Hypothesizing Missing Causal Variables with LLMs

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

Abstract

Scientific discovery is a catalyst for human intellectual advances, driven by the cycle of hypothesis generation, experimental design, data evaluation, and iterative assumption refinement. This process, while crucial, is expensive and heavily dependent on the domain knowledge of scientists to generate hypotheses and navigate the scientific cycle. Central to this is causality, the ability to establish the relationship between the cause and the effect. Motivated by the scientific discovery process, in this work, we formulate 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 complete the partial graph. We design a benchmark with varying difficulty levels and knowledge assumptions about the causal graph. With the growing interest in using Large Language Models (LLMs) to assist in scientific discovery, we benchmark open-source and closed models on our testbed. 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. We also observe surprising results where some of the open-source models outperform the closed GPT-4 model.

1 Introduction

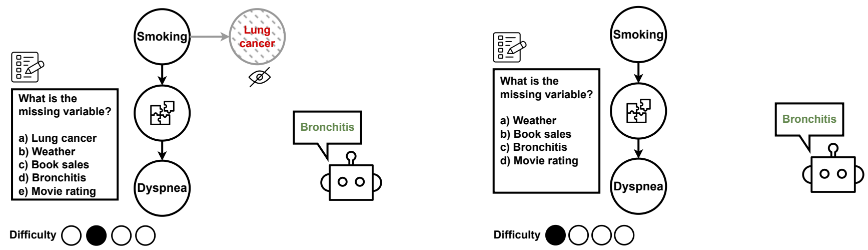
Scientific discovery is a dynamic process driven by inquiry, hypothesis formulation, and data collection [42]. Scientists refine hypotheses based on experimental data, form sub-questions, and iterate until the research question is resolved [21]. Causality plays a key role in assessing hypotheses, allowing interpretation beyond correlations. Tools like Randomised Control Trials (RCTs) establish causal relationships [20], but the process heavily relies on expert guidance for hypothesis formation and experimental design [21]. However, domain knowledge can be challenging to formalize and collect [21].

Recent advancements in Large Language Models (LLMs) [8, 26] have spurred interest in using them for scientific discovery [2]. LLMs have excelled in tasks like reasoning [36, 44] and are being explored in natural sciences [2]. Despite their capabilities, LLMs have limitations such as hallucinations, requiring human oversight [2]. Prior work suggests using LLMs as creative proposers of solutions with task-specific verification [31, 43, 29].

Given the importance of causality in scientific discovery, we explore how LLMs can aid causal reasoning. LLMs have shown state-of-the-art results in identifying pairwise causal relationships using variable names [21] and refining them with causal discovery algorithms [1, 3, 41]. However, these methods presuppose predefined variables and often involve costly data collection.

Our work extends LLMs' role in causal reasoning to hypothesize missing variables in partially known causal graphs, simulating a realistic scientific discovery process. This approach complements existing causal methods, leveraging LLMs' general and domain knowledge without requiring them to determine pairwise causal relations or perform numerical calculations, sidestepping their limitations in these tasks [48, 18].

NeurIPS 2024 Workshop on Causality and Large Models (CaLM).



(a) Task 1 (b) Task 2

Figure 1: Leveraging LLM to indentify 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 distractor (b).

In summary, our main contributions are:

- We introduce a new task of LLM-assisted causal variable identification and hypothesizing.
- We propose a benchmark for hypothesizing missing variables based on a diverse set of existing causal graph datasets.
- We design experimental tests with varying difficulty levels and knowledge assumptions, such as open-world and closed-world settings, the number of missing variables, etc., and gather insights on LLMs' capabilities and weaknesses.
- We benchmark several state-of-the-art models and analyze their performance with respect to variable types.

2 LLMs for Identifying and Hypothesizing Causal Variables

In this work, we leverage language models (LLMs) to identify and hypothesize variables in a causal Directed Acyclic Graph (DAG). Following the approach of hypothesizing causal graphs from partially known structures [12], we assume that some elements of the graph are known and aim to incorporate additional variables to refine the causal structure.

A causal DAG models relationships among N variables $\mathbf{V} = \{V_1, \ldots, V_N\}$, encoded as a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{E}$ is a set of directed edges that form no cycles. Given a partially known DAG $\mathcal{G}^* = (\mathrm{V}^*, \mathbf{E})$ with $\mathrm{V}^* \subseteq \mathbf{V}$, our goal is to identify missing variables, $\mathrm{V}_{\text{missing}} = \mathbf{V} \setminus \mathrm{V}^*$, to expand $\mathcal{G}^*$ to $\mathcal{G}$. This implies that all causal relationships among variables in V^* are fully represented in $\mathcal{G}^*$.

Our methodology explores LLMs' ability to hypothesize causal variables through progressively challenging scenarios. Initially, we provide a partially known DAG and multiple-choice options for missing variables. Complexity is increased by removing multiple nodes, and eventually, we transition to an open-ended scenario where no ground truth is provided, requiring LLMs to hypothesize missing variables without explicit hints. Given LLMs' limitation to textual input, we represent the graph $\mathcal{G}^*$ using a prompt template $P_{\text{LLM}}(\cdot)$, enabling LLMs to parse the causal relationships within the DAG.

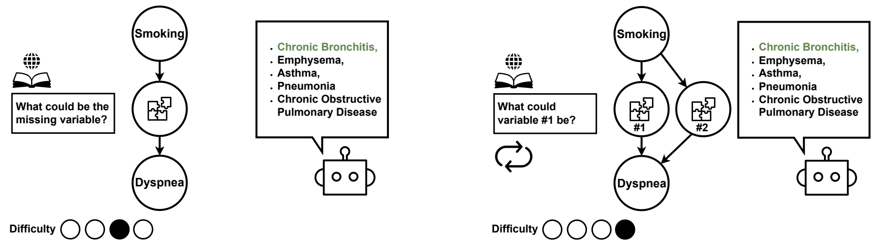
2.1 Task 1: Out-of-Context Controlled Variable Identification

This task (depicted in Figure 1a) 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 carefully chosen to be irrelevant to the given DAG and its context. Let v_x^* represent the variable selected by the LLM to complete $\mathcal{G}^*$.

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

2.2 Task 2: In-Context Controlled Variable Identification

In practical applications, such as healthcare [30] and finance [14], dealing with missing data and unobserved latent variables is a major challenge [38, 6]. Therefore, identifying the missing variables



(a) Task 3

(b) Task 4

Figure 2: Leveraging LLM to hypothesize missing variables in a causal DAG: (a) single variable, (b) iterative hypothesizing of multiple mediators.

and their underlying causal mechanism is an important task. To simulate this, a more challenging task is introduced (see Figure 1b). 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 that include 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. 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 $\mathcal{G}^*$.

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

2.3 Task 3: Hypothesizing in Open World

So far, we have described the testbeds for 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 2a. 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.

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

2.4 Task 4: Iteratively Hypothesizing in Open World

We extend the search space by relaxing the number of missing variables, with partial DAGs derived by removing one or more nodes: $\mathcal{G}^* = \mathcal{G} \setminus \{v_{x_1}, \ldots, v_{x_M}\}$. Open-world results show that LLMs excel in identifying mediators, prompting us to iteratively hypothesize mediators in a causal DAG given a treatment and effect.

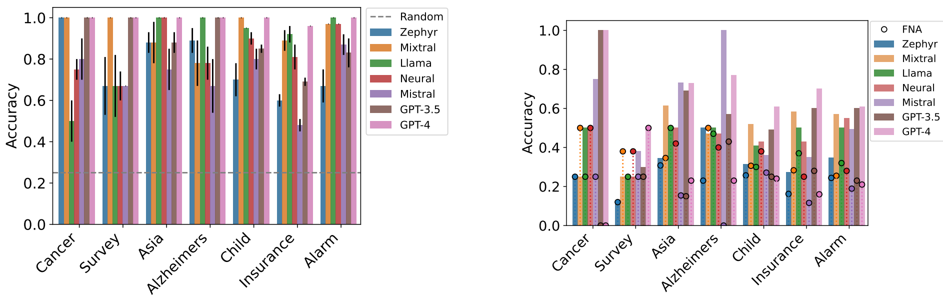
The task setup (Fig. 2b) involves a partial graph $\mathcal{G}^*$ with observed treatment and outcome variables, aiming to hypothesize mediators $M = \{v_{m_1}, \ldots, v_{m_H}\}$ that link treatment v_t to outcome v_y. In each iteration, the LLM hypothesizes a mediator, updates the graph, and uses the new partial graph to identify subsequent mediators.

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

To study the influence of mediator order, we introduce the Mediation Influence Score (MIS), which measures the importance of each mediator via the Natural Direct Effect (NDE) and Natural Indirect Effect (NIE). MIS quantifies the mediator's impact relative to the direct effect:

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

Mediators are generated based on MIS scores, prioritizing those with higher influence.



(a) Task 1 Result.

(b) Task 2 Result.

Figure 3: LLM accuracy in identifying missing causal variables from multiple choices with out-of-context (a) and in-context distractors (b).

3 Evaluation and Results

We evaluate a variety of causal datasets spanning diverse domains. We use the semi-synthetic datasets from BNLearn repository, see Appendix []. We evaluate our setups across different open-source and closed models.

3.1 Task 1: Out-of-Context Controlled Variable Identification

This task establishes a baseline to evaluate LLMs' fundamental abilities in causal reasoning with partial causal graphs. The input includes the ground truth variable name, out-of-context multiple choices for the missing variable v_x, and the partial DAG $\mathcal{G}^*$. The model's accuracy in identifying v_x is computed as:

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

Results. Figure 3a shows the accuracy of various LLMs in identifying missing variables. GPT-4 and Mixtral perform best, achieving perfect accuracy on most datasets, followed by GPT-3.5, except on Insurance and Alarm datasets. Models like Mistral, Llama-70, and Zephyr show varied success, with Insurance proving the most challenging, likely due to its complex DAG structure. All models outperform the random baseline, suggesting they can identify missing causal variables in a partial graph $\mathcal{G}^*$. However, this high accuracy may be driven by the task's simplicity, relying on dataset context rather than true causal reasoning. To probe deeper, the next task introduces in-domain choices to better evaluate LLMs' ability to discern causal variables beyond obvious correlations.

3.2 Task 2: In-Context Controlled Variable Identification

This task presents a more complex scenario to test LLMs' causal reasoning by incorporating two missing nodes in the partial graph. The input includes out-of-context choices, the ground truth variable, and one missing node as an in-context distractor, requiring the model to reason about indirect causal relationships.

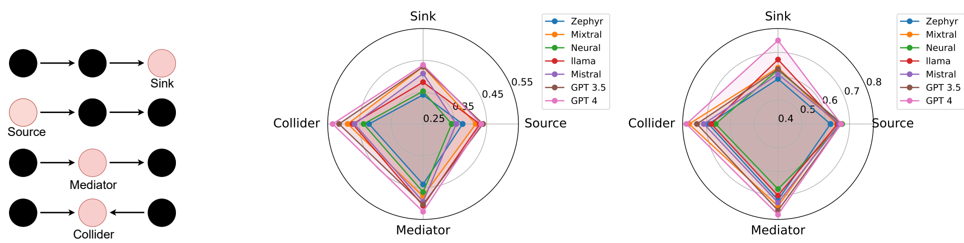
We evaluate performance using two metrics: accuracy and False Node Accuracy (FNA), the latter measuring confusion in selecting the in-context variable over the ground truth:

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

Results. Figure 3b shows both accuracy and FNA across datasets. Ideally, accuracy should be 1.0 and FNA 0.0, with random chance at 0.2. Most models on larger datasets exceed random performance. GPT-3.5 and GPT-4 consistently demonstrate high accuracy and low FNA, indicating their ability to reason causally and identify missing nodes without confusion from in-context distractors. In contrast, open-source models show variable performance. For example, Mistral excels on the Cancer dataset but struggles with the more complex Alarm dataset. Overall, most LLMs can identify causal variables even with multiple missing nodes and in-context distractions.

	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	0.61	0.34	0.60	0.45	0.66	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	0.31	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	0.44	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	0.45	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	0.34	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	0.39	**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**	**0.47**	0.98	0.36	0.53	**0.52**	**0.56**	0.49	**0.75**	**0.50**	**0.73**

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



(a) Types of nodes in a DAG. (b) Semantic similarity. (c) LLM-as-Judge.

Figure 4: Visualising each model's performances, averaged across the different datasets, for Sink, Source, Mediator, and Collider nodes.

3.3 Task 3: Hypothesizing in Open World

In this task, we simulate a scenario where a user provides a partial causal graph without multiple choices, expecting the LLM to complete the causal DAG by hypothesizing missing variables. The model is prompted for $k = 5$ suggestions for the missing node v_x.

To evaluate the suggestions, we employ two metrics: semantic similarity and LLM-as-Judge.
Semantic Similarity: Measures the cosine similarity of the model's suggestions $v^*_{x_{1:5}}$ with the ground truth v_x (details in Appendix B.4).
LLM-as-Judge: This metric assesses the quality of suggestions through a two-step process, comparing them against ground truth variables for contextual semantic similarity (details in B.5).
Results. Model performances using both metrics are reported in Table 1. We analyze each metric across different node types (sources, sinks, colliders, and mediators), with results shown in Figure 4. For detailed performance per dataset, see Figure 17.

GPT-4 and Mistral achieve higher semantic similarity and LLM-as-Judge scores across most datasets. GPT-3.5 also performs well. Semantic similarity is stricter than LLM-as-Judge, as it cannot fully capture contextual information (see example in Table 7). Both metrics show a fair correlation. Models perform better on colliders and mediators, indicating proficiency in reasoning about common causes and indirect relationships. However, they struggle with sink nodes, suggesting difficulty in reasoning about potential outcomes. Source nodes also present challenges, particularly in datasets like Survey and Alzheimer's. Additionally, model performance improves with more suggestions (k) and correlates with the number of edges in the graph, indicating that more context enhances reasoning abilities. Overall, LLMs show promise in hypothesizing mediators and colliders in a partial causal DAG, suggesting their potential utility in real-world applications.

3.4 Task 4: Iteratively Hypothesizing in Open World

In our previous experiment, we found that LLMs excel at identifying mediators when treatments and outcomes are provided. This is particularly relevant in medical settings, where understanding mediators can reveal causal mechanisms.

For unordered mediator evaluation, we hypothesize iteratively in random order. The evaluation mirrors the open-world setting, averaging semantic similarity across all mediators. For ordered

	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 2: Sim: semantic similarity for iteratively hypothesizing mediator nodes in random order. Δ measures prediction changes according to MIS.

evaluation based on the Mediation Influence Score (MIS), we introduce a metric Δ, which measures how the order of mediator realization influences predictions. We prompt the LLM in both ascending and descending orders of significance, calculating Δ as the change in semantic similarity. We focus on datasets with multiple mediators, including Asia, Child, Insurance, and Alarm, which range from 1 to 10 mediators. **Results.** Results are summarized in Table 11, with variances in Appendix D.1. In complex scenarios with multiple missing nodes, LLMs maintain performance. While GPT-4 shows consistent performance, it excels specifically in the Insurance and Alarm datasets. Positive Δ values indicate that using MIS for prompting enhances semantic similarity between hypotheses and ground truth. Overall, LLMs effectively hypothesize multiple mediators in a DAG, and incorporating domain knowledge about mediators can further boost performance.

3.5 Hypothesizing Confounder

In causal inference, backdoor paths can confound the estimation of causal effects, leading to bias if not properly addressed. Thus, hypothesizing and controlling for confounders is crucial. We extracted confounder subgraphs from the Sachs [32], Alarm, and Insurance graphs. As shown in Table 3, some confounders were easily hypothesized by LLMs, achieving perfect accuracy, while others, particularly in the genomic domain of Sachs, posed challenges due to potentially limited domain knowledge. Notably, GPT-4, while a large model, did not always perform best across all datasets, underscoring the necessity for diverse benchmarks to fully assess performance. While LLMs typically excel at hypothesizing colliders, results for confounders varied, highlighting the complexity of this task.

	Sachs	Alarm	Ins
Zephyr	0.10 ± 0.01	0.45 ± 0.05	0.53 ± 0.06
Mixtral	**0.95** ± 0.10	**0.85** ± 0.09	0.63 ± 0.07
Neural	0.30 ± 0.03	0.45 ± 0.05	0.61 ± 0.06
Llama	0.20 ± 0.02	0.47 ± 0.05	0.63 ± 0.06
Mistral	0.20 ± 0.02	**0.85** ± 0.09	0.61 ± 0.06
GPT-3.5	0.40 ± 0.04	0.49 ± 0.05	0.67 ± 0.07
GPT-4	**0.95** ± 0.10	0.73 ± 0.07	**0.78** ± 0.08

Table 3: Evaluating Confounders.

4 Conclusion

Most causality literature assumes necessary data is available and focuses on establishing causal relationships. Generating hypotheses about missing variables is usually done by human experts. We introduce a novel task where LLMs generate hypotheses about missing variables in causal graphs. Our formalized tests vary in difficulty and knowledge level, benchmarking models on identifying missing variables from various distractors and hypothesizing in open-world settings. Our findings suggest LLMs are effective for generating hypotheses, especially for mediators, which are often less known than treatments and outcomes.

5 Acknowledgements

This work was partially funded by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

[1] Ahmed Abdulaal, adamos hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *ICLR*, 2024. URL `https://openreview.net/forum?id=pAoqRlTBtY`.

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

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

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

[5] Ingo A Beinlich, Henri Jacques Suermondt, R Martin Chavez, and Gregory F Cooper. 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, 1989.

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

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

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

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

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

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

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

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

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

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

[16] 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. Mistral 7b. *arXiv*, 2023.

[17] 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. Mixtral of experts. *arXiv*, 2024.

[18] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing causal reasoning in language models. In *NeurIPS*, 2023.

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

[20] John Kendall. Designing a research project: randomised controlled trials and their principles. *Emergency medicine journal: EMJ*, 20(2):164, 2003.

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

[22] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[23] Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

[24] Steffen L Lauritzen and David J Spiegelhalter. 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, 1988.

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

[26] OpenAI. Gpt-4 technical report. *arXiv*, 2023.

[27] Judea Pearl. *Causality*. Cambridge university press, 2009.

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

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

[30] James Robins. 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, 1986.

[31] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, pages 1–3, 2023.

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

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

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

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

[36] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR*, 2023.

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

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

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

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

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

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

[43] Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. Hypothesis search: Inductive reasoning with language models. *arXiv*, 2023.

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

[45] Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really good logical reasoners? a comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841*, 2023.

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

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

[48] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *TMLR*, 2023.

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

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

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

A 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.

Sources are variables v_s with no incoming edges. Mathematically sources are $d_{\text{in}}(v_s) = 0$ where d_{in} is the in-degree of the graph.
Sinks are variables v_k with no outgoing edges. Sinks are $d_{\text{out}}(v_k) = 0$ where d_{out} is the out-degree of the graph.
Treatment are variables v_t, characterized as nodes $d_{\text{in}}(v_t) = 0$ that are being intervened upon.
Outcome are variables v_y, characterized as the nodes $d_{\text{out}}(v_y) = 0$ that are observed for interventions from the treatments.
Mediator are variables v_m that have both incoming and outgoing edges ($d_{\text{in}}(v_m) > 0$ and $d_{\text{out}}(v_m) > 0$), acting as intermediaries in the causal pathways between treatment and outcome. Hence v_k is a mediator if it is both a child of v_i and a parent of v_j.
Confounder are variables v_k that influence both treatment and outcome, exhibiting edges directed towards the treatment and outcome nodes ($d_{\text{out}}(v_k) \geq 2$. Hence v_k is a confounder if it is a parent of both v_i and v_j.
Collider are variables v_l that have two edges meeting, and have an in-degree greater than one $d_{\text{in}}(v_l) > 1$. Hence v_k is a collider if it is a child of both v_i and v_j.

Mediation Analysis. Mediation analysis quantifies the treatment's effect on the outcome through a mediator variable. This effect is decomposed into the Natural Direct Effect (NDE) and the Natural Indirect Effect (NIE). The NDE represents the treatment's effect on the outcome without mediation, while the NIE represents the effect mediated by the mediator variable. Futher explanation can be found in Appendix D.

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

B Implementation

B.1 Experimental setup

We evaluate a variety of causal datasets spanning diverse domains. We use the semi-synthetic datasets from BNLearn repository - Cancer:$\mathcal{G}(5, 4)$ [23], Survey:$\mathcal{G}(6, 6)$ [33], Asia:$\mathcal{G}(8, 8)$ [24], Child:$\mathcal{G}(20, 25)$ [35], Insurance:$\mathcal{G}(27, 52)$ [7], and Alarm:$\mathcal{G}(37, 46)$ [5]. We also evaluate our approach on a realistic Alzheimer's Disease dataset:$\mathcal{G}(9, 16)$ [1], developed by five domain experts. These datasets span across 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 [1], 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.

We evaluate our setups across different open-source and closed models. The models we use are GPT-3.5 [8], GPT-4 [26], LLama2-chat-7b [39], Mistral-7B-Instruct-v0.2 [16], Mixtral-7B-Instruct-v0.1 [17], Zephyr-7b-Beta [40] and Neural-chat-7b-v3-1 [15].

Dataset	V	E	Description
Cancer	5	4	Factors around lung cancer
Survey	6	6	Factors for choosing transportation
Asia	8	8	Factors affecting dysponea
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.

B.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.

GPT-3.5 GPT-4 were accessed via API. 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.

B.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.

B.4 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 Predisposit
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 Consciousne
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 Performan
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
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 Leve
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.

B.5 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 [51], 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($\mathcal{G}^*, V_{\text{GT}}, P$, LLM)
    Initialize $qualityRatings$ as an empty list
    for each node v_{GT} in **V do**
        $suggestions \leftarrow$ GenerateSuggestions($\mathcal{G}^*, P$, LLM)
        $bestSuggestion \leftarrow$ SelectBestSuggestion($suggestions, v_{\text{GT}}$, 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($suggestions, v_{\text{GT}}$,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.

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

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.

to judge how these biases might affect the final score. Hence for robust evaluation we also evaluate using the semantic similarity.

B.6 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

B.7 Related Work

Our work is based on the framework of causality as proposed by Pearl [27]. The intersection of language and causality is explored in [11, 13, 37**?**] to extract causal relationships from a large corpus of text. With the advancements in LLMs and their ability to process large contexts, there has been an interest in using them for causal reasoning [21]. Some works have focused on commonsense causality [9, 34] and temporal causal reasoning [50, 49]. More recently Kıcıman et al. [21], Long et al. [25**?**] introduced a method to discover causal structures by prompting LLMs with variable names. Ban et al. [4], Vashishtha et al. [41], Ban et al. [3] extended this work by introducing ancestral constraints to refine the causal structures derived from LLMs. Abdulaal et al. [1] combined data-based deep structural causal models, such as [47], with LLMs generated causal structure. Beyond using the ingested information for causal tasks, Jin et al. [19] focused on pure causal inference using LLMs. Recent work attempted to train causal transformers [**? ?**], however, in this work we aimed to test the hypothesizing abilities of generalist LLMs. In contrast to previous work, we focus on the

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

novel task of identifying and hypothesizing missing variables, a task that comes before data collection and evaluation, with LLMs as assistants. Additionally, existing works tested inductive hypothesis generation with LLMs [10, 28, 45, 46, 29], although, we look at causal hypothesis generation.

C 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

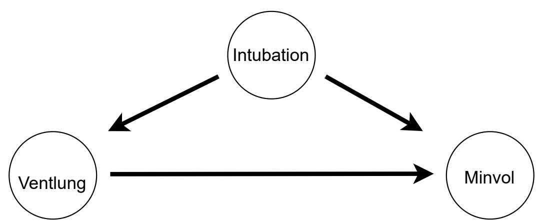


Figure 5: Alarm 1

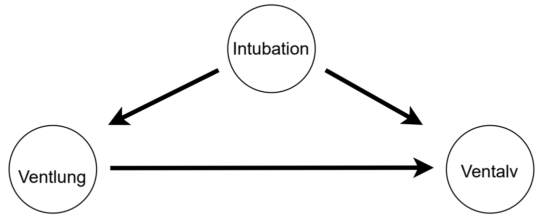


Figure 6: Alarm 2

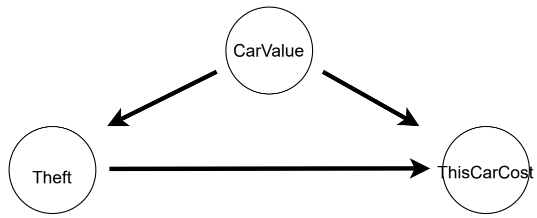


Figure 7: Insurance 1

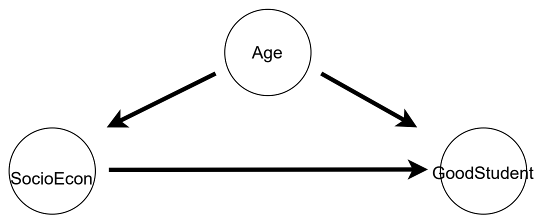


Figure 8: Insurance 2

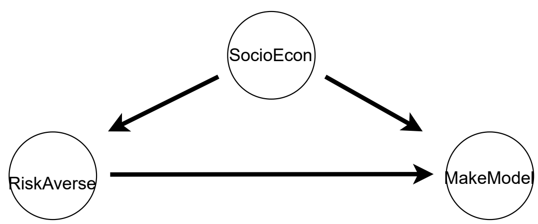


Figure 9: Insurance 3

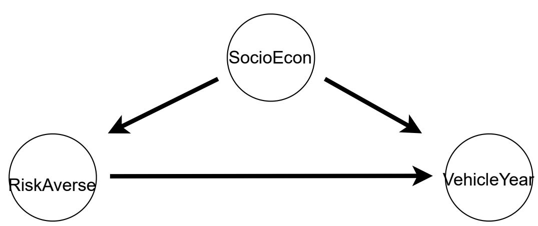


Figure 10: Insurance 4

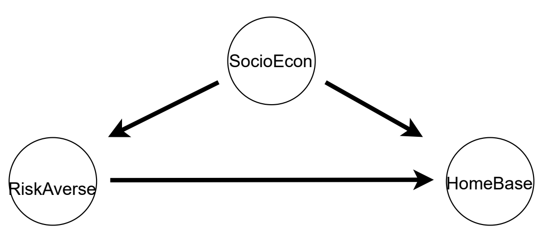


Figure 11: Insurance 5

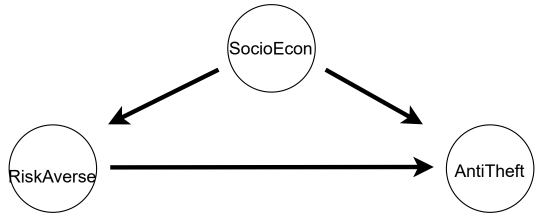


Figure 12: Insurance 6

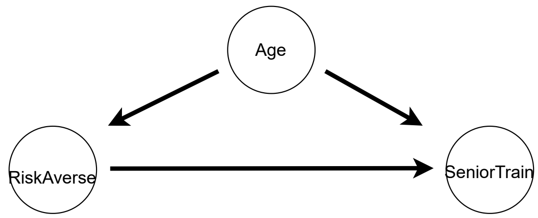


Figure 13: Insurance 7

D Further results

D.1 Variances

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

	Cancer		Survey		Asia		Alzheimers		Child		Insurance		Alarm		Avg	
	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J	Sim	LLM-J
Zephyr	0.36 ±0.04	0.61 ±0.06	0.34 ±0.07	0.60 ±0.05	0.45 ±0.05	0.66 ±0.04	0.35 ±0.03	0.75 ±0.03	0.51 ±0.02	0.70 ±0.04	0.45 ±0.04	0.44 ±0.05	0.46 ±0.03	0.69 ±0.02	0.42 ±0.04	0.63 ±0.04
Mixtral	0.41 ±0.03	0.66 ±0.04	0.39 ±0.05	0.66 ±0.06	**0.66** ±0.02	0.75 ±0.03	0.31 ±0.04	0.77 ±0.02	**0.53** ±0.03	**0.77** ±0.02	0.46 ±0.03	**0.56** ±0.04	**0.50** ±0.03	0.72 ±0.06	0.46 ±0.03	0.70 ±0.05
Neural	0.38 ±0.02	0.77 ±0.05	0.43 ±0.02	0.55 ±0.03	0.53 ±0.03	0.55 ±0.04	0.44 ±0.05	0.71 ±0.03	0.48 ±0.04	0.70 ±0.03	0.47 ±0.04	0.43 ±0.05	0.47 ±0.02	0.67 ±0.03	0.45 ±0.03	0.63 ±0.04
Llama	0.40 ±0.03	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.

D.2 Analysis of difference across tasks

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 14. 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.

	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.

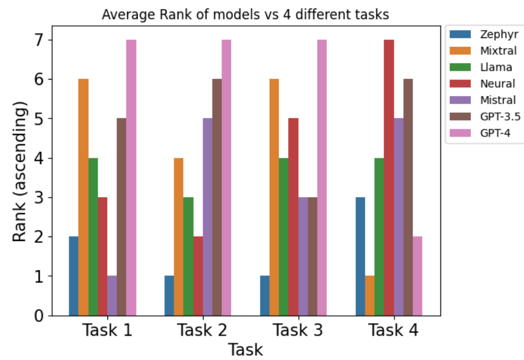
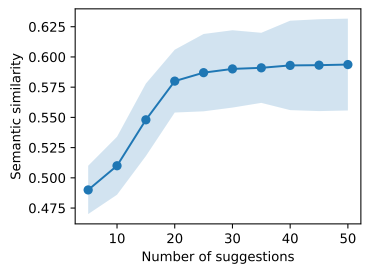
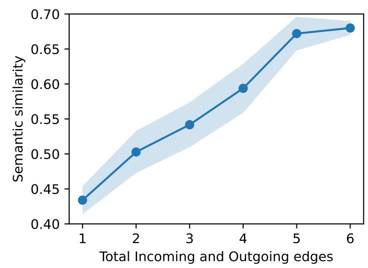


Figure 14: Average Rank of each model against the different tasks. We ranked the mode since the metrics are different to evaluate each task averaged across datasets

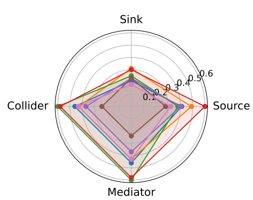


(a)

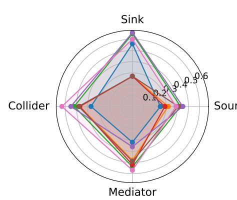


(b)

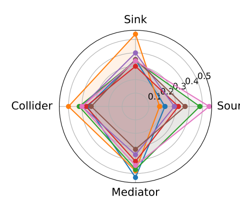
Figure 15: L: Plot of semantic similarity with an increasing number of suggestions for GPT-4 on the Alarm dataset. R: Plot of semantic similarity against the total number of incoming and outgoing edges for GPT-4 on the Alarm dataset.



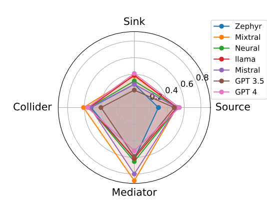
(a) Cancer



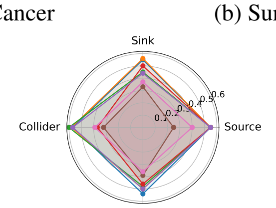
(b) Survey



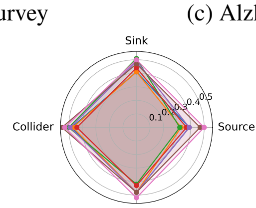
(c) Alzheimers



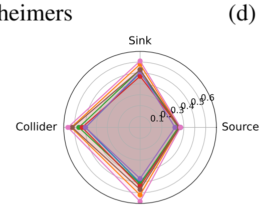
(d) Asia



(e) Child



(f) Insurance



(g) Alarm

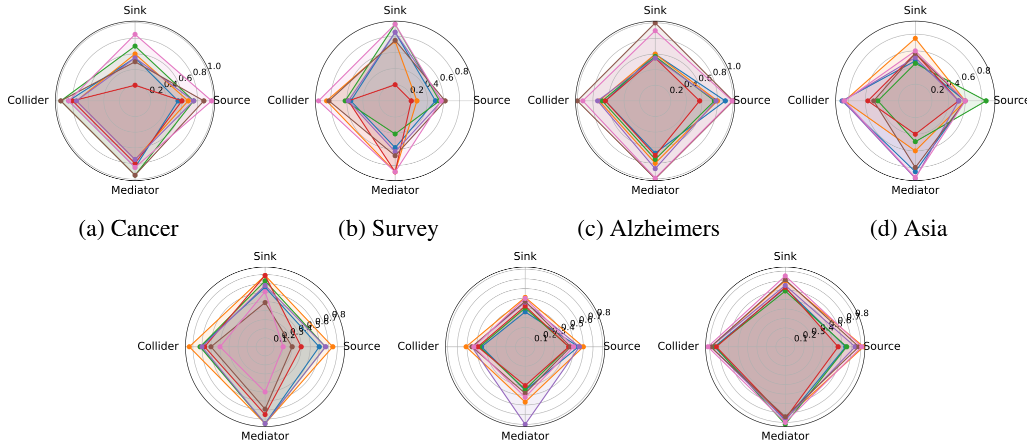
Figure 16: Detailed spider plots for Semantic similarity

D.3 Breaking down the performance

D.4 Fine grained model performance

D.5 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



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

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

Figure 17: Detailed spider plots for LLM-as-judge metric

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.

D.6 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.

D.7 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 [22] also in metadata-based causal reasoning [41]. 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.

D.8 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

E 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 15

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.

F Causal graphs

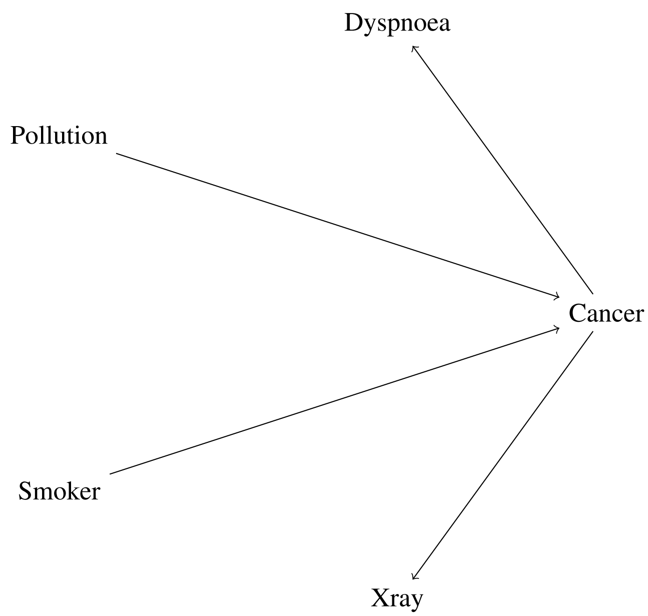


Figure 18: Cancer DAG

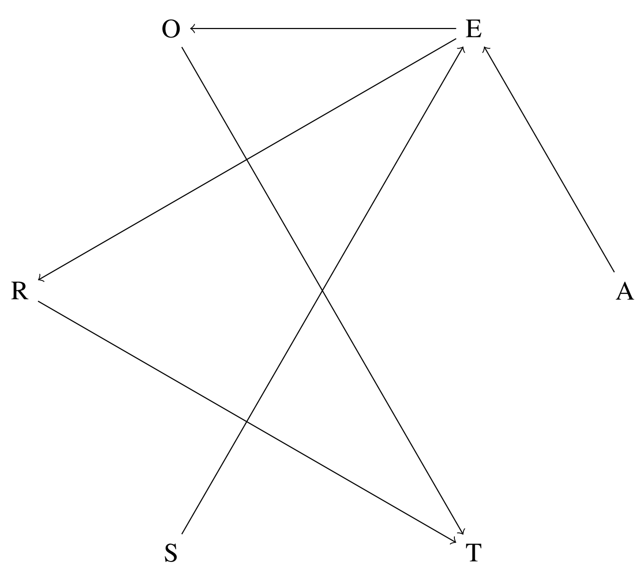


Figure 19: Survey DAG

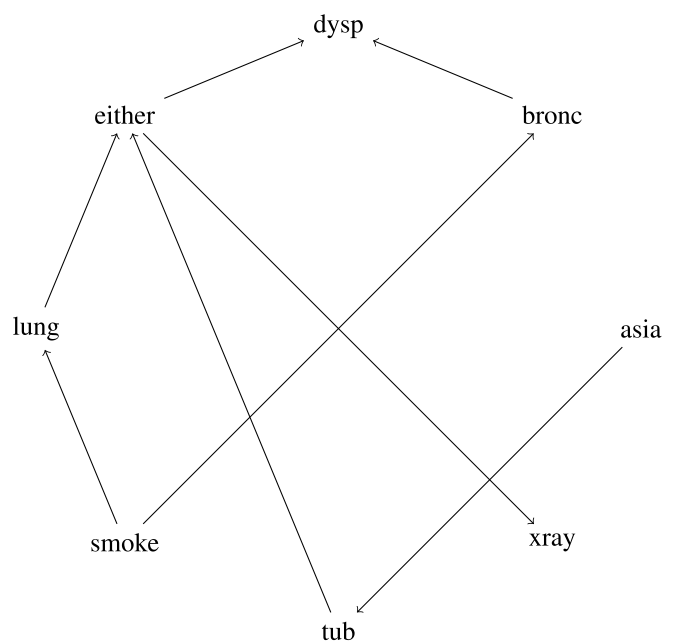


Figure 20: Asia DAG

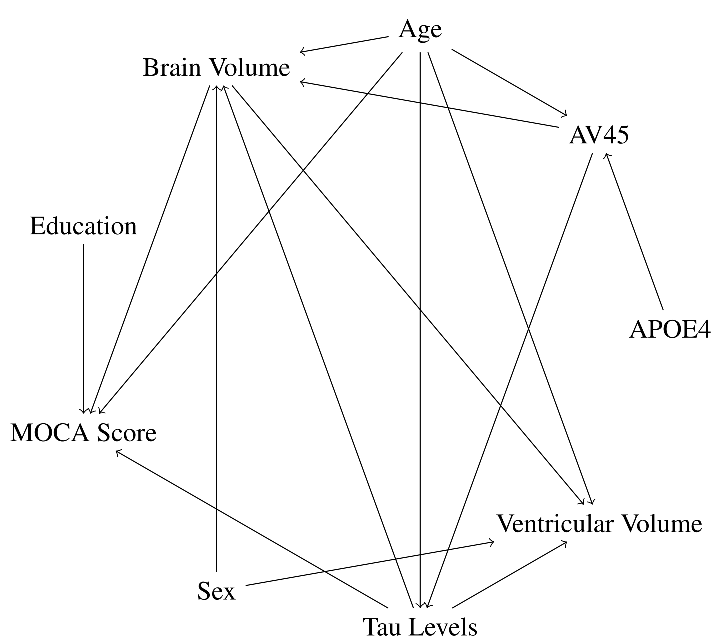


Figure 21: Alzheimer's DAG

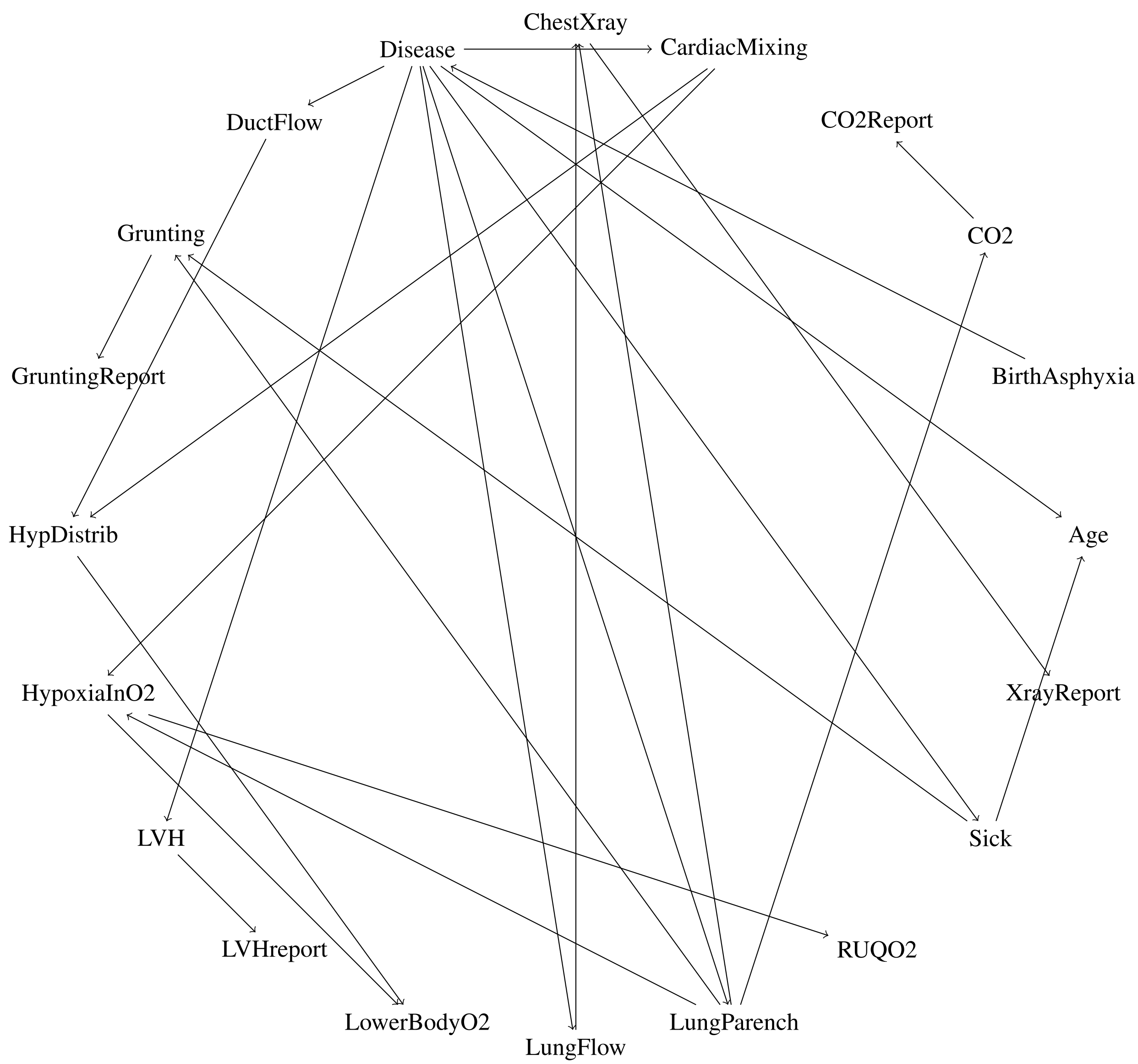


Figure 22: Child DAG

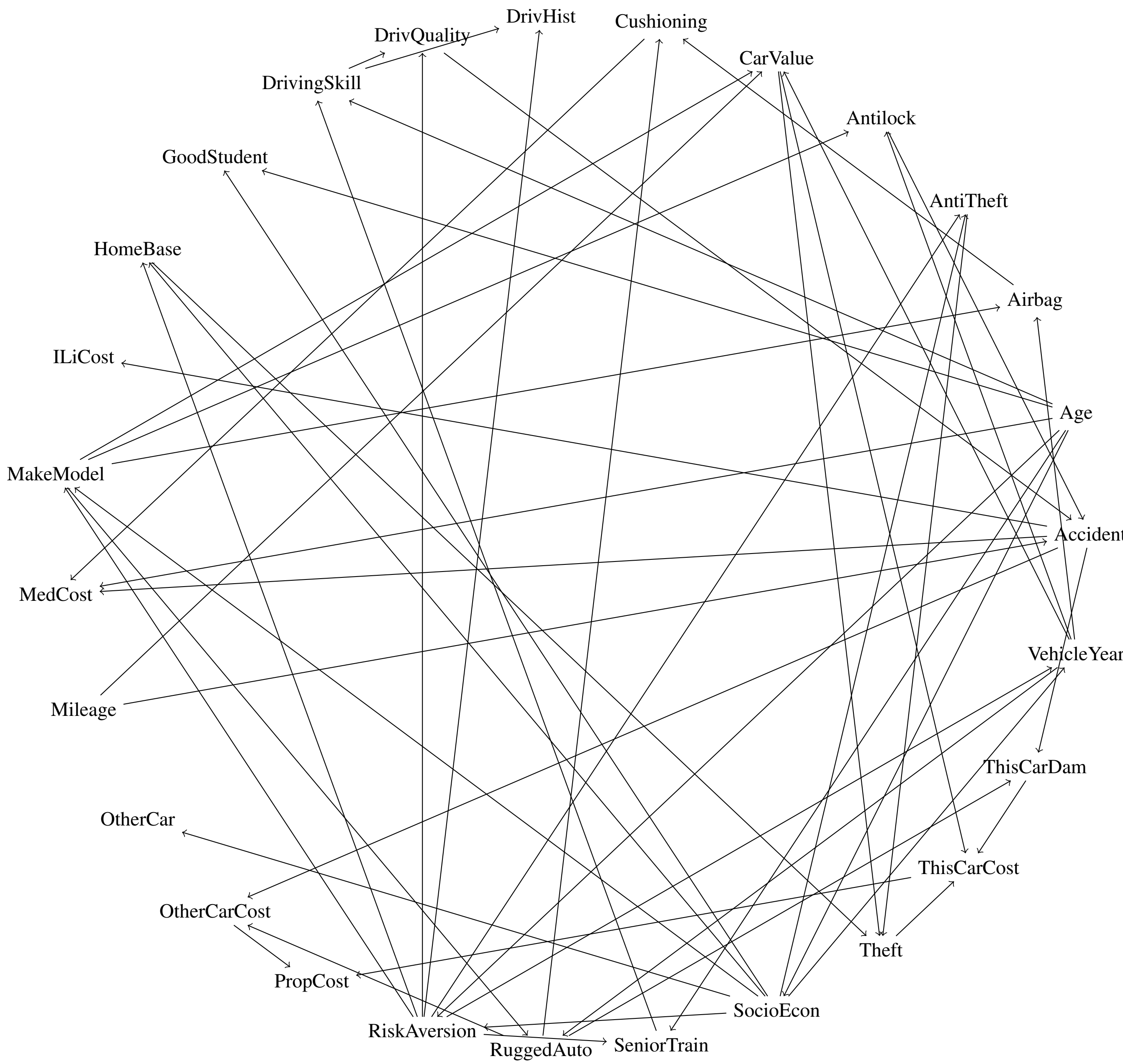


Figure 23: Insurance DAG

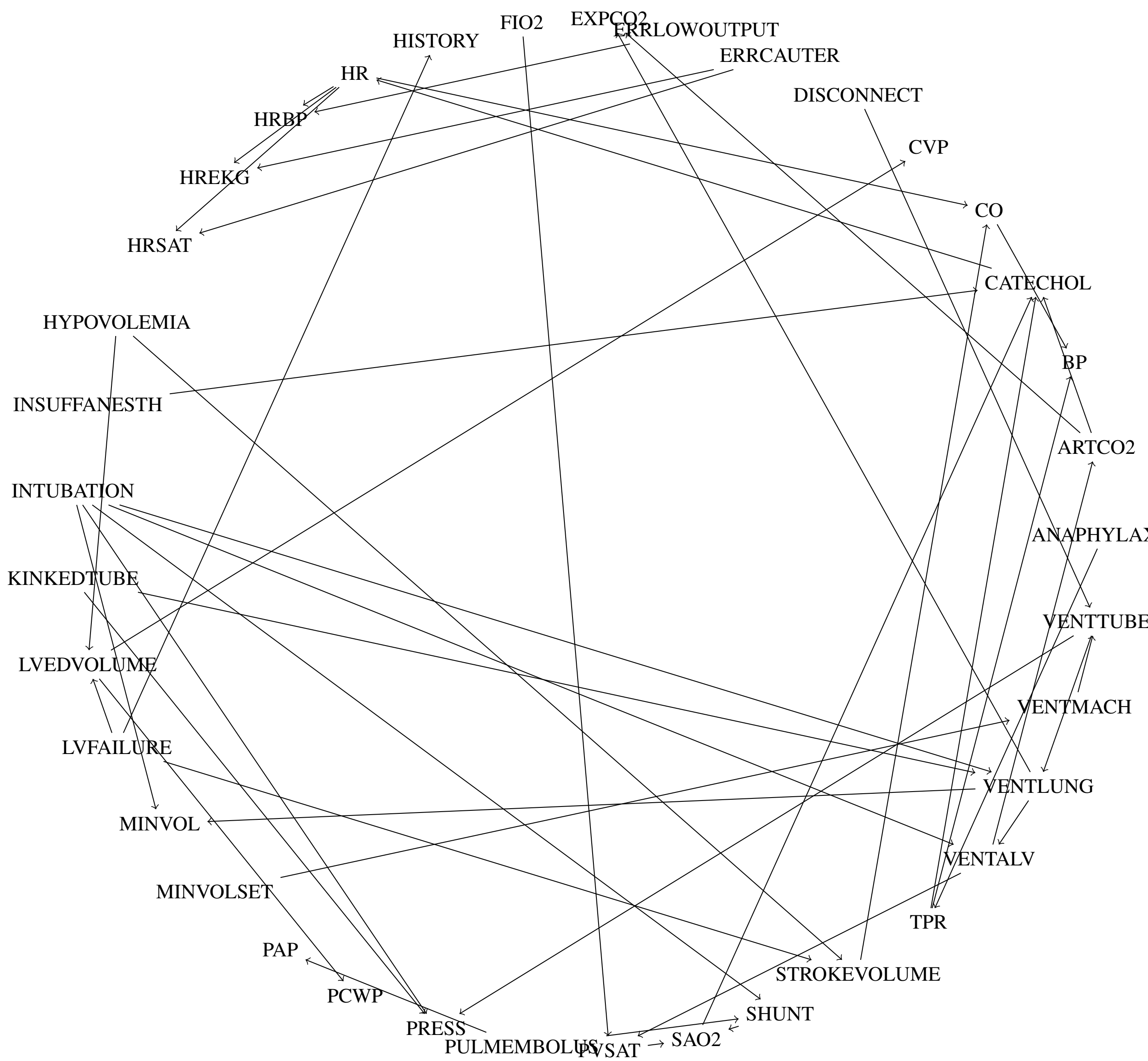


Figure 24: Alarm DAG

G 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.

H 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.

I NDE and NIE

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 part of the causal do-calculus introduced by [27]. We consider binary causal DAGs, i.e., each variable can either take 0 or 1 as values.

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

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

I.1 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.