Causal Discovery with Language Models as Imperfect Experts

Stephanie Long [*1] **Alexandre Piché** [*2 3 4] **Valentina Zantedeschi** [*2] **Tibor Schuster** [1] **Alexandre Drouin** [2 4]

Abstract

Understanding the causal relationships that underlie a system is a fundamental prerequisite to accurate decision-making. In this work, we explore how expert knowledge can be used to improve the data-driven identification of causal graphs, beyond Markov equivalence classes. In doing so, we consider a setting where we can query an expert about the orientation of causal relationships between variables, but where the expert may provide erroneous information. We propose strategies for amending such expert knowledge based on consistency properties, e.g., acyclicity and conditional independencies in the equivalence class. We then report a case study, on real data, where a large language model is used as an imperfect expert.

1. Introduction

Understanding the cause-and-effect relationships that underlie a complex system is critical to accurate decision-making. Unlike any statistical association, causal relationships allow us to anticipate the system's response to interventions. Currently, randomized control trials (RCTs) serve as the gold standard for establishing causation (Peters et al., 2017). However, RCTs can be costly and oftentimes impractical or unethical. As such, there has been growing interest in *causal discovery*, which aims to uncover causal relationships from data collected by passively observing a system (see (Glymour et al., 2019) for a review).

Causal discovery methods have been successfully applied in various fields, including genetics (Sachs et al., 2005) and climate science (Runge et al., 2019). Nevertheless, a fundamental limitation of such methods is their ability to only recover the true graph of causal relationships up to a set of equivalent solutions known as the *Markov equivalence class* (MEC), leading to uncertainty in downstream applications, such as estimating the effect of interventions (Maathuis et al., 2009).

One approach to reducing such uncertainty is the incorporation of expert knowledge, e.g., to rule out the existence of certain edges and reduce the set of possible solutions (Meek, 1995). However, such methods typically assume that the knowledge provided by the *expert is correct*. In this work, we consider a more realistic case, where the *expert is potentially incorrect*. Our approach leverages such *imperfect experts*, e.g., large language models, to reduce uncertainty in the output of a causal discovery algorithm by orienting edges, while maintaining fundamental consistency properties, such as the acyclicity of the causal graph and the conditional independencies in the MEC.

Contributions:

- We formalize the use of *imperfect experts* in causal discovery as an optimization problem that minimizes the size of the MEC while ensuring that the true graph is still included (Section 3).
- We propose a greedy approach that relies on Bayesian inference to optimize this objective by incrementally incorporating expert knowledge (Section 4).
- We empirically evaluate the performance of our approach, on real data, with an expert that returns correct orientations with some fixed probability (Section 5).
- We then empirically assess if the approach holds when taking a large language model as the expert – with mitigated results (Section 5).

2. Background and Related Work

We now review key background concepts and related work.

Causal Bayesian networks: Let $\mathbf{X} \coloneqq (X_1,...,X_d)$ be a vector of d random variables with distribution $p(\mathbf{X})$ and $G^\star \coloneqq \langle \mathcal{V}_{G^\star}, \mathcal{E}_{G^\star} \rangle$ be a directed acyclic graph (DAG) with vertices $\mathcal{V}_{G^\star} = \{v_1,...,v_d\}$ and edges $\mathcal{E}_{G^\star} \subset \mathcal{V}_{G^\star} \times \mathcal{V}_{G^\star}$. Each vertex $v_i \in \mathcal{V}_{G^\star}$ corresponds to a random variable X_i and a directed edge $(v_i, v_j) \in \mathcal{E}_{G^\star}$ represents a direct causal relationship from X_i to X_j. We assume that $p(\mathbf{X})$ is *Markovian* with respect to $G^\star$, i.e.,

$$p(X_1,...,X_d) = \prod_{i=1}^{d} p(X_i \,|\, \mathrm{pa}_i^{G^\star}),$$

[*]Equal contribution [1]McGill University [2]ServiceNow Research [3]Université de Montréal [4]Mila - Quebec AI Institute. Correspondence to: Valentina Zantedeschi <vzantedeschi@gmail.com>.

Accepted to ICML workshop on Structured Probabilistic Inference & Generative Modeling, Honolulu, Hawaii, USA. PMLR 202, 2023.

where $\text{pa}_i^{G^\star}$ denotes the parents of X_i in $G^\star$.

Causal discovery: This task consists of recovering $G^\star$ from data, which are typically sampled from $p(\mathbf{X})$ (Glymour et al., 2019). Existing methods can be broadly classified as being: i) constraint-based (Spirtes et al., 2000; 2013), which use conditional independence tests to rule-out edges, or ii) score-based (Chickering, 2002; Zheng et al., 2018), which search for the DAG that optimizes some scoring function. One common limitation of these approaches is their inability to fully identify the true underlying graph $G^\star$ beyond its *Markov equivalence class* (MEC) (Peters et al., 2017).

Equivalence classes: The MEC, $\mathcal{M}_{G^\star}$, is a set of graphs that includes $G^\star$ and all other DAGs with equivalent conditional independences. These may have different edge orientations, leading to uncertainty in downstream tasks, such as treatment effect estimation (Maathuis et al., 2009). One common approach to reducing the size of $\mathcal{M}_{G^\star}$ is to include interventional data (Eberhardt et al., 2005; Brouillard et al., 2020; Mooij et al., 2020). However, similar to RCTs, the collection of such data may not always be feasible or ethical. An alternative approach, which we adopt in this work, is to eliminate graphs that are deemed implausible by an expert.

Expert knowledge: Previous work has considered experts that give: (i) forbidden edges (Meek, 1995), (ii) (partial) orderings of the variables (Scheines et al., 1998; Andrews et al., 2020), (iii) ancestral constraints (de Campos & Castellano, 2007; Li & Beek, 2018; Chen et al., 2016), and (iv) constraints on interactions between types of variables (Brouillard et al., 2022) (see Constantinou et al. (2023) for a review). Typically, all DAGs that are contradicted by the expert are discarded, resulting in a new equivalence class $\mathcal{M}^E \subseteq \mathcal{M}_{G^\star}$. One pitfall is that, realistically, an expert is unlikely to always be correct, and thus, $G^\star$ might be discarded, i.e., $G^\star \notin \mathcal{M}^E$. In this work, we attempt to reduce $\mathcal{M}_{G^\star}$ as much as possible, while ensuring $G^\star \in \mathcal{M}^E$ with high probability, in the presence of *imperfect expert knowledge*. We note that our work is akin to Oates et al. (2017), but a key difference is that they assume a *directionally informed* expert, i.e., that cannot misorient edges in $G^\star$. Moreover, their approach is *expert-first*, i.e., data is used to expand an initial graph given by an expert, while our approach is *data-first*, i.e., the expert is used to refine the solution of a causal discovery algorithm.

Large language models: In situations where access to human experts is limited, Large Language Models (LLMs), such as GPT-4 (OpenAI, 2023a), offer promising alternatives. Recent studies have demonstrated that certain LLMs possess a rich knowledge base that encompasses valuable information for causal discovery (Choi et al., 2022; Long et al., 2023; Hobbhahn et al., 2022; Willig et al., 2022; Kıcıman et al., 2023; Tu et al., 2023), achieving state-of-the-art accuracy on datasets such as the Tübingen pairs (Mooij et al., 2016). In this work, we investigate the use of LLMs as *imperfect experts* within the context of causal discovery. Unlike prior approaches, which typically assume the correctness of extracted knowledge,[1] we propose strategies to use, potentially incorrect, LLM knowledge to eliminate some graphs in $\mathcal{M}_{G^\star}$, while ensuring that $G^\star \in \mathcal{M}^E$ with high probability.

3. Problem Setting

We now formalize our problem of interest. Let $G^\star$ represent the true causal DAG, as defined in Section 2, and let $\mathcal{M}_{G^\star}$ be its MEC. We assume that $\mathcal{M}_{G^\star}$ is known, e.g., that it has been obtained via some causal discovery algorithm. Further, we assume the availability of *metadata* $\{\mu_1,...,\mu_d\}$, where each μ_i provides some information about X_i, e.g., a name, a brief description, etc. We then assume access to an expert, who consumes such metadata and makes decisions:

Definition 3.1. (Expert) An expert is a function that, when queried with the metadata for a pair of variables (μ_i,μ_j), returns a hypothetical orientation for the $X_i - X_j$ edge:

$$E(\mu_i,\mu_j) = \begin{cases} \rightarrow & \text{if it believes that } (v_i,v_j) \in \mathcal{E}_{G^\star} \\ \leftarrow & \text{if it believes that } (v_j,v_i) \in \mathcal{E}_{G^\star} \end{cases}. \quad (1)$$

Of note, $E(\mu_i,\mu_j)$ can be incorrect (*imperfect expert*) and thus, our problem of interest consists in elaborating strategies to maximally make use of such imperfect knowledge.

Let $\mathcal{U}(\mathcal{M}_{G^\star})$ be the set of indices of all pairs of variables related by an edge whose orientation is ambiguous in $\mathcal{M}_{G^\star}$:

$$\mathcal{U}(\mathcal{M}_{G^\star}) := \{(i,j) \,|\, i<j \text{ and } \exists G,G' \in \mathcal{M}_{G^\star} \text{ s.t. } (v_i,v_j) \in \mathcal{E}_G \wedge (v_j,v_i) \in \mathcal{E}_{G'}\}. \quad (2)$$

We aim to elaborate a strategy S that uses the expert's knowledge to orient edges in $\mathcal{U}(\mathcal{M}_{G^\star})$ and obtain a new equivalence class $\mathcal{M}^{E,S}$, such that uncertainty is reduced to the minimum, i.e., $|\mathcal{M}^{E,S}| \ll |\mathcal{M}_{G^\star}|$, but $G^\star$ still belongs to $\mathcal{M}^{E,S}$ with high probability, that is:

$$\min \left|\mathcal{M}^{E,S}\right| \quad \text{such that } p\big(G^\star \in \mathcal{M}^{E,S}\big) \geq 1-\eta, \quad (3)$$

where $\eta \in [0,1]$ quantifies tolerance to the risk that the true graph $G^\star$ is not in the resulting equivalence class. This problem can be viewed as a trade-off between reducing uncertainty, by shrinking the set of plausible DAGs, and the risk associated with making decisions based on an imperfect expert.

4. Strategies for Imperfect Experts

Instead of blindly accepting expert orientations, we leverage the consistency information provided by the true MEC to estimate which decisions are most likely incorrect. Indeed,

[1] The Bayesian approach of Choi et al. (2022) is an exception.

among all possible combinations of edge orientations, only a few are possible, since many of them would create cycles or introduce new v-structures. The different strategies that we now propose for solving Problem (3) leverage such consistency imperatives, as well as Bayesian inference, to increase robustness to errors in expert knowledge.

Noise model: First, let us define the noise model that, we assume, characterizes mistakes made by the expert. Figure 1 shows the dependency graph for the decision process of a type of imperfect expert that we dub "ε-expert". For any pair $p_i = (p_{i1}, p_{i2}) \in \mathcal{U}(\mathcal{M}_{G^\star})$, we use the notation O_{p_i} to denote the *unknown* true edge orientation and $E_{p_i} := E(\mu_{p_{i1}}, \mu_{p_{i2}})$ denotes the orientation given by the expert. Further, for any subset of indices $I \subseteq \mathcal{U}(\mathcal{M}_{G^\star})$, we use $O_I := \{O_{p_i}\}_{i=1}^{|I|}$; the same applies to E_I. Notice that (i) true edge orientations are, in general, interdependent because of the aforementioned consistency properties of the MEC, and that (ii) edges already oriented in $\mathcal{M}_{G^\star}$ are not represented since they are constants (i.e., the expert is not queried for those). In this model, we assume that, for any $p_i \in \mathcal{U}(\mathcal{M}_{G^\star})$, the expert's response depends only on the true value O_{p_i}, i.e., $p(E_{p_i} \mid O_{\mathcal{U}(\mathcal{M}_{G^\star})}) = p(E_{p_i} \mid O_{p_i})$ and is incorrect with constant probability ε.

We now define the components of our Bayesian approach.

Prior: We consider a simple prior that encodes the knowledge given by the true MEC. It boils down to an uniform prior over the graphs in $\mathcal{M}_{G^\star}$, effectively assigning no mass to any edge combination that is not consistent (creates a cycle or a new v-structure). Thus, the prior for a partial edge orientation O_I corresponds to its frequency in the graphs of $\mathcal{M}_{G^\star}$:

$$p(O_I) = \sum_{\mathbf{o}_{\neg I}} p\big(O_I, O_{\mathcal{U}(\mathcal{M}_{G^\star}) \setminus I} = \mathbf{o}_{\neg I}\big),$$

where we marginalize over all possible combinations of values, $\mathbf{o}_{\neg I}$, for the remaining unoriented edges $O_{\mathcal{U}(\mathcal{M}_{G^\star}) \setminus I}$.

Posterior: The posterior probability that orientations for all edges in $\mathcal{U}(\mathcal{M}_{G^\star})$ are correct, given all observed expert decisions $E_{\mathcal{U}(\mathcal{M}_{G^\star})}$, is then given by:

$$p\big(O_{\mathcal{U}(\mathcal{M}_{G^\star})} \mid E_{\mathcal{U}(\mathcal{M}_{G^\star})}\big) = \frac{p\big(E_{\mathcal{U}(\mathcal{M}_{G^\star})} \mid O_{\mathcal{U}(\mathcal{M}_{G^\star})}\big)\, p\big(O_{\mathcal{U}(\mathcal{M}_{G^\star})}\big)}{p\big(E_{\mathcal{U}(\mathcal{M}_{G^\star})}\big)}, \quad (4)$$

where, for the ε-expert noise model, the likelihood is s.t.,

$$p(E_{\mathcal{U}(\mathcal{M}_{G^\star})} \mid O_{\mathcal{U}(\mathcal{M}_{G^\star})}) = \prod_{p_i \in \mathcal{U}(\mathcal{M}_{G^\star})} p(E_{p_i} \mid O_{p_i}).$$

In contrast, due to interdependencies between the true edge orientations, the posterior probability cannot similarly be factorized and, in general, $p(O_{p_i} \mid E_{\mathcal{U}(\mathcal{M}_{G^\star})}) \neq p(O_{p_i} \mid E_{p_i})$. Note that the posterior for a subset edges $I \subseteq \mathcal{U}(\mathcal{M}_{G^\star})$, e.g, oriented by an iterative strategy, can be obtained via simple marginalization. Finally, note that the posterior can be used to estimate $p\big(G^\star \in \mathcal{M}^{E,S}\big)$, since any mistake in orienting $p_i \in \mathcal{U}(\mathcal{M}_{G^\star})$ results in excluding $G^\star$ from $\mathcal{M}^{E,S}$.

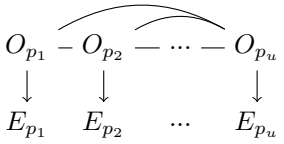

Figure 1. The ϵ-expert's dependency graph between true edge orientations (O_{p_i}) and expert decisions (E_{p_i}), where $u = |\mathcal{U}(\mathcal{M}_{G^\star})|$.

Greedy approach: We now propose a greedy strategy for optimizing Problem (3) that iteratively orients edges in $\mathcal{U}(\mathcal{M}_{G^\star})$. Let $\mathcal{M}^{(t)}$ denote the MEC at the t-th iteration of the algorithm and let $\mathcal{M}^{(t)}_{p_i}$ denote the MEC resulting from additionally orienting $p_i \in \mathcal{U}(\mathcal{M}^{(t)})$ at step t and propagating any consequential orientations using Meek (1995)'s rules. The algorithm starts with $\mathcal{M}^{(1)} = \mathcal{M}_{G^\star}$. We consider two strategies to greedily select the best p_i:

1. S_{size}: selects the edge that leads to the smallest equivalence class:
$$\operatorname*{argmin}_{p_i} |\mathcal{M}^{(t)}_{p_i}|$$
2. S_{risk}: selects the edge that leads to the lowest risk of excluding $G^\star$ from the equivalence class:
$$\operatorname*{argmin}_{p_i} \left[1 - p\left(O_{\mathcal{U}(\mathcal{M}_{G^\star}) \setminus \mathcal{U}\left(\mathcal{M}^{(t)}_{p_i}\right)} \mid E_{\mathcal{U}(\mathcal{M}_{G^\star})}\right)\right]$$

This procedure is repeated while $p\Big(G^\star \in \mathcal{M}^{(t)}_{p_i}\Big)$, estimated according to Equation (4), is greater or equal to $1 - \eta$.

5. Results and Discussion

We now evaluate the ability of our approach to leverage imperfect expert knowledge using real-world causal Bayesian networks from the *bnlearn* repository (Scutari, 2010).

Networks: We considered the following networks: (i) `Asia` (Lauritzen & Spiegelhalter, 1988), (ii) `ALARM` (Beinlich et al., 1989), (iii) `CHILD` (Spiegelhalter & Cowell, 1992), and (iv) `Insurance` (Binder et al., 1997). For each network, we extracted variable descriptions from the related publication and used them as metadata μ_i (see Appendix C).

Experts: We considered two kinds of expert: (i) ε-experts, as defined in Section 4, with various levels ε, and (ii) LLM-based experts based on GPT-3.5 (Ouyang et al., 2022). Details about prompting can be found at Appendix D. For each kind of expert, we considered both strategies: S_{size} and S_{risk}. Moreover, for the LLM-based expert, we also considered a naive strategy that consists of simply orienting all edges according to the expert, as in Long et al. (2023).

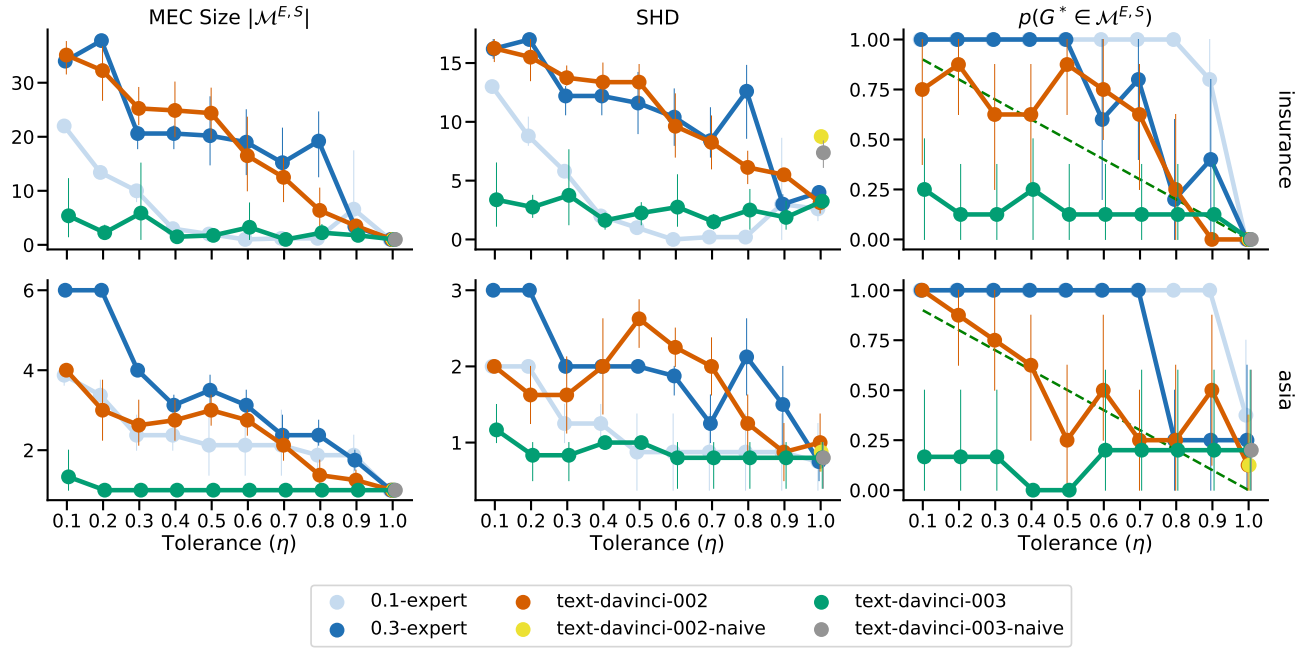

Figure 2. Results for `Insurance` and `Asia` and strategy S_{risk}. For the ε-experts, we consider $\varepsilon \in \{0.1, 0.3\}$. For the LLM-based experts, we consider the `text-davinci-{002, 003}` versions of GPT-3.5. We also report results for *naive* variants that follow Long et al. (2023) and do not make use of our greedy approach. Error bars show a 95% confidence interval. For all experts, except for `text-davinci-003`, both the MEC size and SHD decrease as tolerance is increased and $p(G^\star \in \mathcal{M}^{E,S}) \geq 1-\eta$.

Metrics: The expert/strategy combinations were evaluated based on: (i) the resulting size of their equivalence class, $|\mathcal{M}^{E,S}|$, (ii) the structural Hamming distance (SHD) between the completed partially DAG (CP-DAG; see Glymour et al. (2019)) of $\mathcal{M}^{E,S}$ and the true graph $G^\star$, (iii) an empirical estimate of $p(G^\star \in \mathcal{M}^{E,S})$, taken over repetitions of the experiment.

Protocol: For each Bayesian network, we extracted the MEC $\mathcal{M}_{G^\star}$ based on the structure of $G^\star$. This simulates starting from the ideal output of a causal discovery algorithm. We then attempted to reduce the size of $\mathcal{M}_{G^\star}$ by querying each expert according to the greedy approach in Section 4 for $\eta \in [0.1, 1]$, where $\eta = 1$ corresponds to disregarding the constraint of Problem 4. Each $\varepsilon-$expert experiment was repeated 5 times and LLM-expert experiment was repeated 8 times. Given that the LLM-experts have a deterministic output for a given prompt, we randomized the causation verb in order to introduce stochasticity (see Appendix D). Figure 2 shows the results of our experiments for `Insurance` and `Asia` with strategy S_{risk}. Results for other networks and strategy S_{size} are in Appendix A.

Results for ε-experts: On all networks, our approach, combined with both strategies, decreases the MEC's size for all noise levels (ε), while keeping the true graph in $\mathcal{M}^{E,S}$ with probability at least $1-\eta$, as predicted by our theoretical results. Consequently, the SHD also decreases as the tolerance to risk increases. This highlights the effectiveness of our approach when the expert satisfies the noise model of Section 4.

Results for the LLM-expert: Overall, we observe a clear reduction in SHD for $\mathcal{M}^{E,S}$ compared to the starting point $\mathcal{M}_{G^\star}$. This shows that some causally-relevant knowledge can be extracted from LLMs, which is in line with the conclusions of recent work.

On all datasets, the LLM-based experts achieve SHDs that are on par or better than those of their naive counterparts (Long et al., 2023) for $\eta = 1$, while additionally enabling the control of the probability of excluding $G^\star$. Further, on every dataset, except for `ALARM`, each LLM-based expert performs comparably to at least one of the ε-experts. For `ALARM`, we observe that $G^\star$ is excluded from $\mathcal{M}^{E,S}$, even for small tolerance η. This can be explained by ambiguities in the metadata, which are sometimes ambiguous even for human experts (see Appendix C).

Finally, another key observation is the poor uncertainty calibration of `text-davinci-003` compared to `text-davinci-002`, which is in line with observations made by OpenAI (2023b). The `text-davinci-003` model is often over-confident in its answers, which leads it to underestimate the probability of excluding $G^\star$ from $\mathcal{M}^{E,S}$. Consequently, even for small tolerance η, the resulting equivalence classes contain incorrectly oriented edges.

6. Conclusion

This work studied how imperfect expert knowledge can be used to refine the output of causal discovery algorithms. We proposed a greedy algorithm that iteratively rejects graphs from a MEC, while controlling the probability of excluding the true graph. Our empirical study revealed that our approach is effective when combined with experts that satisfy our assumptions. However, its performance was mitigated when a LLM was used as the expert. Nevertheless, our results show the clear potential of LLMs to aid causal discovery and we believe that further research in this direction is warranted. Possible extensions to this work include the exploration of noise models better-suited for LLMs, as well as alternative methods for querying such models (e.g., different prompt styles, better uncertainty calibration, etc.,). Further, our approach could be coupled with Bayesian causal discovery methods, replacing our MEC-based prior by one derived from a learned posterior distribution over graphs.

References

Andrews, B., Spirtes, P., and Cooper, G. F. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4002–4011. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/andrews20a.html.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Beinlich, I. A., Suermondt, H. J., Chavez, R. M., and Cooper, G. F. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. pp. 247–256, 1989.

Binder, J., Koller, D., Russell, S., and Kanazawa, K. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2-3):213–244, 1997.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Brouillard, P., Taslakian, P., Lacoste, A., Lachapelle, S., and Drouin, A. Typing assumptions improve identification in causal discovery. In *Conference on Causal Learning and Reasoning*, pp. 162–177. PMLR, 2022.

Chen, E. Y.-J., Shen, Y., Choi, A., and Darwiche, A. Learning bayesian networks with ancestral constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Choi, K., Cundy, C., Srivastava, S., and Ermon, S. Lmpriors: Pre-trained language models as task-specific priors. *arXiv preprint arXiv: 2210.12530*, 2022.

Constantinou, A. C., Guo, Z., and Kitson, N. K. The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, pp. 1–50, 2023.

de Campos, L. M. and Castellano, J. G. Bayesian network learning algorithms using structural restrictions. *International Journal of Approximate Reasoning*, 45(2):233–254, 2007.

Eberhardt, F., Glymour, C., and Scheines, R. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Conference on Uncertainty in Artificial Intelligence*, 2005.

Glymour, C., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00524.

Hobbhahn, M., Lieberum, T., and Seiler, D. Investigating causal understanding in llms. 2022.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z. H., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Lauritzen, S. L. and Spiegelhalter, D. J. Local computation with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 50(2):157–224, 1988.

Li, A. and Beek, P. Bayesian network structure learning with side constraints. In *International conference on probabilistic graphical models*, pp. 225–236. PMLR, 2018.

Long, S., Schuster, T., and Piché, A. Can large language models build causal graphs? *arXiv preprint arXiv: 2303.05279*, 2023.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A): 3133 – 3164, 2009. doi: 10.1214/09-AOS685. URL https://doi.org/10.1214/09-AOS685.

Meek, C. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pp. 403–410, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016. URL http://jmlr.org/papers/v17/14-518.html.

Mooij, J. M., Magliacane, S., and Claassen, T. Joint causal inference from multiple contexts. *The Journal of Machine Learning Research*, 21(1):3919–4026, 2020.

Oates, C., Kasza, J., Simpson, J., and Forbes, A. Repair of partly misspecified causal diagrams. *Epidemiology*, 28, 2017.

OpenAI. Gpt-4 technical report, 2023a.

OpenAI, R. Gpt-4 technical report. *arXiv*, pp. 2303–08774, 2023b.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

Scutari, M. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. doi: 10.18637/jss.v035.i03.

Spiegelhalter, D. J. and Cowell, R. G. Learning in probabilistic expert systems. pp. 447–466, 1992.

Spirtes, P., Glymour, C., and Scheines, R. Constructing bayesian network models of gene expression networks from microarray data. 2000.

Spirtes, P. L., Meek, C., and Richardson, T. S. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

Tu, R., Ma, C., and Zhang, C. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. 2023.

Willig, M., Zečević, M., Dhami, D. S., and Kersting, K. Can foundation models talk causality? *arXiv preprint arXiv:2206.10591*, 2022.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

A. Additional Experimental Results

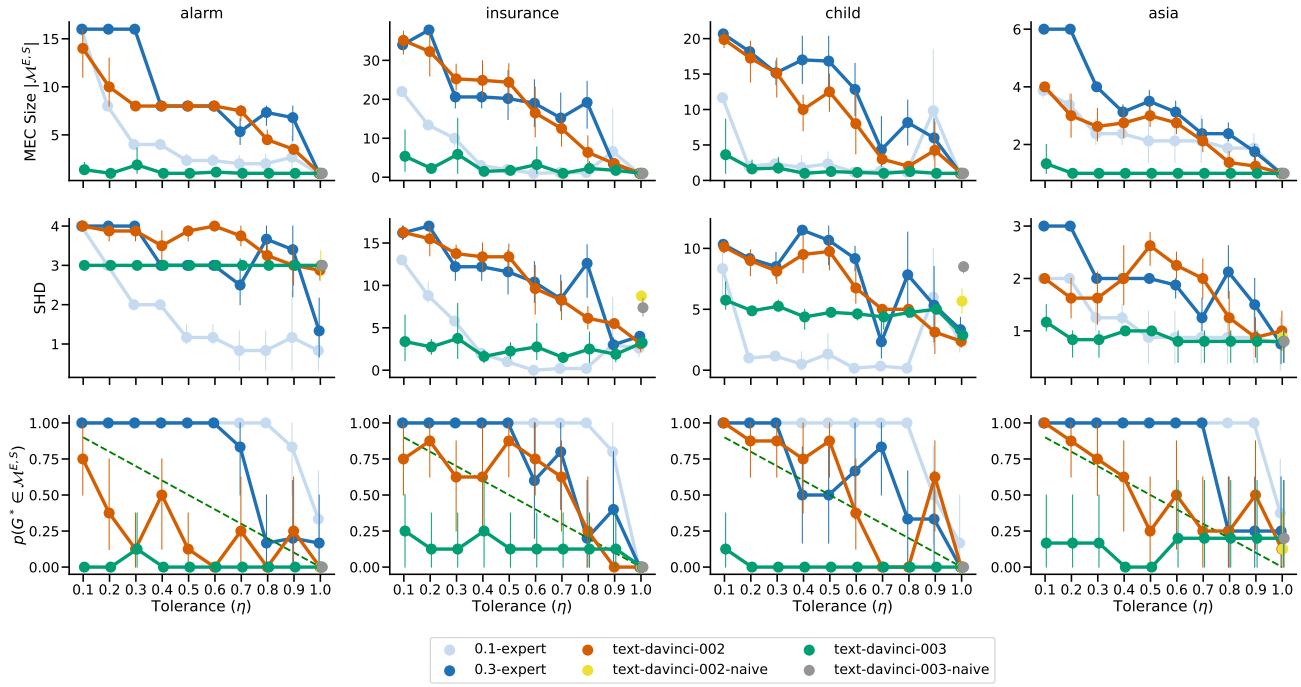

Figure 3. All results for strategy S_{risk}. We observe that the MEC size consistently decreases as the tolerance level is increased.

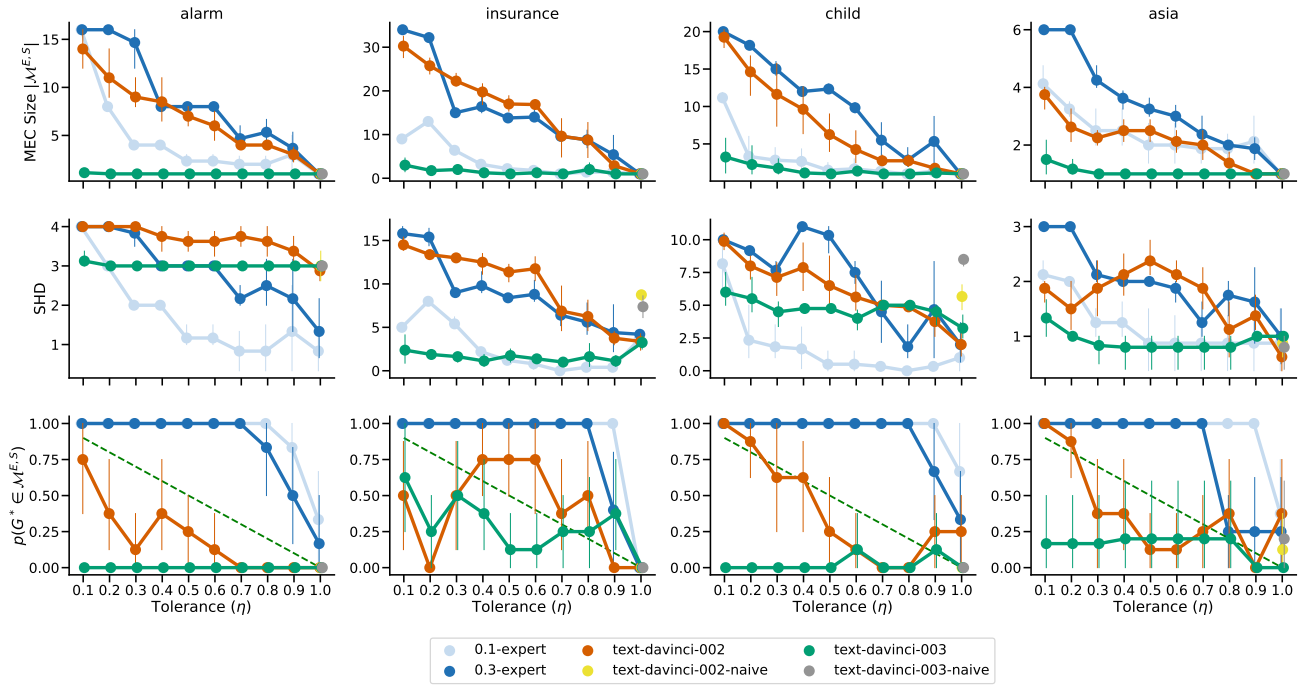

Figure 4. All results for strategy S_{size}. We observe that the MEC size consistently decreases as the tolerance level is increased.

Table 1. Characteristics of included causal networks

Dataset	# Nodes	# Edges	Parameters
Asia	8	8	18
CHILD	20	25	230
Insurance	27	52	1008
ALARM	37	46	509

B. Implementation details

The code for this work is available at https://github.com/StephLong614/Causal-disco.

C. Details for the BnLearn causal Bayesian networks

Acquisition: All Bayesian network structures were acquired from https://www.bnlearn.com/bnrepository/.

Variable metadata: For each of the included causal Bayesian networks, we extracted a *code book*, i.e., a list of variable names and an associated description (e.g., 'birth asphyxia': 'lack of oxygen to the blood during birth'), from the associated original paper. For instance, for the `CHILD` network, this information was extracted from Spiegelhalter & Cowell (1992). All code books are available at: https://github.com/StephLong614/Causal-disco/tree/main/codebooks.

Metadata pitfalls: Certain Bayesian networks contain edge orientations between variable pairs that appear incongruent with intuitive reasoning. For example, in the CHILD Network (Figure 6), the edge orientation between *disease* and *age* exhibits a counterintuitive direction: $disease \rightarrow age$. Implying the causal relationship of "disease causes age" rather than the more intuitive and expected "age causes disease".

D. Details for LLM-based experts

D.1. Querying for edge orientations

In order to obtain a probability distribution over the orientations of an edge, we use a prompt similar to Bai et al. (2022). We use the following prompt format:

```
 Among these two options which one is the most likely true:
(A) {μ_i} {verb_k} {μ_j}
(B) {μ_j} {verb_k} {μ_i}
The answer is:
```

where verb_k is randomized at each decision and the variables .

For example, if we wanted to elicit a prediction for the direction of an edge between variables with metadata μ_i: "lung cancer", μ_j: "cigarette smoking", and causation verb verb_k: "causes" we would use the following prompt:

```
 Among these two options which one is the most likely true:
(A) lung cancer causes cigarette smoking
(B) cigarette smoking causes lung cancer'
The answer is:
```

We then compute the log probability of the responses `(A)` and `(B)`, and use the softmax to obtain a probability distribution over the directions of the edge (Kadavath et al., 2022). Since we rely on scoring, instead of generation, the output of the LLM-expert is deterministic given a fixed prompt. To foster randomness in the LLM-expert outputs, we randomly draw verb_k from the following verbs of causation: provokes, triggers, causes, leads to, induces, results in, brings about, yields, generates, initiates, produces, stimulates, instigates, fosters, engenders, promotes, catalyzes, gives rise to, spurs, and sparks.

E. Causal Bayesian networks included

We included the following causal Bayesian networks in this work. All were extracted from the *bnlearn* Repository https://www.bnlearn.com/bnrepository/.

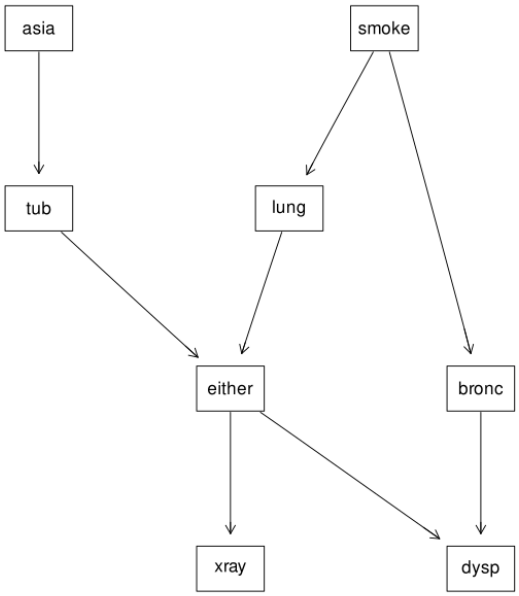

Figure 5. **Asia** Bayesian network representing a fictitious medical illustrating possible causes of shortness-of-breath (dyspnoae) (Lauritzen & Spiegelhalter, 1988). Abbreviations: *asia = visit to Asia?; tub = Tuberculosis; either = either tuberculosis or lung cancer; lung = lung cancer; bronc = bronchitis; dysp = dyspnoae.*]

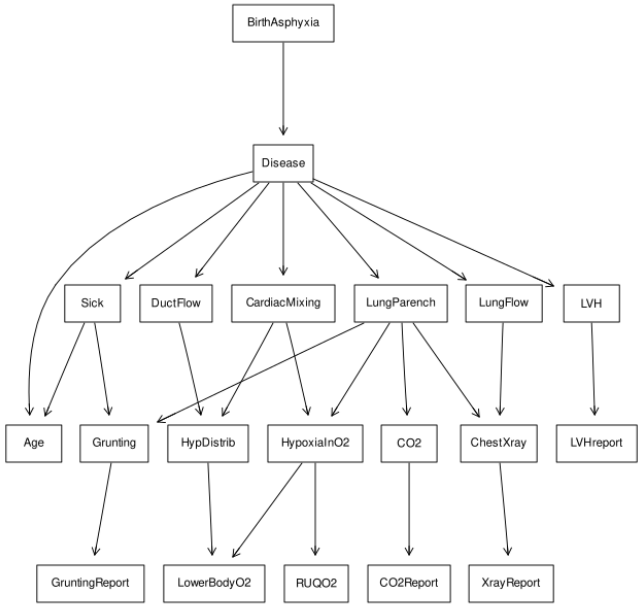

Figure 6. **CHILD** Bayesian network which represents the presentation of six possible conditions that lead to "blue babies" i.e., birth asphyxia (Spiegelhalter & Cowell, 1992). Abbreviations: *LungParench = Lung parenchyma, LVH = left ventricular hypertrophy; HypDistrib = hypoxia distribution; RUQO2 = right upper quad oxygen level.*

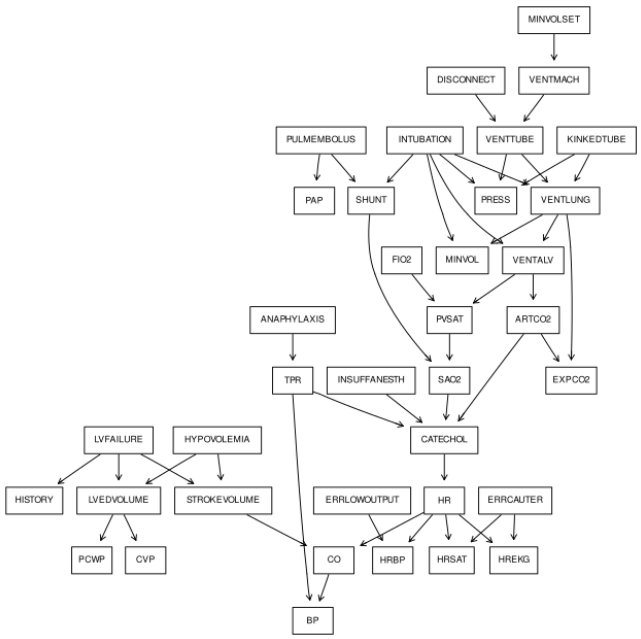

Figure 7. **ALARM** Bayesian network representing a diagnostic application for patient monitoring which includes 8 diagnoses, 16 findings, and 13 intermediate variables (Beinlich et al., 1989). Abbreviations: *MINVOLSET = minute ventilation; VENTMACH = ventilation machine; PULMEMBOLOUS = pulmonary embolism; PAP = pulmonary artery pressure; FIO2 = fraction of inspired oxygen; MINVOL = minute volume; VENTALV = alveolar ventilation; PVSAT = pulmonary artery oxygen saturation ; ARTCO2 = arterial CO2; TPR = total peripheral resistance; SAO2 = oxygen saturation; EXPCO2 = expelled CO2; LVFAILURE = left ventricular failure; CATECHOL = catecholamine; LVEDVOLUME = left ventricular end-diastolic volume; HR = heart rate; ERR = error; PCWP = pulmonary capillary wedge pressure; CVP = central venous pressure; CO = cardiac output; HRBP = rate blood pressure*; HRSAT = heart rate saturation

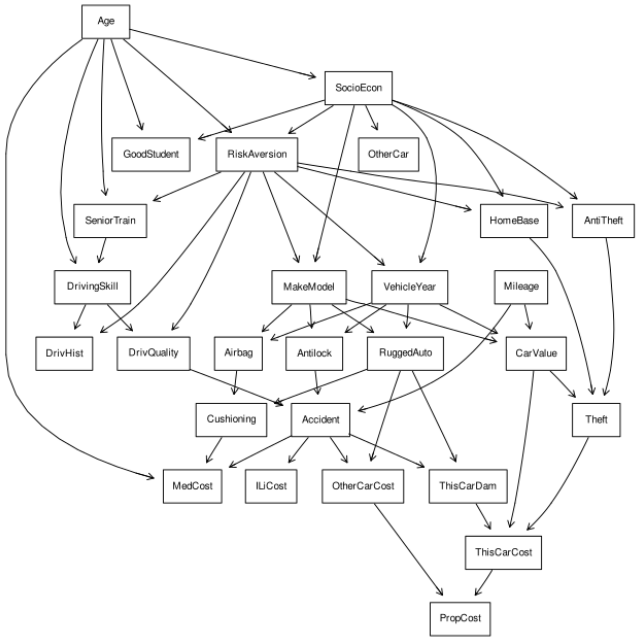

Figure 8. **Insurance** Bayesian network illustrating factors that affect expected claim costs for a car insurance policyholder (Binder et al., 1997). Abbreviations: *DrivHist = driving history; ILiCost = insurance liability cost; PropCost = property cost*