSimilarity \leftarrow 0$
        Initialize $selectedMediator \leftarrow$ null
        for each $suggestion$ in $suggestions$ **do**
            $similarityScore \leftarrow$ Calculate semantic similarity for $suggestion$
            if $similarityScore > highestSimilarity$ **then**
                $highestSimilarity \leftarrow similarityScore$
                $selectedMediator \leftarrow suggestion$
            end if
        end for
        Update $\mathcal{G}^* \leftarrow \mathcal{G}^* \cup \{selectedMediator\}$
    end for
    return $\mathcal{G}^*$
end procedure

Algorithm 4 Ordered Mediator Generation and Evaluation Based on MIS

Input: Partial graph $\mathcal{G}^*$, Treatment v_t, Outcome v_y, Set of potential mediators M, Number of suggestions k
Output: Δ - measure of the influence of mediator ordering
procedure CALCULATEMIS(v_t, v_y, M)
    Initialize MISList as an empty list
    for each mediator v_{m_i} in M **do**
        Calculate NIE(v_{m_i}) and NDE(v_{m_i})
        MIS(v_{m_i}) $\leftarrow \frac{\text{NIE}(v_{m_i})}{\text{NDE}(v_{m_i})}$
        Append MIS(v_{m_i}) to MISList
    end for
    return MISList
end procedure
procedure GENERATEMEDIATORSORDERED($\mathcal{G}^*, v_t, v_y, M, k$)
    MISList $\leftarrow$ CALCULATEMIS(v_t, v_y, M)
    Sort M in descending order of MISList to get M_{desc}
    Sort M in ascending order of MISList to get M_{asc}
    $averageDesc \leftarrow$ GENERATEANDEVALUATE($\mathcal{G}^*, M_{\text{desc}}, k$)
    $averageAsc \leftarrow$ GENERATEANDEVALUATE($\mathcal{G}^*, M_{\text{asc}}, k$)
    $\Delta \leftarrow \frac{|averageDesc - averageAsc|}{averageDesc}$
    return Δ
end procedure
function GENERATEANDEVALUATE($\mathcal{G}^*, M_{\text{order}}, k$)
    Initialize similarityScores as an empty list
    for each mediator v_{m_i} in M_{order} **do**
        Perform the same steps as in the refined random order mediator generation
        (Generate k suggestions, select the most similar, update $\mathcal{G}^*$)
        Append the highest similarity score to similarityScores
    end for
    return Average of similarityScores
end function

B 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-3.5	0.34	0.39	0.48	0.48	1.0	0.58	0.20	0.48	0.47	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 8: 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-3.5	0.40	0.50	0.48	0.30	1.0	0.75	0.40	0.75	0.60	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 9: LLM judge

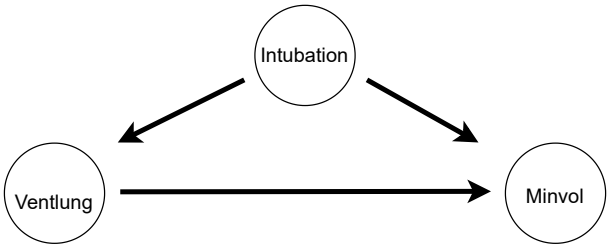


Figure 6: Alarm 1

C Further results

C.1 Variances

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

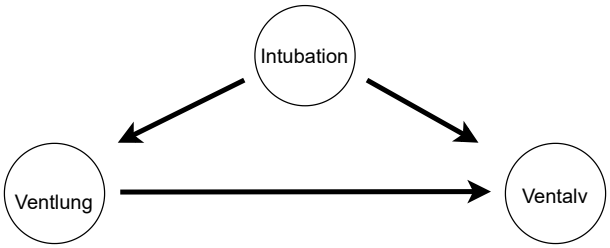


Figure 7: Alarm 2

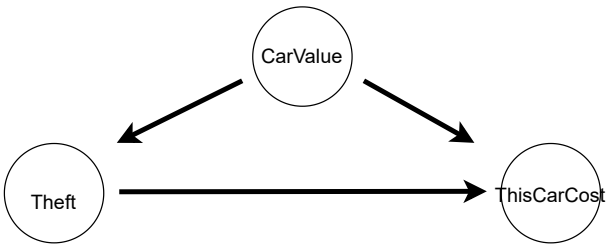


Figure 8: Insurance 1

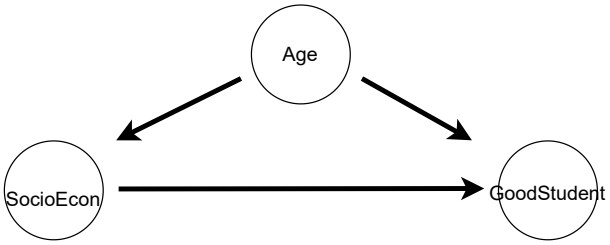


Figure 9: Insurance 2

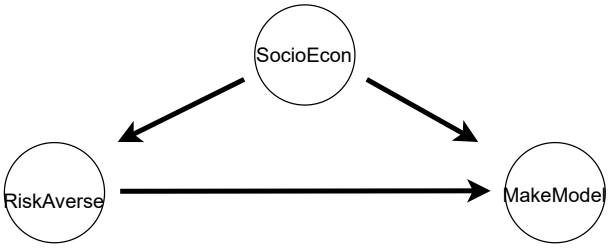


Figure 10: Insurance 3

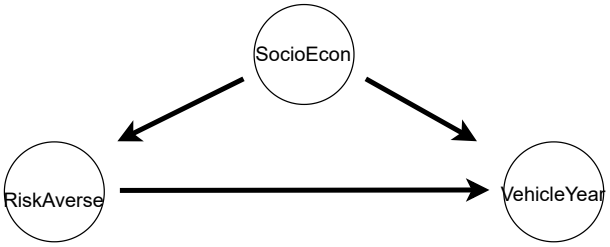


Figure 11: Insurance 4

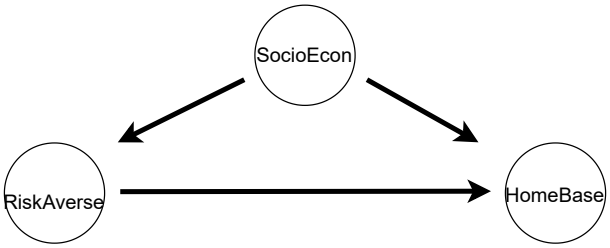


Figure 12: Insurance 5

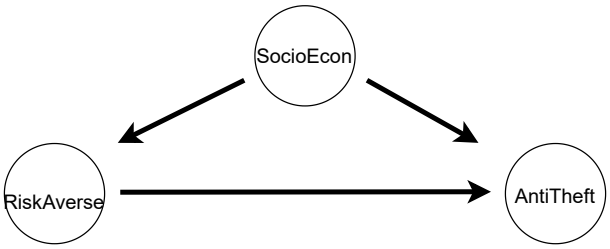


Figure 13: Insurance 6

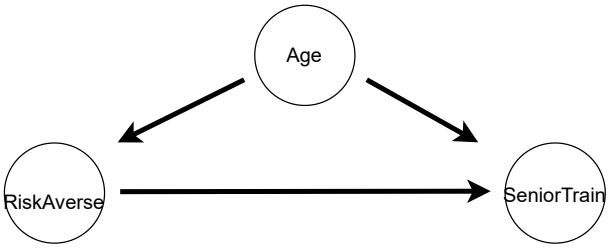


Figure 14: Insurance 7

	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 ±0.03	0.48 ±0.05	0.40 ±0.04	0.54 ±0.05	0.53 ±0.03	0.58 ±0.06	0.45 ±0.05	0.61 ±0.03	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-3.5	0.48 ±0.03	0.74 ±0.04	0.42 ±0.00	**0.79** ±0.03	0.47 ±0.04	0.61 ±0.04	0.39 ±0.05	**1.00** ±0.00	0.36 ±0.05	0.60 ±0.05	0.47 ±0.07	0.52 ±0.02	0.48 ±0.04	0.73 ±0.05	0.44 ±0.04	0.71 ±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 10: 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 ±0.05	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 ±0.01	0.48 ±0.05	0.13 ±0.02	0.47 ±0.04	0.11 ±0.01
GPT-3.5	0.48 ±0.05	0.01 ±0.01	0.36 ±0.04	0.25 ±0.03	0.48 ±0.05	0.17 ±0.02	0.51 ±0.05	0.02 ±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 11: 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.

C.2 Analysis of difference across tasks

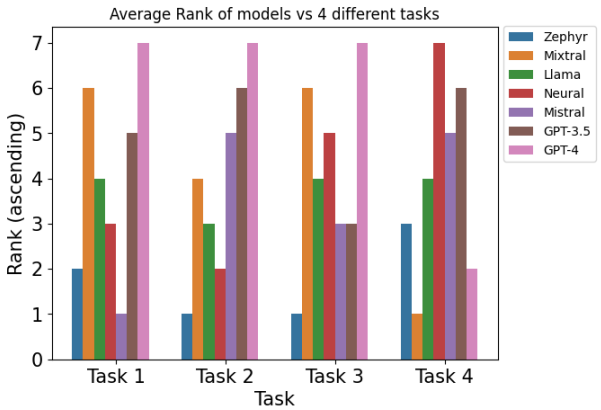


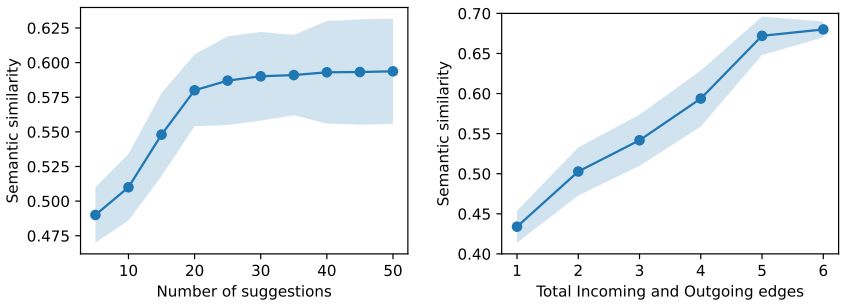
Figure 15: Average Rank of each model against the different tasks. We ranked the mode since the metrics are different to evaluate each task averaged across datasets

Since the metrics are different to evaluate each task, it is not meaningful or straightforward to compare the raw results. It must also be noted that the tasks are not linear. To address this, we rank the model performances across all models and datasets and present these rankings in Figure 15. This allows us to compare the relative performance of the models across different tasks.

As we observe from the graph, GPT-4 model shows consistently top performances in Tasks 1-3,

however, it has one of the lowest performances for Task 4. GPT-3.5 shows a strong performance in Task 2 and 4, ranking 2nd, but drops in Tasks 1 and 3. We observe that Zephyr, Neural and Mistral show consistently average performances. These observations motivate the significance of the tasks proposed in our benchmark. They highlight the variability in model performance across different tasks and emphasize the need for comprehensive and diverse benchmarks to fully assess the capabilities of these models.

C.3 Breaking down the performance



(a) (b)

Figure 16: L: Plot of semantic similarity with an increasing number of suggestions for GPT-4 on the Alarm dataset. R: Plot of semantic similarity against the total number of incoming and outgoing edges for GPT-4 on the Alarm dataset.

C.4 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 12: Model-Mixtral to evaluate the effect of context given in the prompt.

C.5 Using explanations

While using LLMs for hypothesizing the missing nodes withing the causal graph for the open world setting, introduced an additional question to prompt the model to provide explanations for each of their 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 13: 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 ambiguous hypothesis by the LLM that may have same semantics, which again breaks the semantic similarity metric. This further motivates 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 13 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.

C.6 Chain of thought

In recent times, 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. 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 14: Model-Mixtral to evaluate the effect of COT given in the prompt.

C.7 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

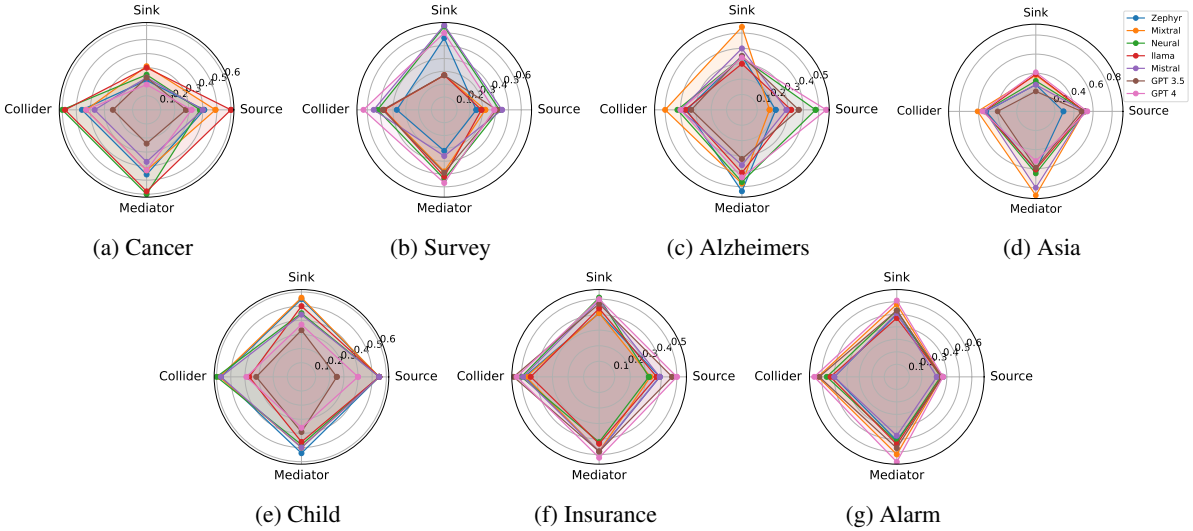
C.8 Results on Neuropathic Dataset

We added a new dataset, the neuropathic pain dataset (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 dataset 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-3.5	0.82	0.53	0.31	0.47	0.72
GPT-4	0.94	0.68	0.24	0.51	0.76

Table 15: Comparison of model performances across tasks on Neuropathic dataset.

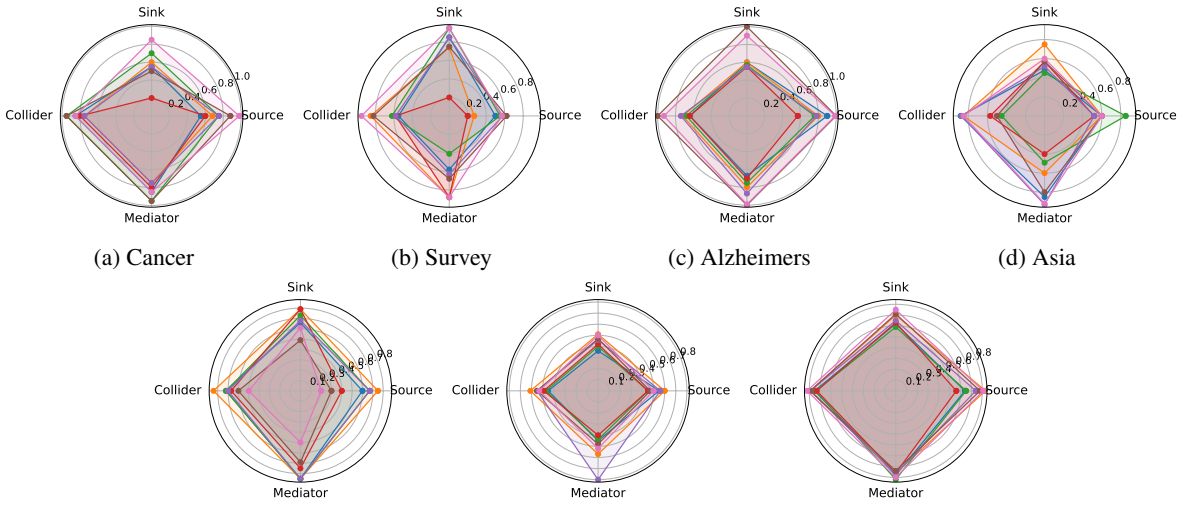
C.9 Fine grained model performance



(a) Cancer (b) Survey (c) Alzheimers (d) Asia

(e) Child (f) Insurance (g) Alarm

Figure 17: Detailed spider plots for Semantic similarity



(a) Cancer (b) Survey (c) Alzheimers (d) Asia

(e) Child (f) Insurance (g) Alarm

Figure 18: Detailed spider plots for LLM-as-judge metric

D Finetuning and Few-shot prompting

D.1 Finetuning

we aim to assess the LLM's causal reasoning via prompting. 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 dataset 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 16: 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 datasets which are contextually distant from the test domain, given the diversity of datasets 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.

D.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

Dataset	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 17: Fewshot prompting results.

medical dataset 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).

E Causal graphs

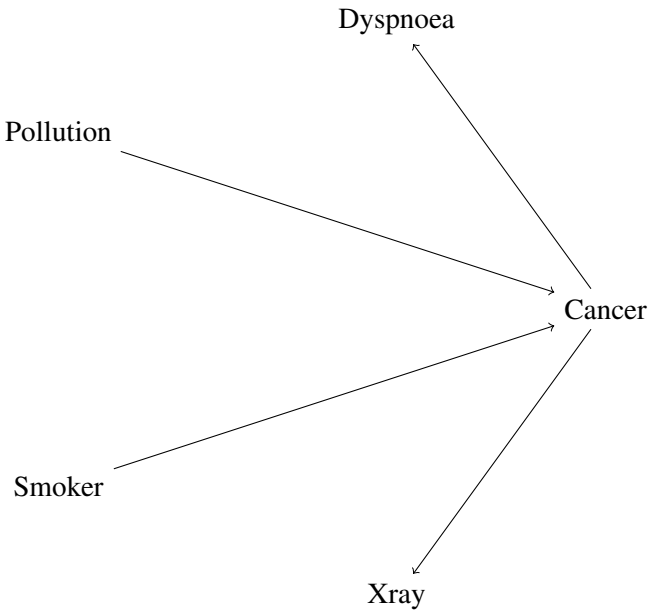


Figure 19: Cancer DAG

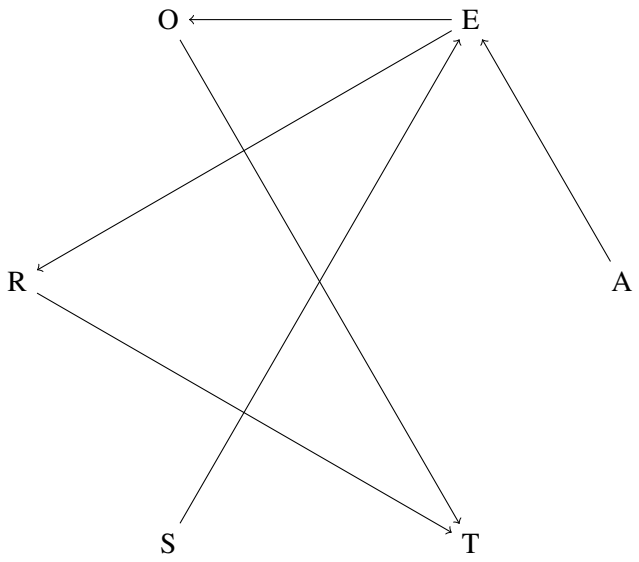


Figure 20: Survey DAG

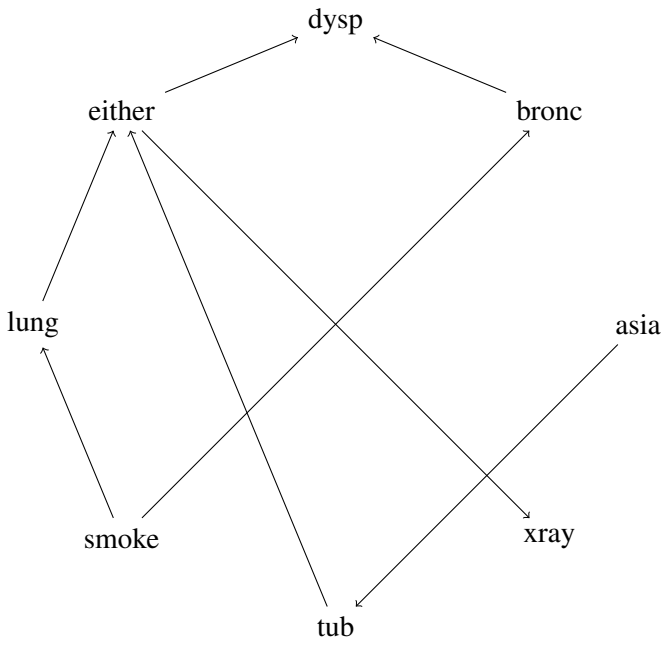


Figure 21: Asia DAG

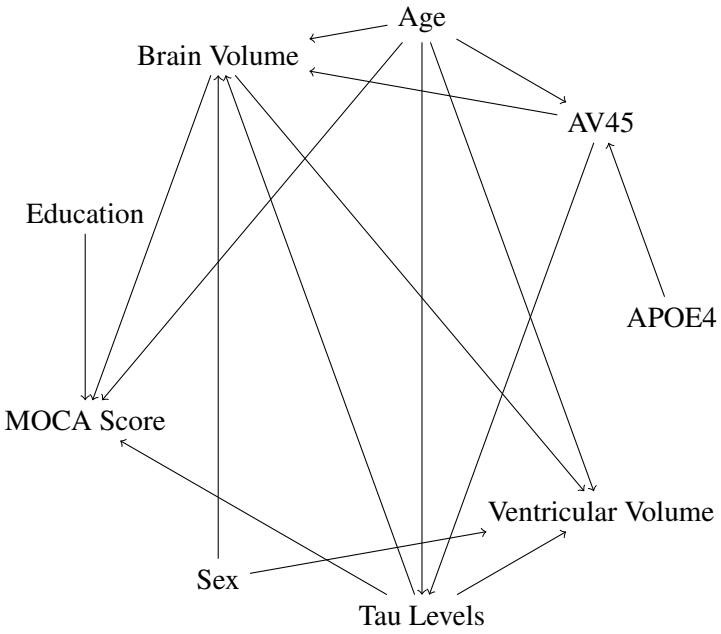


Figure 22: Alzheimer's DAG

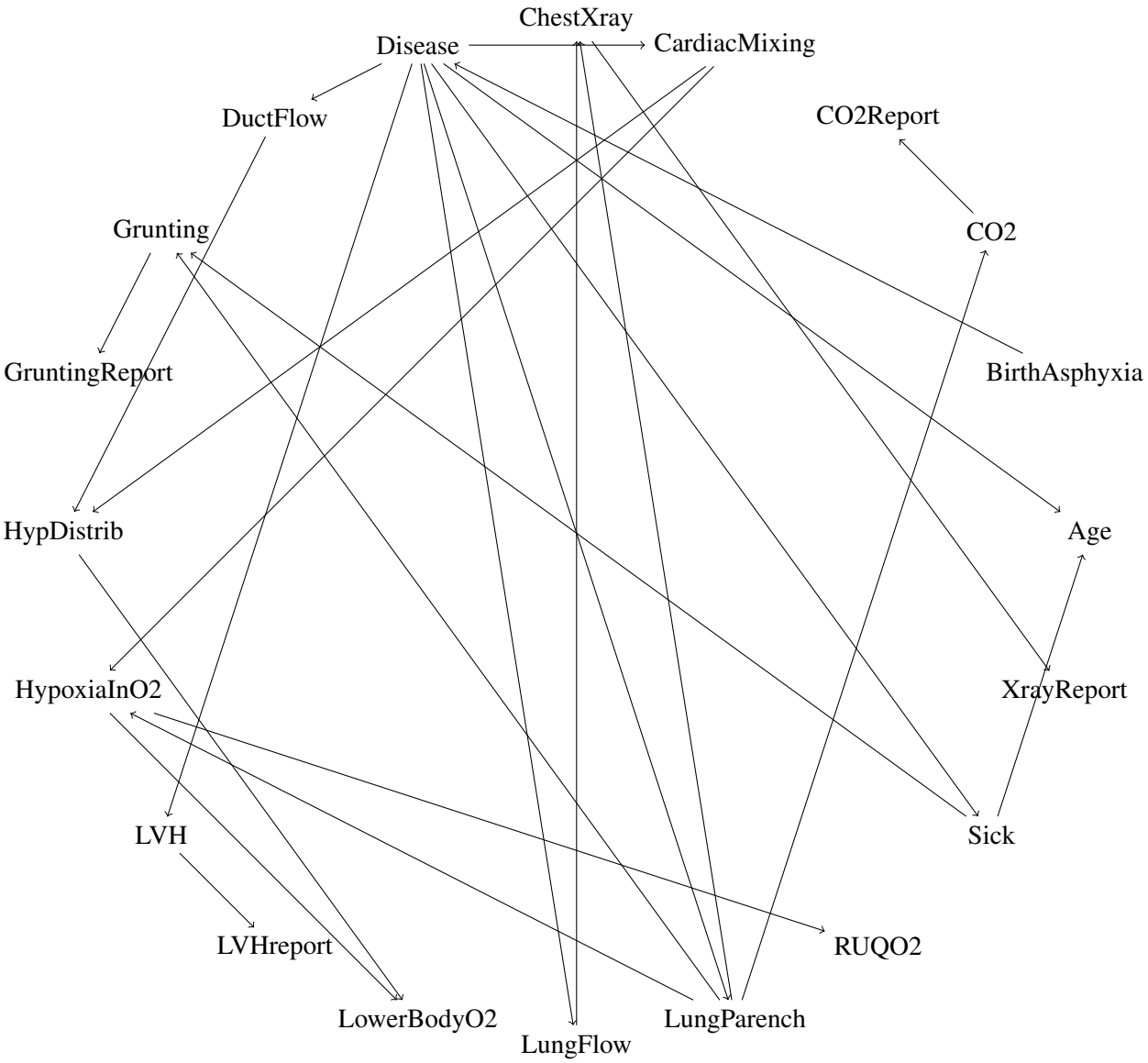


Figure 23: Child DAG

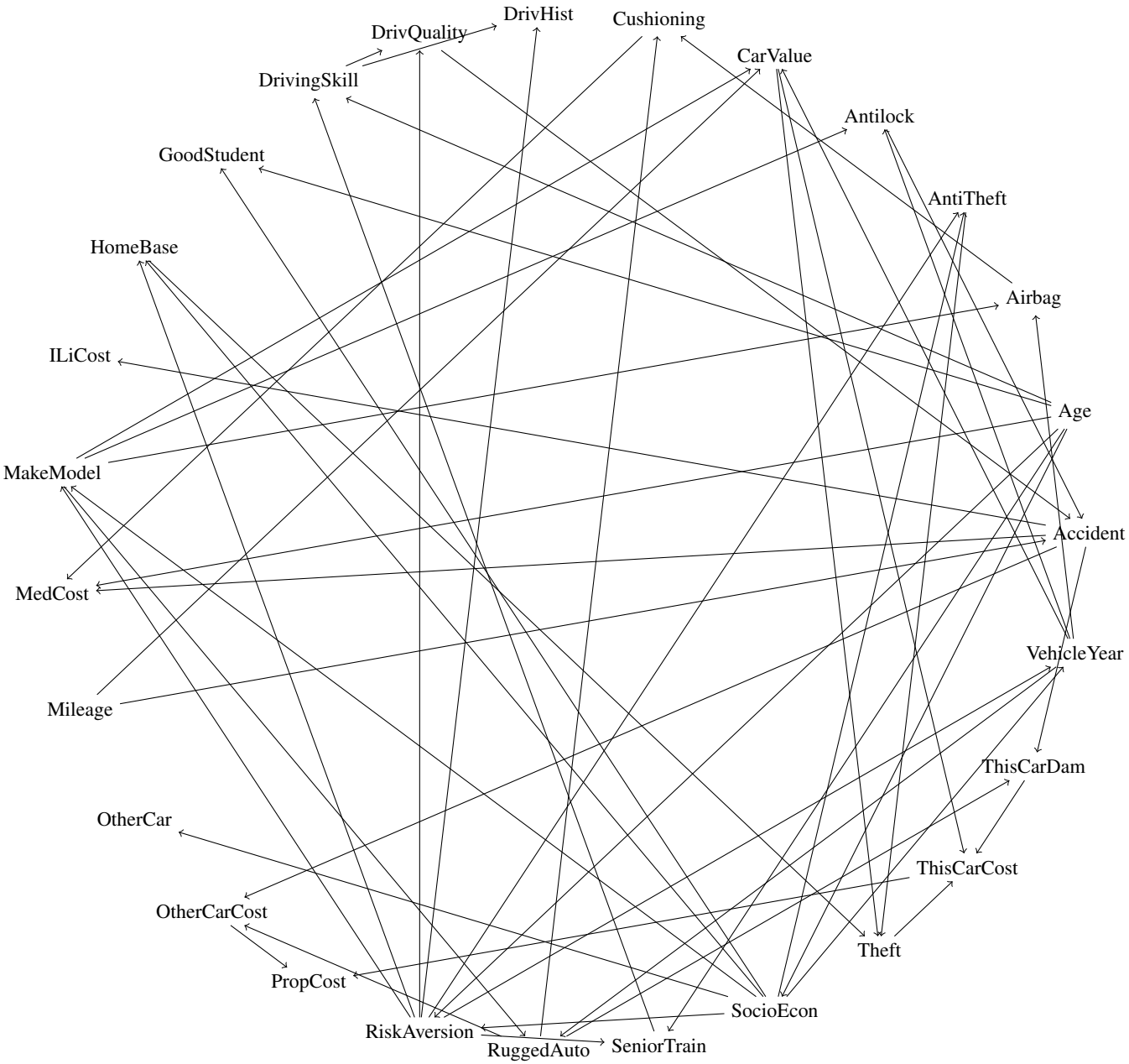


Figure 24: Insurance DAG

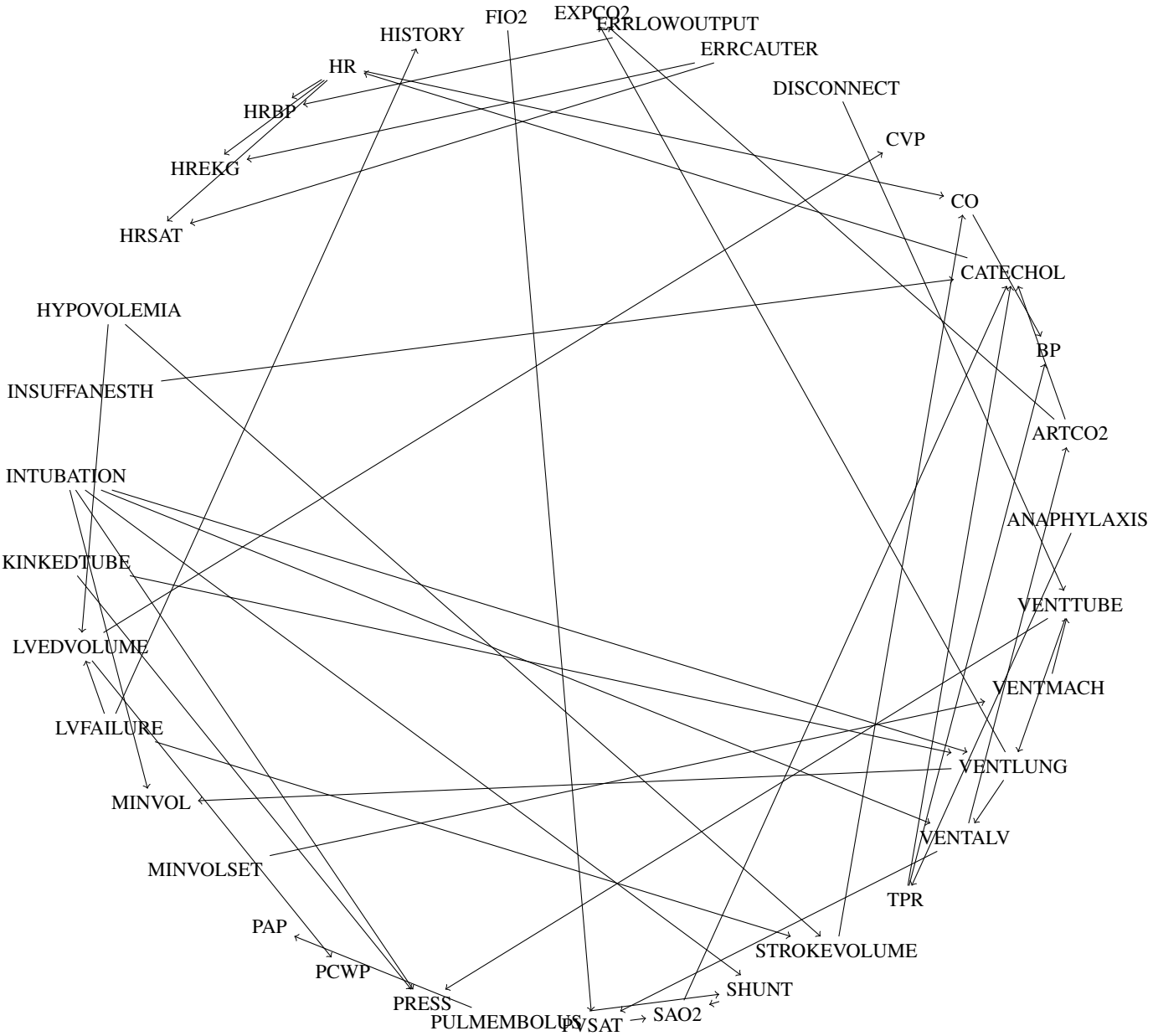


Figure 25: Alarm DAG

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].

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: 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: 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: 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: 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: 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 >. < x > causes < expelled co2 >. < x > causes < x3 >. < x3 > causes < x2 >. < x1 > causes < blood pressure >. < disconnection > causes < x7 >. < 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 < x9 >. < x2 > causes < heart rate blood pressure >. < x2 > causes < heart rate displayed on ekg monitor >. < x2 > causes < oxygen saturation >. < x2 > causes < x1 >. < hypovolemia > causes < left ventricular end-diastolic volume >. < hypovolemia > causes < stroke volume >. < insufficient anesthesia > causes < x3 >. < intubation > causes < x5 >. < intubation > causes < minute volume >. < intubation > causes < x4 >. < intubation > causes < shunt - normal and high >. < intubation > causes < breathing pressure >. < kinked chest tube > causes < x5 >. < 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 < x6 >. < sudden blockage in the pulmonary arteries > causes < shunt - normal and high >. < sudden blockage in the pulmonary arteries > causes < pulmonary artery pressure >. < x9 > causes < x8 >. < x8 > causes < x3 >. < shunt - normal and high > causes < x8 >. < stroke volume > causes < x1 >. < total peripheral resistance > causes < x3 >. < total peripheral resistance > causes < blood pressure >. < x4 > causes < x >. < x4 > causes < x9 >. < x5 > causes < expelled co2 >. < x5 > causes < minute volume >. < x5 > causes < x4 >. < x6 > causes < x7 >. < x7 > causes < x5 >. < x7 > causes < breathing pressure >. 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> [ventilation rate, respiratory effort, airway resistance, lung compliance, alveolar gas exchange] </Answer>
Reasons for the answer:
1. **Ventilation Rate**: The variable X causes both expelled CO2 and another variable (X3), which indirectly influences heart rate, blood pressure, and oxygen saturation. The rate of ventilation directly affects the amount of CO2 expelled from the body, making it a plausible candidate for X.
2. **Respiratory Effort**: This suggestion is based on the relationship between X and expelled CO2, as well as the influence of X on another variable (X3) that affects cardiovascular and respiratory parameters. Respiratory effort, which encompasses the work of breathing, could affect CO2 expulsion and have downstream effects on the cardiovascular system.
3. **Airway Resistance**: Airway resistance could be a candidate for X because it directly influences the amount of CO2 that can be expelled (a function of airflow) and could also affect other physiological variables through its impact on ventilation efficiency.
4. **Lung Compliance**: Lung compliance, or the ease with which the lungs can expand, directly affects ventilation and, consequently, the expulsion of CO2. It could also indirectly influence other variables in the graph through its impact on gas exchange efficiency.
5. ****Alveolar Gas Exchange****: This process directly determines the amount of CO2 expelled and could influence the variable X3, affecting cardiovascular parameters. Efficient gas exchange is crucial for maintaining appropriate oxygen and CO2 levels in the blood, impacting heart rate, blood pressure, and oxygen saturation.
These suggestions are not explicitly mentioned in the input graph but are closely related to the physiological processes and outcomes described. They offer plausible explanations for the causal relationships involving the variable X, considering the context of patient monitoring in an ICU setting.

Prompt: Strictly follow the format mentioned otherwise you will be disqualified.', 'ello. 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 >. < **Alveolar Gas Exchange** > causes < expelled co2 >. < **Alveolar Gas Exchange** > causes < x2 >. < x2 > causes < x1 >. < x > causes < blood pressure >. < disconnection > causes < x6 >. < 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 < x8 >. < x1 > causes < heart rate blood pressure >. < x1 > causes < heart rate displayed on ekg monitor >. < x1 > causes < oxygen saturation >. < x1 > causes < x >. < hypovolemia > causes < left ventricular end-diastolic volume >. < hypovolemia > causes < stroke volume >. < insufficient anesthesia > causes < x2 >. < intubation > causes < x4 >. < intubation > causes < minute volume >. < intubation > causes < x3 >. < intubation > causes < shunt - normal and high >. < intubation > causes < breathing pressure >. < kinked chest tube > causes < x4 >. < 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 < x5 >. < sudden blockage in the pulmonary arteries > causes < shunt - normal and high >. < sudden blockage in the pulmonary arteries > causes < pulmonary artery pressure >. < x8 > causes < x7 >. < x7 > causes < x2 >. < shunt - normal and high > causes < x7 >. < stroke volume > causes < x >. < total peripheral resistance > causes < x2 >. < total peripheral resistance > causes < blood pressure >. < x3 > causes < **Alveolar Gas Exchange** >. < x3 > causes < x8 >. < x4 > causes < expelled co2 >. < x4 > causes < minute volume >. < x4 > causes < x3 >. < x5 > causes < x6 >. < x6 > causes < x4 >. < x6 > causes < breathing pressure >. 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> [cardiac output, heart function, blood flow, vascular resistance, cardiac cycle] </Answer>
Reasons for the answer:
1. **Cardiac Output**: The variable X causes blood pressure and is caused by stroke volume. In physiological terms, cardiac output is a primary factor influencing blood pressure, as it represents the volume of blood the heart pumps per minute. Since stroke volume (the amount of blood pumped by the left ventricle of the heart in one contraction) contributes to cardiac output, it's logical to suggest "cardiac output" as a potential identity for X.
2. **Heart Function**: Given that X is related to blood pressure and influenced by stroke volume, it's reasonable to infer that X pertains to the overall function of the heart. Heart function encompasses various aspects of how the heart works, including pumping efficiency, which directly affects blood pressure.
3. **Blood Flow**: Since X has a direct relationship with blood pressure and an indirect relationship with factors affecting the heart's pumping ability, "blood flow" could be a plausible suggestion. Blood flow is crucial for maintaining blood pressure and is inherently linked to the heart's ability to pump blood effectively.
4. **Vascular Resistance**: X's impact on blood pressure and its connection to total peripheral resistance suggest that it might involve vascular resistance. Vascular resistance is the resistance that must be overcome to push blood through the circulatory system and is a key determinant of blood pressure.
5. **Cardiac Cycle**: The cardiac cycle involves all the events related to the flow of blood through the heart during one complete heartbeat, including systole (contraction) and diastole (relaxation). Given X's influence on blood pressure and its relationship with stroke volume, considering the cardiac cycle as a potential identity for X makes sense, as it encompasses the mechanisms affecting blood pressure regulation.

Prompt: 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 >. < < **Alveolar Gas Exchange** > causes < expelled co2 >. < < **Alveolar Gas Exchange** > causes < x1 >. < x1 > causes < x >. < **Cardiac Output** > causes < blood pressure >. < disconnection > causes < x5 >. < 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 < x7 >. < x > causes < heart rate blood pressure >. < x > causes < heart rate displayed on ekg monitor >. < x > causes < oxygen saturation >. < x > causes < **Cardiac Output** >. < hypovolemia > causes < left ventricular end-diastolic volume >. < hypovolemia > causes < stroke volume >. < insufficient anesthesia > causes < x1 >. < intubation > causes < x3 >. < intubation > causes < minute volume >. < intubation > causes < x2 >. < intubation > causes < shunt - normal and high >. < intubation > causes < breathing pressure >. < kinked chest tube > causes < x3 >. < 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 < x4 >. < sudden blockage in the pulmonary arteries > causes < shunt - normal and high >. < sudden blockage in the pulmonary arteries > causes < pulmonary artery pressure >. < x7 > causes < x6 >. < x6 > causes < x1 >. < shunt - normal and high > causes < x6 >. < stroke volume > causes < **Cardiac Output** >. < total peripheral resistance > causes < x1 >. < total peripheral resistance > causes < blood pressure >. < x2 > causes < < **Alveolar Gas Exchange** >. < x2 > causes < x7 >. < x3 > causes < expelled co2 >. < x3 > causes < minute volume >. < x3 > causes < x2 >. < x4 > causes < x5 >. < x5 > causes < x3 >. < x5 > causes < breathing pressure > 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> [patient anxiety level, patient consciousness level, patient pain level, patient respiratory rate, patient skin temperature] </Answer>
Reasons for the answer:
1. **Patient Anxiety Level**: The variable X influences heart rate, blood pressure, oxygen saturation, and cardiac output, which are all physiological parameters that can be affected by a patient's psychological state, such as anxiety. Anxiety can lead to increased heart rate and blood pressure, changes in oxygen saturation, and alterations in cardiac output due to the body's stress response.
2. **Patient Consciousness Level**: Consciousness level can impact heart rate, blood pressure, and oxygen saturation due to the brain's control over these vital signs. Changes in consciousness level, such as becoming more alert or lapsing into unconsciousness, can significantly affect these parameters, which are all influenced by variable X.
3. **Patient Pain Level**: Pain can cause physiological responses that include increased heart rate, elevated blood pressure, changes in oxygen saturation due to hyperventilation or altered breathing patterns, and potentially affect cardiac output as the body reacts to the stress of pain.
4. **Patient Respiratory Rate**: Given that X influences expelled CO2 and is affected by alveolar gas exchange, it's plausible that X could relate to the patient's respiratory rate. Respiratory rate directly impacts alveolar gas exchange, which in turn affects expelled CO2 levels, oxygen saturation, and indirectly influences heart rate and blood pressure through changes in blood chemistry.
5. **Patient Skin Temperature**: While not directly related to the cardiovascular or respiratory systems, skin temperature can be an indirect indicator of changes in these systems. For example, changes in cardiac output and peripheral resistance can affect blood flow to the skin, thereby influencing skin temperature. Variable X's impact on cardiac output and blood pressure suggests it could also indirectly affect skin temperature.
These suggestions are based on the understanding that variable X has a broad impact on both cardiovascular and respiratory parameters, indicating it could be a factor that influences or reflects the overall physiological state of the patient.

Prompt: 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 >. < < **Alveolar Gas Exchange** > causes < expelled co2 >. < < **Alveolar Gas Exchange** > causes < x >. < x > causes <Patient Respiratory Rate>. < **Cardiac Output** > causes < blood pressure >. < disconnection > causes < x4 >. < 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 < x6 >. <Patient Respiratory Rate> causes < heart rate blood pressure >. <Patient Respiratory Rate> causes < heart rate displayed on ekg monitor >. <Patient Respiratory Rate> causes < oxygen saturation >. <Patient Respiratory Rate> causes < **Cardiac Output** >. < hypovolemia > causes < left ventricular end-diastolic volume >. < hypovolemia > causes < stroke volume >. < insufficient anesthesia > causes < x >. < intubation > causes < x2 >. < intubation > causes < minute volume >. < intubation > causes < x1 >. < intubation > causes < shunt - normal and high >. < intubation > causes < breathing pressure >. < kinked chest tube > causes < x2 >. < 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 < x3 >. < sudden blockage in the pulmonary arteries > causes < shunt - normal and high >. < sudden blockage in the pulmonary arteries > causes < pulmonary artery pressure >. < x6 > causes < x5 >. < x5 > causes < x >. < shunt - normal and high > causes < x5 >. < stroke volume > causes < **Cardiac Output** >. < total peripheral resistance > causes < x >. < total peripheral resistance > causes < blood pressure >. < x1 > causes < < **Alveolar Gas Exchange** >. < x2 > causes < x6 >. < x3 > causes < expelled co2 >. < x3 > causes < minute volume >. < x3 > causes < x1 >. < x3 > causes < x4 >. < x4 > causes < x3 >. < x4 > causes < breathing pressure > 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> [oxygen delivery, arterial blood gas, lung compliance, blood oxygen level, respiratory efficiency] </Answer>
Reasons for the answer:
1. **Oxygen Delivery**: Given that "alveolar gas exchange" causes "X" and "X" in turn affects "patient respiratory rate," it's logical to infer that "X" could be related to how effectively oxygen is being delivered throughout the body. Oxygen delivery is a critical aspect of respiratory function and would directly impact respiratory rate.
2. **Arterial Blood Gas (ABG)**: ABG tests measure the levels of oxygen and carbon dioxide in the blood and can indicate how well oxygen is being transported and carbon dioxide is being expelled, which are functions influenced by alveolar gas exchange. Since "X" is caused by "alveolar gas exchange" and affects "patient respiratory rate," it's plausible that "X" could involve the measurement or state of arterial blood gases.
3. **Lung Compliance**: This refers to the elasticity or flexibility of the lungs, which directly impacts alveolar gas exchange and, consequently, the respiratory rate. Since "X" is linked to both "alveolar gas exchange" and "patient respiratory rate," it could suggest a relationship with the mechanical properties of the lung, such as compliance.
4. **Blood Oxygen Level**: This is a direct outcome of effective alveolar gas exchange and would naturally influence the patient's respiratory rate as the body adjusts to maintain adequate oxygenation. Given the causal links mentioned, "X" could very well represent the level of oxygen present in the blood.
5. **Respiratory Efficiency**: This encompasses how effectively the respiratory system can supply oxygen to and remove carbon dioxide from the body. It's influenced by alveolar gas exchange and directly affects the respiratory rate, making it a plausible candidate for "X."
These suggestions are not explicitly mentioned in the input graph but are logically inferred based on the described causal relationships and a basic understanding of respiratory physiology.

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.