
Can Large Language Models Help Experimental Design for Causal Discovery?

Junyi Li^{*1} Yongqiang Chen^{*23} Chenxi Liu⁴ Qianyi Cai¹ Tongliang Liu⁵ Bo Han⁴ Kun Zhang²³
Hui Xiong¹⁶

Abstract

Designing proper experiments and selecting optimal intervention targets is a longstanding problem in scientific or causal discovery. Identifying the underlying causal structure from observational data alone is inherently difficult. Obtaining interventional data, on the other hand, is crucial to causal discovery, yet it is usually expensive and time-consuming to gather sufficient interventional data to facilitate causal discovery. Previous approaches commonly utilize uncertainty or gradient signals to determine the intervention targets. However, numerical-based approaches may yield suboptimal results due to the inaccurate estimation of the guiding signals at the beginning when with limited interventional data. In this work, we investigate a different approach, whether we can leverage Large Language Models (LLMs) to assist with the intervention targeting in causal discovery by making use of the rich world knowledge about the experimental design in LLMs. Specifically, we present Large Language Model Guided Intervention Targeting (LeGIT) – a robust framework that effectively incorporates LLMs to augment existing numerical approaches for the intervention targeting in causal discovery. Across 4 realistic benchmark scales, LeGIT demonstrates significant improvements and robustness over existing methods and even surpasses humans, which demonstrates the usefulness of LLMs in assisting with experimental design for scientific discovery.

1. Introduction

Science originates along with discovering new causal knowledge with *interventional experiments inspired by observations* (Kuhn & Hawkins, 1963). The art of finding causal

^{*}Equal contribution ¹The Hong Kong University of Science and Technology (Guangzhou) ²MBZUAI ³CMU ⁴HKBU ⁵USYD ⁶HKUST. Correspondence to: Hui Xiong <xionghui@ust.hk>.

relations from different interventions is then summarized and improved with statistical methods (Pearl & Mackenzie, 2018; Spirtes et al., 2010; Glymour et al., 2019). Identifying and utilizing causal relations is fundamental to numerous applications, including biology (Vowels et al., 2022) and financial systems (Dong et al., 2023). Despite the wide deployment of causal discovery methods, uncovering the underlying causal connections merely based on observational data alone is typically challenging due to limitations in identifiability. It usually requires additional interventional data obtained by perturbing part of the causal system to overcome the issue (Spirtes et al., 2000).

However, collecting interventional data is expensive and time-consuming, as it usually involves a physical process of a real-world system (Cherry & Daley, 2012). Consequently, *both the number of samples and the intervention targets are significantly limited in the experimental design in the real world* (Tong & Koller, 2001). Previous approaches usually rely on uncertainty (Lindley, 1956) or information theoretic metric to maximize the utility of an experiment (Tigas et al., 2022). Recently, leveraging gradient signals for intervention targeting has gained significant success (Olko et al., 2023), as it naturally fits into various gradient-based causal discovery methods. Despite some success, both uncertainty-based and gradient-based approaches may still suffer from suboptimality, as the estimation of the signals is usually noisy. Especially when with limited interventional data, the inaccurate estimation of the scores can easily mislead the intervention targeting and the subsequent causal discovery. The emergence of large language models (LLMs) (OpenAI, 2023) provides an opportunity to incorporate rich world knowledge about experimental design into the intervention targeting process, and raises an intriguing research question:

Can we leverage LLMs for intervention targeting and do LLMs really help with it?

Recent explorations into the use of LLMs for various causal learning and reasoning tasks suggest that these models may already encapsulate substantial domain knowledge (Kiciman et al., 2023; Lampinen et al., 2023; Abdulaal et al., 2024; Li et al., 2024). LLMs have demonstrated the ability to process the meta-information encoded in natural language and leverage the meta-information to reason for the causal-

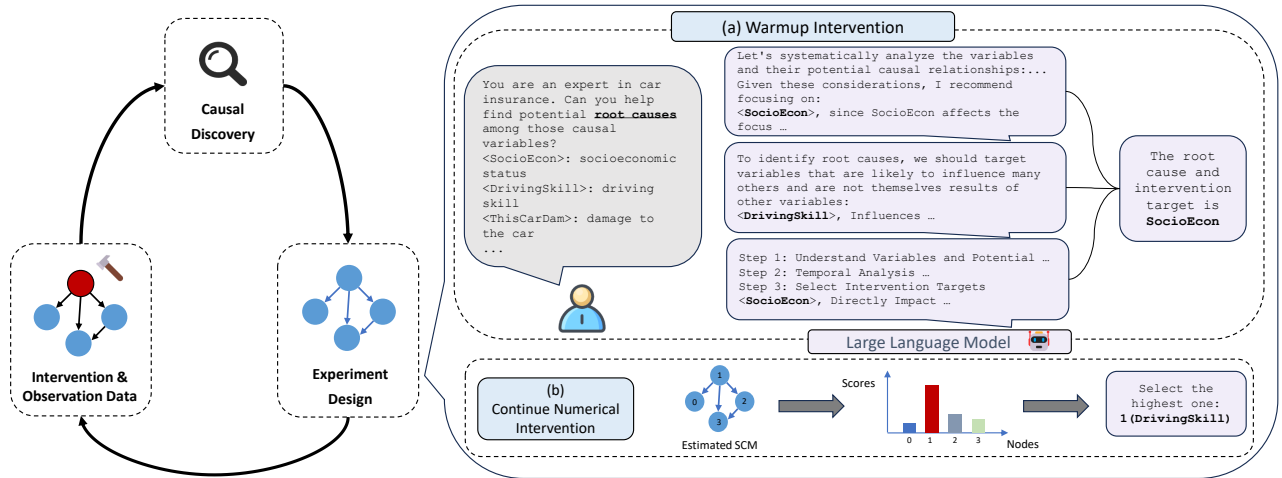


Figure 1. Illustration of the LeGIT framework. The left side represents the loop of Online Causal Discovery, while the right side illustrates the experiment design process. In Step (a), Large Language Models (LLMs) warm up the causal discovery process by leveraging world knowledge and aligning it with the experiment’s meta-information. This enables the identification of clear causal structures, which, in Step (b), guide previous methods to pinpoint informative intervention targets effectively.

ity, which was considered restricted to humans (Gopnik et al., 2004; Trott et al., 2022; Sahu et al., 2022). Furthermore, LLMs have exhibited remarkable potential in advancing complex scientific discovery (AI4Science & Quantum, 2023). Additionally, discussions about the limitations of LLMs in understanding causality were also raised in the community (Zečević et al., 2023; Jin et al., 2023; Zhang et al., 2023a). This underscores the need for a robust approach to leverage the knowledge in LLMs about experimental design while mitigating the risks of being misled by their hallucinations regarding causality (Zhang et al., 2023b).

To this end, we present a new framework called Large Language Model Guided Intervention Targeting (LeGIT), designed to maximize while robustly leveraging the knowledge in LLMs to assist with the intervention targeting. Shown as in Fig. 1, at the beginning of the causal discovery, the numerical-based methods have limited numerical knowledge about the underlying causal system to use due to the limited data. Consequently, the estimated signals tend to be noisy and misleading. In contrast, LLMs can leverage the meta-information about the causal system and relate the learned world knowledge to identify high-potential intervening targets. After obtaining a relatively clearer causal graph, LLMs may not be able to provide sufficient guidance. Therefore, similar to humans, LeGIT leverages numerical methods to select the intervening targets. Our contributions can be summarized as follows:

- To the best of our knowledge, we are *the first to investigate the use of LLMs in the experimental design* to select intervention targets for causal discovery.
- We propose a novel framework called LeGIT that combines the advantages of both previous numerical methods as well as LLMs to facilitate the intervening targeting.

- We conduct extensive experiments with 4 real-world based benchmarks and verify that LeGIT can outperform previous numerical-based methods and even humans.
- We highlight the promise of LLMs in scientific discovery, that LLMs can effectively incorporate world knowledge, making them cost-efficient complements to humans.

2. Preliminaries

We begin by briefly introducing the preliminaries and notation in online causal discovery (Olko et al., 2023).

Causal relations among variables can be modeled using Structural Causal Models (SCMs)(Pearl & Mackenzie, 2018; Glymour et al., 2019), where each variable X_i is generated by $X_i = f_i(PA_i, U_i)$, with PA_i its causal parents and U_i independent noise. These relations can be represented by a directed acyclic graph (DAG) $G = (V, E)$, where nodes correspond to variables and edges represent direct causal links. The joint distribution factorizes as $P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i|PA_i)$. However, observational data alone can only identify the DAG up to a Markov Equivalence Class (MEC)(Spirtes et al., 2000).

To recover the true DAG from the MEC, online causal discovery incorporates interventional data (Tong & Koller, 2001; Hauser & Bühlmann, 2011; Ke et al., 2019). As outlined in Algorithm 1, a causal discovery algorithm \mathcal{A} iteratively updates its structure using both observational and interventional data. Interventions, modeled as replacing $P(X_i|PA_i)$ with $\hat{P}(X_i|PA_i)$, yield modified distributions $P_i(X) = \hat{P}(X_i|PA_i) \prod_{j \neq i} P(X_j|PA_j)$. We use hard interventions for simplicity. The online discovery proceeds in T rounds: Initially, a causal graph model ϕ_0 is fitted using observational data. In each subsequent round, an interven-

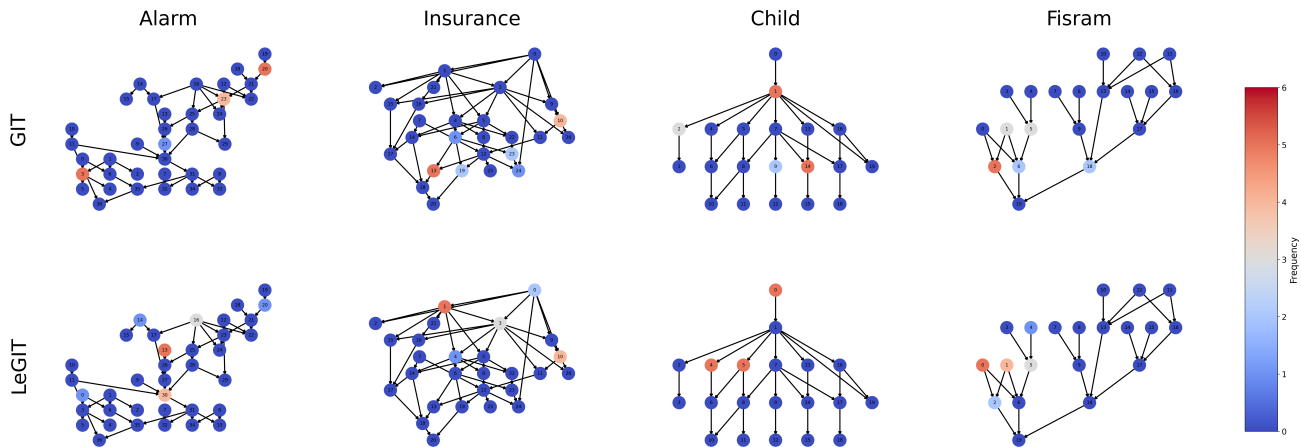


Figure 2. Intervention targets selected by LLM-based and gradient-based methods at the initial stage of online causal discovery.

tion target I is selected using a targeting method, and new interventional samples are collected to update the DAG.

Intervention targeting methods include AIT, which uses an F -test (Scherrer et al., 2021), and Bayesian Optimal Experimental Design, which selects targets via posterior inference over DAGs (Tigas et al., 2022). Gradient-based Intervention Targeting (GIT)(Olko et al., 2023) instead leverages gradient signals—via hallucinated gradients(Ash et al., 2020)—to evaluate interventions, offering improved performance and natural integration with gradient-based methods such as ENCO (Lippe et al., 2022), which we focus on in this study.

3. Methodology

3.1. Challenges in Existing Intervention Targeting

Despite the success of GIT method, GIT is highly sensitive to the accuracy of the gradient estimation and estimated causal graphs, which can be extremely noisy in the early rounds of an experiment. Therefore, we might mistakenly choose a variable that exerts minimal influence on the system, wasting valuable intervention budgets and misdirecting subsequent learning steps. To demonstrate the above issue, we consider four realistic causal discovery BN benchmarks (Scutari, 2010), i.e., *Alarm*, *Insurance*, *Child*, and *Fisram* and plot the distribution of the intervention target at the initial stage.

As given in Fig. 2, it can be found that the success of GIT varies across different datasets. Intuitively, at the beginning of the intervention, intervening on variables that affect lots of other variables can bring more information about the system (Lindley, 1956; Agrawal et al., 2019). In the Alarm dataset, the selected intervention targets are influential nodes. However, in the Insurance, Child, and Fisram dataset, the selected nodes only influence a few other nodes. Intervening on such targets with limited influence may lead to significant resource waste and further misdirect subse-

quent online causal discovery rounds.

In comparison, we design prompts to query LLMs about root causes in the system using only meta-information, such as variable descriptions. The prompt templates are shown in Fig.27. To evaluate the effectiveness of LLM-based intervention targeting, we visualize the intervention targets suggested by LLMs in Fig.2. Notably, even with limited input, LLMs can leverage their broad world knowledge to identify key influential nodes.

3.2. LLM-Guided Intervention Targeting

Motivated by the above experiments, we propose our framework, **Large Language Model Guided Intervention Targeting (LeGIT)**, which integrates the strengths of both numerical methods and LLMs to improve intervention targeting. The complete algorithmic LeGIT is provided in the Algorithm 2 Appendix B. LeGIT consists of four stages.

Warmup Stage At the start of online causal discovery, numerical estimations can be noisy and may mislead the process. To mitigate this, we first prompt LLMs to leverage their pre-trained knowledge and analyze variable descriptions to suggest influential candidates. The prompt template is shown in Fig.27. This process yields an initial set of intervention targets, denoted as \mathcal{D}_{warmup} . From \mathcal{D}_{warmup} , we select T_{warmup} variables to build a preliminary map of the causal system. To ensure robustness, we apply the self-consistency prompting technique(Wang et al., 2022) when determining the final targets.

Bootstrapped Stage Although the first warmup stage yields a basic structure of the underlying causal system, due to the intrinsic limitations of LLMs such as limited context length (Liu et al., 2023) and hallucination (Zhang et al., 2023b), LLMs may only focus on a subset of the variables and find the influential nodes therein. Nevertheless, when the number of causal variables is large, LLMs tend to give

Can Large Language Models Help Experimental Design for Causal Discovery?

Table 1. Average SHD, SID, and BSF with standard deviation (over 5 seeds) for real-world data ($T = 33$ rounds, $|D_{int}^I| = 32$, $N = 1056$).

Methods	Alarm			Insurance			Child			Fisram		
	SHD↓	SID↓	BSF↑	SHD↓	SID↓	BSF↑	SHD↓	SID↓	BSF↑	SHD↓	SID↓	BSF↑
CBED	28.20 ± 4.31	213.80 ± 42.44	0.8053 ± 0.06	21.60 ± 4.63	260.00 ± 31.83	0.7529 ± 0.04	5.40 ± 2.06	44.40 ± 18.51	0.9150 ± 0.05	3.60 ± 1.62	27.20 ± 6.31	0.9655 ± 0.02
AIT	32.80 ± 8.42	204.60 ± 52.09	0.7214 ± 0.05	24.20 ± 7.47	312.40 ± 87.50	0.6711 ± 0.11	9.00 ± 3.29	52.20 ± 21.03	0.8752 ± 0.04	8.00 ± 3.63	53.00 ± 29.23	0.9125 ± 0.05
Random Choice	38.80 ± 3.54	204.40 ± 58.15	0.7430 ± 0.08	26.00 ± 3.63	323.80 ± 14.96	0.7137 ± 0.02	5.40 ± 1.20	51.00 ± 17.11	0.9396 ± 0.04	4.40 ± 3.38	41.00 ± 37.38	0.9633 ± 0.02
Round Robin	25.00 ± 1.26	118.60 ± 21.78	0.9301 ± 0.02	17.40 ± 4.54	<u>232.20 ± 27.23</u>	<u>0.8042 ± 0.02</u>	3.40 ± 2.50	23.00 ± 14.39	0.9824 ± 0.02	4.80 ± 1.72	57.20 ± 19.61	0.9147 ± 0.02
Degree Prob	29.40 ± 4.67	144.60 ± 49.77	0.7798 ± 0.06	25.80 ± 2.93	305.20 ± 17.45	0.7054 ± 0.03	6.20 ± 2.48	36.20 ± 16.35	0.8842 ± 0.06	6.60 ± 1.20	40.60 ± 9.73	0.9147 ± 0.02
GIT	<u>19.60 ± 3.77</u>	131.40 ± 47.66	<u>0.9295 ± 0.02</u>	<u>16.40 ± 3.14</u>	243.80 ± 28.72	0.7960 ± 0.02	<u>2.80 ± 0.75</u>	<u>20.40 ± 12.50</u>	0.9841 ± 0.01	2.00 ± 1.67	27.00 ± 21.94	0.9944 ± 0.00
Human	22.60 ± 5.43	133.20 ± 27.01	0.8976 ± 0.02	<u>14.20 ± 3.43</u>	232.20 ± 40.74	0.8065 ± 0.03	2.00 ± 0.63	18.80 ± 8.42	0.9944 ± 0.00	<u>1.60 ± 1.02</u>	<u>21.20 ± 12.61</u>	0.9955 ± 0.00
LeGIT	17.40 ± 3.61	121.00 ± 38.27	0.9362 ± 0.02	12.60 ± 0.80	200.60 ± 35.32	0.8205 ± 0.01	2.20 ± 0.98	20.60 ± 5.61	<u>0.9858 ± 0.02</u>	1.20 ± 0.98	15.80 ± 8.23	0.9885 ± 0.02

an incomplete set of influential nodes. Therefore, we further incorporate a second warmup stage, to bootstrap the use of LLM’s world knowledge in early intervention targeting.

More concretely, we leverage the intermediate causal discovery results $\phi_{T_{warmup}}$ after the T_{warmup} rounds and examine the left variables that have not been involved in $\phi_{T_{warmup}}$. Then, we further prompt LLMs to give more focus on the left set of variables and to find the influential variables that were missing in previous rounds.

Double Selection Stage After getting the warmup and bootstrapped intervention target, we perform a double selection to ensure the robustness of the discovered causal structure while minimizing unnecessary interventions.

Continual Intervention Stage After the three warmup stages, we have already obtained relatively clearer yet complicated causal graphs. Even for humans, it is hard to determine the best experimental design. Therefore, we switch to using the numerical-based methods to continue to consume the remaining intervention budgets.

We discuss Theoretical and Practical part in Appendix D .

4. Experiments

In this section, we conduct extensive experiments to evaluate LeGIT on real-world datasets and compare LeGIT against various baselines in intervention selection and humans. Further details available in Appendix C.

Datasets We use four real-world based benchmark from the BN repository (abn; Scutari, 2010): *Fisram*, *Child*, *Insurance*, and *Alarm*. More details are given in Appendix C.7.

Baselines We compare LeGIT against different active online causal discovery algorithms GIT (Olko et al., 2023), AIT (Scherrer et al., 2021), CBED (Tigas et al., 2022), and four additional baselines following different heuristics: Random Choice, Round Robin, Degree Prob Sample, Human.

Implementation We conduct experiments using an observational dataset of $|D_{obs}| = 5000$ samples and perform $T = 33$ rounds of interventions, each collecting $|D_{int}^I| = 32$ interventional samples, totaling $N = 1056$ interventional samples. Other details are given in Appendix C.6.

Metrics We evaluate the performance of different meth-

ods using three metrics following the common practice: SHD (Tsamardinos et al., 2006), SID (Peters & Bühlmann, 2015), and BSF (Constantinou, 2019). A detailed description can be found in the Appendix C.5.

Empirical Results As shown in Table 1, it can be found that LeGIT achieves state-of-the-art causal discovery performances, with consistent improvements against the adopted gradient-based methods and *even human baseline* across all metrics and benchmarks. The superior SHD scores demonstrate that LeGIT is highly effective in accurately reconstructing the underlying graph structures, minimizing the number of erroneous edge modifications required. The consistently low SHD and SID scores, coupled with high BSF values, underscore the efficacy of LeGIT in accurately learning network structures and providing tangible benefits. Compared to heuristic-based methods like Random Choice and Round Robin, LeGIT offers a more strategic and data-driven approach, leading to better performance metrics. While Human interventions remain strong competitors, LeGIT bridges the gap between automated methods and expert-driven processes. This positions LeGIT as an effective tool for structure learning, capable of delivering expert-level performance without manual choices. More analysis and results are given in Appendix C.

5. Conclusions

In this work, we investigated the feasibility of incorporating LLMs into the intervention targeting for experimental design in causal discovery. We introduced a novel framework called LeGIT, which combines the best of previous numerical-based approaches and the rich knowledge in LLMs. Specifically, LeGIT leverages LLMs to warm up the online causal discovery procedure by identifying the influential root cause variables to begin the intervention. Empirically, we verified the effectiveness of LeGIT leveraging LLMs to warm up the online causal discovery can achieve the SOTA performance across multiple realistic causal discovery benchmarks. Notably, LeGIT also outperforms humans in intervention targeting, highlighting the high potential and strong effectiveness of LeGIT. The findings with LeGIT demonstrate that LLMs offer a scalable and cost-efficient approach to enhance experimental design, paving the way for new research directions in causal analysis and scientific discovery.

References

- Australasian bayesian network modelling society. <https://www.abnms.org/>.
- Abdulaal, A., adamos hadjivasiliou, Montana-Brown, N., He, T., Ijishakin, A., Drobnjak, I., Castro, D. C., and Alexander, D. C. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2024.
- Agrawal, R., Squires, C., Yang, K., Shanmugam, K., and Uhler, C. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3400–3409. PMLR, 2019.
- AI, M. Meta llama 3. <https://ai.meta.com/blog/meta-llama-3/>, 2024. Accessed: 2024-05-20.
- AI4Science, M. R. and Quantum, M. A. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*, 2023.
- Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=ryghZJBKPS>.
- Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data. *ArXiv*, abs/2007.01754, 2020. URL <https://api.semanticscholar.org/CorpusID:220347136>.
- Chen, S., Xu, M., Wang, K., Zeng, X., Zhao, R., Zhao, S., and Lu, C. Clear: Can language models really understand causal graphs? *arXiv preprint arXiv:2406.16605*, 2024.
- Cherry, A. B. C. and Daley, G. Q. Reprogramming cellular identity for regenerative medicine. *Cell*, 148:1110–1122, 2012. URL <https://api.semanticscholar.org/CorpusID:9993432>.
- Constantinou, A. C. Evaluating structure learning algorithms with a balanced scoring function. *arXiv preprint arXiv:1905.12666*, 2019.
- DeepSeek-AI. Deepseek-v3 technical report, 2024. URL <https://arxiv.org/abs/2412.19437>.
- DeepSeek-AI. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Dong, X., Dai, H., Fan, Y., Jin, S., Rajendran, S., and Zhang, K. On the three demons in causality in finance: Time resolution, nonstationarity, and latent factors. *ArXiv*, abs/2401.05414, 2023. URL <https://api.semanticscholar.org/CorpusID:266933380>.
- Elidan, G. Bayesian network repository. <https://www.cs.huji.ac.il/w~galel/Repository/>, 2025. Available at <https://www.cs.huji.ac.il/w~galel/Repository/>.
- Glymour, C., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111 1:3–32, 2004.
- Hauser, A. and Bühlmann, P. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.*, 13:2409–2464, 2011. URL <https://api.semanticscholar.org/CorpusID:16393667>.
- Houlsby, N., Huszar, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745, 2011. URL <http://arxiv.org/abs/1112.5745>.
- Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M. T., and Schölkopf, B. Can large language models infer causation from correlation? *arXiv preprint*, arXiv:2306.05836, 2023.
- Jiralerspong, T., Chen, X., More, Y., Shah, V., and Bengio, Y. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*, 2024.
- Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Pal, C. J., and Bengio, Y. Learning neural causal models from unknown interventions. *ArXiv*, abs/1910.01075, 2019. URL <https://api.semanticscholar.org/CorpusID:203626996>.
- Kiciman, 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.
- Kuhn, T. S. and Hawkins, D. The structure of scientific revolutions. *American Journal of Physics*, 31:554–555, 1963.

- Lampinen, A. K., Chan, S. C., Dasgupta, I., Nam, A. J. H., and Wang, J. X. Passive learning of active causal strategies in agents and language models. In *Advances in Neural Information Processing Systems*, 2023.
- Li, P., Wang, X., Zhang, Z., Meng, Y., Shen, F., Li, Y., Wang, J., Li, Y., and Zhu, W. Realtcd: Temporal causal discovery from interventional data with large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, pp. 4669–4677, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3680042. URL <https://doi.org/10.1145/3627673.3680042>.
- Lindley, D. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1956. URL <https://api.semanticscholar.org/CorpusID:123582195>.
- Lippe, P., Cohen, T., and Gavves, E. Efficient neural causal discovery without acyclicity constraints. In *International Conference on Learning Representations*, 2022.
- Liu, C., Chen, Y., Liu, T., Gong, M., Cheng, J., Han, B., and Zhang, K. Discovery of the hidden world with large language models. *ArXiv*, abs/2402.03941, 2024. URL <https://api.semanticscholar.org/CorpusID:267499909>.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *arXiv preprint*, arXiv:2307.03172, 2023.
- Long, S., Piché, A., Zantedeschi, V., Schuster, T., and Drouin, A. Causal discovery with language models as imperfect experts. *arXiv preprint*, arXiv:2307.02390, 2023a.
- Long, S., Schuster, T., and Piché, A. Can large language models build causal graphs? *arXiv preprint*, arXiv:2303.05279, 2023b.
- Olko, M., Zając, M., Nowak, A., Scherrer, N., Annadani, Y., Bauer, S., Kuciński, Ł., and Miłoś, P. Trust your ∇ : Gradient-based intervention targeting for causal discovery. *Advances in Neural Information Processing Systems*, 36:50617–50647, 2023.
- OpenAI. Gpt-4 technical report, 2023.
- OpenAI. Hello, gpt-4o! <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed: 2024-05-20.
- Pearl, J. and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.
- Peters, J. and Bühlmann, P. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3): 771–799, 2015.
- Sahu, P., Cogswell, M., Gong, Y., and Divakaran, A. Unpacking large language models with conceptual consistency. *ArXiv*, abs/2209.15093, 2022. URL <https://api.semanticscholar.org/CorpusID:252668345>.
- Scherrer, N., Bilaniuk, O., Annadani, Y., Goyal, A., Schwab, P., Schölkopf, B., Mozer, M. C., Bengio, Y., Bauer, S., and Ke, N. R. Learning neural causal models with active interventions. *arXiv preprint arXiv:2109.02429*, 2021.
- Scutari, M. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(3):1–22, 2010.
- Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press, 2000. ISBN 978-0-262-19440-2.
- Spirtes, P., Glymour, C., Scheines, R., and Tillman, R. Automated Search for Causal Relations: Theory and Practice. 2010. URL https://kilthub.cmu.edu/articles/journal_contribution/Automated_Search_for_Causal_Relations_Theory_and_Practice/6490961.
- Tigas, P., Annadani, Y., Jesson, A., Schölkopf, B., Gal, Y., and Bauer, S. Interventions, where and how? experimental design for causal models at scale. *Advances in neural information processing systems*, 35:24130–24143, 2022.
- Tong, S. and Koller, D. Active learning for structure in bayesian networks. In *International Joint Conference on Artificial Intelligence*, 2001. URL <https://api.semanticscholar.org/CorpusID:8155128>.
- Trott, S., Jones, C. J., Chang, T. A., Michaelov, J. A., and Bergen, B. K. Do large language models know what humans know? *Cognitive science*, 47 7:e13309, 2022. URL <https://api.semanticscholar.org/CorpusID:252089182>.
- Tsamardinos, I., Brown, L., and Aliferis, C. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 10 2006. doi: 10.1007/s10994-006-6889-7.

- Vashishtha, A., Reddy, A. G., Kumar, A., Bachu, S., Balasubramanian, V. N., and Sharma, A. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., family=Walt, given=Stéfan J., p. d. u., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, , Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., family=Mulbregt, given=Paul, p. u., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., family=Miranda Cardoso, given=José Vinícius, p. u., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., Vázquez-Baeza, Y., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. 17(3):261–272. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2. URL <https://doi.org/10.1038/s41592-019-0686-2>.
- Vowels, M. J., Camgoz, N. C., and Bowden, R. D’ya like dags? a survey on structure learning and causal discovery. *ACM Computing Survey*, 55(4), 2022. ISSN 0360-0300.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Zečević, M., Willig, M., Dhami, D. S., and Kersting, K. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.
- Zhang, C., Bauer, S., Bennett, P., Gao, J., Gong, W., Hilmkil, A., Jennings, J., Ma, C., Minka, T., Pawlowski, N., and Vaughan, J. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint, arXiv:2304.05524*, 2023a.
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., and Shi, S. Siren’s song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint, arXiv:2309.01219*, 2023b.

A. Related Work

Intervention Targeting/Experiment Design Scientific progress in causal discovery is often driven by interventional experiments inspired by observational insights (Kuhn & Hawkins, 1963). Traditional methods focused on designing effective experiments to establish causal links, while statistical approaches aimed to automate causal inference from observational data (Pearl & Mackenzie, 2018; Spirtes et al., 2000). However, observational data alone is insufficient for identifying causal structures, and interventional data is costly to collect (Spirtes et al., 2000). To address these challenges, several methods for optimal intervention design have been developed.

Active Intervention Targeting (AIT) selects intervention targets using an F -test inspired criterion, evaluating discrepancies in interventional sample distributions from a posterior distribution of graphs (Scherrer et al., 2021). Causal Bayesian Experimental Design (CBED) uses Bayesian Optimal Experimental Design to select interventions that maximize mutual information (MI) between new data and existing graph beliefs, with MI estimated via a BALD-like method (Tigas et al., 2022; Houlisby et al., 2011). Gradient-based Intervention Targeting (GIT) (Olko et al., 2023) leverages gradient information to determine interventions that maximize impact on causal parameter updates, which is particularly advantageous in low-data settings. In our work, we explore leveraging these advanced intervention strategies within the framework of LLMs to determine whether LLMs can effectively engage in experimental design for causal discovery, pushing the boundaries of what automated, data-driven causal inference can achieve.

Causal Discovery With LLMs Recent advancements in LLMs have opened new opportunities in causal learning and reasoning by incorporating domain knowledge, common sense, and contextual reasoning (Kiciman et al., 2023). LLMs have demonstrated capabilities across Pearl’s ladder of causation—association, intervention, and counterfactuals—bridging gaps that traditional models have with high-level causal reasoning. They have shown promising results in pairwise causal discovery tasks by utilizing semantic information not accessible through numerical data alone (Jiralerspong et al., 2024).

On the other hand, LLMs can sometimes behave like “causal parrots”, repeating learned associations without demonstrating true causal reasoning (Zečević et al., 2023; Chen et al., 2024). Moreover, their performance varies significantly depending on task complexity, with limited success in advanced causal reasoning such as full graph discovery and counterfactual analysis (Zhang et al., 2023a; Jin et al., 2023; Long et al., 2023a). Another promising line of work integrates LLMs with traditional causal discovery methods to leverage their complementary strengths (Long et al., 2023a; Abdulaal et al., 2024; Vashishtha et al., 2023; Liu et al., 2024). This hybrid approach has shown improved performance in constructing causal graphs, benefiting from LLMs’ understanding of language context and traditional methods’ data-driven precision.

While prior studies emphasize the role of LLMs in causal analysis, the question of whether LLMs can meaningfully contribute to experimental design in causal discovery remains largely unaddressed. Experimental design encompasses proposing interventions, predicting outcomes, and assessing experimental strategies—tasks that extend beyond basic causal inference. This paper seeks to bridge this gap by investigating the potential of LLMs to support experimental design, exploring their unique value, and critically evaluating their strengths and limitations in guiding causal experiments.

B. Algorithm Illustration

In this part, we present the online causal discovery algorithm as algorithm. 1, and LeGIT as algorithm. 2.

Algorithm 1 ONLINE CAUSAL DISCOVERY (Olko et al., 2023)

Require: Causal discovery algorithm \mathcal{A} (e.g., ENCO), Number of data acquisition rounds T , Intervention targeting method \mathcal{M} , Observational dataset \mathcal{D}_{obs}

Output: Final parameters of graph model: φ_T and Final estimated CausalDAG: $\mathbb{P}(G)$

- 1: $\mathcal{D}_{int} \leftarrow \emptyset$
 - 2: Fit graph model φ_0 with algorithm \mathcal{A} on \mathcal{D}_{obs}
 - 3: **for** each intervention acquisition round $i = 1, 2, \dots, T$ **do**
 - 4: $I_i \leftarrow$ generate intervention targets using \mathcal{M}
 - 5: $\mathcal{D}_{int}^i \leftarrow$ query for data from interventions I_i
 - 6: $\mathcal{D}_{int} \leftarrow \mathcal{D}_{int} \cup \mathcal{D}_{int}^i$
 - 7: Fit φ_i with algorithm \mathcal{A} on \mathcal{D}_{int} and \mathcal{D}_{obs}
 - 8: **end for**
-

Algorithm 2 LEGIT: LARGE LANGUAGE MODEL GUIDED INTERVENTION TARGETING

Require: Causal discovery algorithm for Intervention Data \mathcal{A} (e.g., ENCO); Intervention Score targeting method \mathcal{M} , (e.g. GIT); LLM for root cause proposal Ψ ; Number of data acquisition rounds T ; Observational dataset \mathcal{D}_{obs} ; Graph Node List V ; Warmup Epoch T_{warmup} ; Bootstrapped Search Epoch $T_{bootstrapped}$

Ensure: Final parameters of graph model: φ_T and CausalDAG: $\mathbb{P}(G)$

- 1: $\mathcal{D}_{warmup} \leftarrow \Psi(V, T_{warmup})$ //Get Warmup List from LLM
 - 2: **for** round $i = 1, 2, \dots, T$ **do**
 - 3: **if** $i \leq T_{warmup}$ **then**
 - 4: $\mathcal{D}_{int}^I \leftarrow \mathcal{D}_{warmup}[i]$
 - 5: **else if** $i = T_{warmup} + 1$ **then**
 - 6: // Get the Isolated Nodes List
 - 7: $V_{isolated} \leftarrow$ isolated node from $\mathbb{P}(G_i)$
 - 8: //Get Bootstrapped warmup Intervention Target from isolated Nodes
 - 9: $\mathcal{D}_{bootstrapped} \leftarrow \Psi(V_{isolated}, T_{bootstrapped})$
 - 10: $\mathcal{D}_{int}^I \leftarrow \mathcal{D}_{bootstrapped}[i - T_{bootstrapped}]$
 - 11: **else if** $T_{warmup} < i \leq T_{warmup} + T_{bootstrapped}$ **then**
 - 12: $\mathcal{D}_{int}^I \leftarrow \mathcal{D}_{bootstrapped}[i - T_{warmup}]$
 - 13: **else if** $T_{warmup} + T_{bootstrapped} < i \leq 2(T_{warmup} + T_{bootstrapped})$ **then**
 - 14: //Double Selection LLM'S List
 - 15: $\mathcal{D}_{int}^I \leftarrow (\mathcal{D}_{warmup} + \mathcal{D}_{bootstrapped})[i - T_{warmup} - T_{bootstrapped}]$
 - 16: **else**
 - 17: $\mathcal{D}_{int}^I \leftarrow$ generate intervention targets using \mathcal{M} ,
 - 18: **end if**
 - 19: $\mathcal{D}_{int} \leftarrow \mathcal{D}_{int} \cup \mathcal{D}_{int}^I$
 - 20: Fit φ_i with algorithm \mathcal{A} on \mathcal{D}_{int} and \mathcal{D}_{obs}
 - 21: **end for**
-

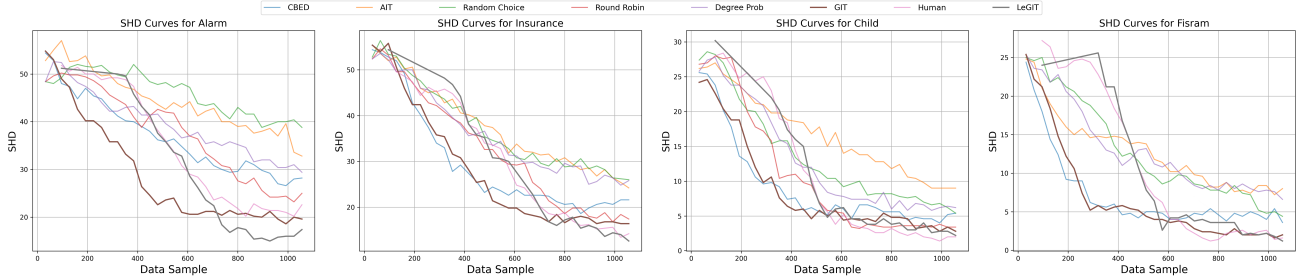


Figure 3. SHD metric for different methods (over 5 seeds) towards different intervention samples. ($T = 33$ rounds, $|D_{int}^I| = 32$, $N = 1056$)

C. Full Experimental Results and settings

C.1. Normal Data Setting Experiment Analysis

Intervention dynamics. In Fig. 3, we further plot the performances of different methods along with the increase of the data samples obtained from different rounds. It can be found that, although at the beginning of the online causal discovery, LeGIT may not demonstrate outstanding SHD results. Along with more data samples coming in, LeGIT converge to a better solution faster than any other methods. In contrast, despite a faster decrease speed of GIT, GIT finally converges to a suboptimal solution due to unsuitable initialization, which verifies our discussion.

Statistical significance. In addition, we also perform the Paired T-test and the Wilcoxon Single-Rank test (Virtanen et al.) against GIT and LeGIT of the result in Table 1. The result is shown in Table 2. Considering the sample size, at a 90% confidence level, we believe that the results of LeGIT outperform GIT on all four datasets. The results clearly indicate that LeGIT outperforms existing baseline methods across all three evaluation metrics.

Table 2. P-value of Paired T-test and Wilcoxon test with GIT and LeGIT SHD results in Table 1.

	Paired T-test	Wilcoxon test
Alarm	0.074	0.100
Insurance	0.060	0.067
Child	0.070	0.083
Fisram	0.099	0.102

C.2. Low Data Experiment Analysis

Table 3. Average SHD, SID, and BSF with standard deviation (over 5 seeds) for real-world data with a low data budget ($T = 33$ rounds, $|D_{int}^I| = 16$, $N = 528$).

Methods	Alarm			Insurance			Child			Fisram		
	SHD↓	SID↓	BSF↑	SHD↓	SID↓	BSF↑	SHD↓	SID↓	BSF↑	SHD↓	SID↓	BSF↑
CBED	32.40 ± 4.36	214.20 ± 69.73	0.8229 ± 0.05	26.40 ± 3.56	327.00 ± 38.46	0.7042 ± 0.02	9.20 ± 3.25	46.60 ± 18.49	0.8735 ± 0.06	5.60 ± 1.02	44.20 ± 21.25	0.9425 ± 0.04
AIT	41.20 ± 5.49	270.00 ± 29.61	0.6728 ± 0.07	37.00 ± 12.26	421.40 ± 82.68	0.5602 ± 0.12	10.00 ± 3.29	73.40 ± 45.64	0.8175 ± 0.11	12.60 ± 1.96	80.60 ± 13.81	0.8643 ± 0.04
Random Choice	40.80 ± 2.71	236.40 ± 12.31	0.7203 ± 0.07	25.60 ± 2.24	311.00 ± 22.17	0.7211 ± 0.03	8.20 ± 2.32	51.60 ± 33.15	0.8300 ± 0.07	5.80 ± 2.79	46.00 ± 22.92	0.9507 ± 0.04
Round Robin	33.60 ± 7.34	169.00 ± 35.69	0.8975 ± 0.06	22.60 ± 3.72	269.20 ± 44.37	0.7749 ± 0.04	4.60 ± 2.42	32.40 ± 24.25	0.9327 ± 0.06	5.00 ± 1.55	45.60 ± 22.60	0.9599 ± 0.02
Degree Prob	42.60 ± 6.34	244.20 ± 35.06	0.6762 ± 0.08	31.80 ± 4.40	351.00 ± 27.64	0.6737 ± 0.05	9.00 ± 2.90	60.80 ± 23.01	0.8512 ± 0.07	11.00 ± 3.74	67.80 ± 20.45	0.8676 ± 0.06
GIT	27.20 ± 4.71	177.80 ± 61.65	0.9025 ± 0.04	22.40 ± 3.72	296.00 ± 44.23	0.7463 ± 0.04	6.00 ± 1.55	33.80 ± 15.75	0.9134 ± 0.05	3.60 ± 3.61	35.20 ± 31.90	0.9731 ± 0.04
Human	24.00 ± 2.28	188.40 ± 27.09	0.9093 ± 0.02	20.40 ± 2.65	280.60 ± 26.04	0.7603 ± 0.04	4.60 ± 2.06	20.60 ± 24.81	0.9482 ± 0.04	3.80 ± 1.47	33.80 ± 15.35	0.9731 ± 0.02
LeGIT	21.00 ± 2.37	159.40 ± 26.81	0.9158 ± 0.01	18.20 ± 1.17	259.00 ± 66.69	0.7894 ± 0.02	4.40 ± 2.15	28.20 ± 15.03	0.9499 ± 0.04	2.20 ± 1.17	29.00 ± 17.30	0.9857 ± 0.02

Furthermore, we conduct additional experiments in an extremely low-data setting, where only 16 interventional data samples are sampled from each round, and other settings are the same as above. This low-data setting is more practically relevant. Additionally, due to the insufficient intervention data, the performance of causal discovery algorithms in estimating effects is diminished (Lippe et al., 2022), which further tests the effectiveness and robustness of the intervention strategy.

The results presented in Table 3, where LeGIT achieves larger improvements under the challenging low-data condition across all datasets. We also provide the SHD curves with respect to different intervention samples under Table 3 settings in Fig.25 of Appendix C.9. These findings serve as strong evidence that reaffirm the effectiveness of LeGIT in real-world experimental design scenarios, where both the number of interventions and the sample size are limited.

The result of the low-data experiment further verifies our discussion that numerical methods suffer from noise or insufficient data, leading to a suboptimal solution. The numerical-based method does not even outperform round-robin on 3 smaller datasets, underscoring its limitations in such scenarios. In contrast, the use of LLMs enables scalable and effective guidance

that complements numerical methods, reducing the risk of suboptimal convergence, and having more stable performance in real-world applications.

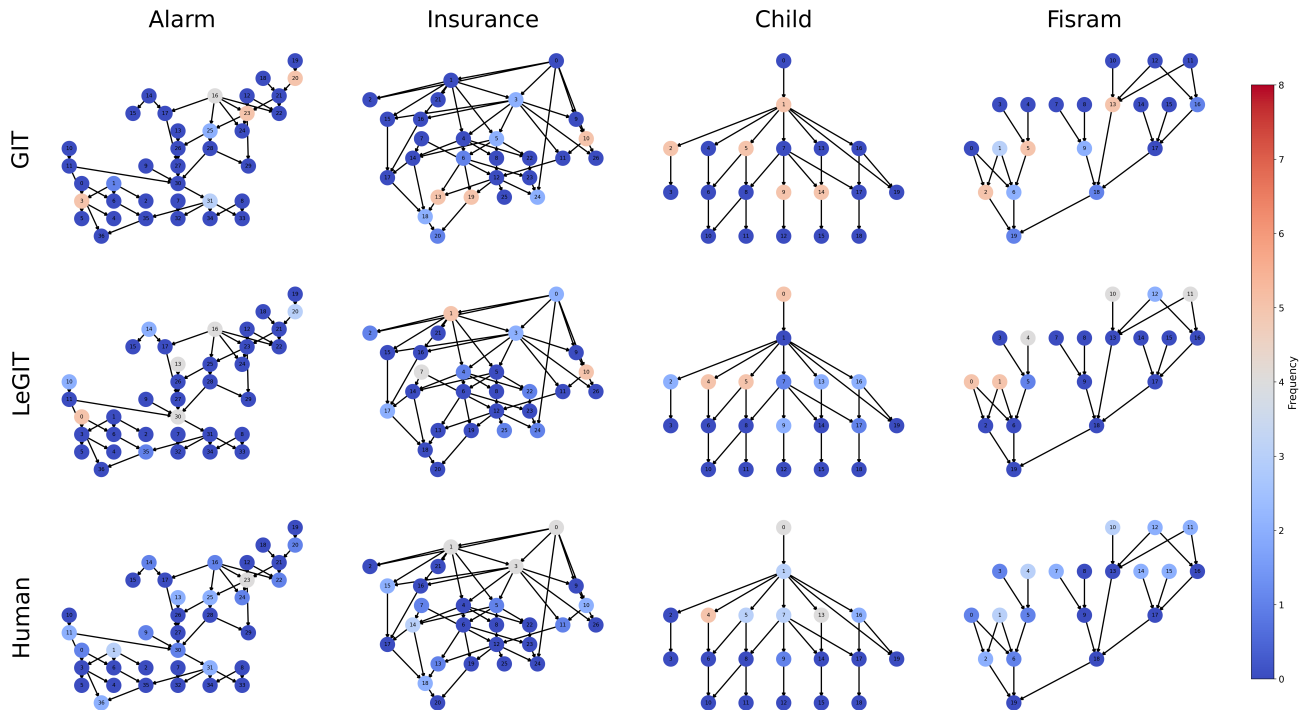


Figure 4. The selected Node Frequency obtained by different strategies on Epoch 0-4 from 5 different seeds under Table 1 setting.

C.3. Detailed comparisons and analyses of initial intervention targeting

Figs. 4 depict the distribution of selected nodes between epochs from 0 to 5. Notably, the numerical methods (GIT) tend to get trapped in the initialization phase of the Insurance, alarm dataset, which consistently select less central nodes in the graph that are often peripheral or leaf nodes. In contrast, our model (LeGIT) that SocioEcon (socioeconomic status, No.1 nodes in graph), plays a crucial role in the insurance system, potentially influencing car choice, driving behavior, and the ability to afford certain safety features.

Compared to humans, LeGIT demonstrates superior performance on two complex datasets: Alarm and Insurance. As the number of variables increases, determining the optimal interventions to reveal the structure of the causal graph becomes combinatorially explosive. For humans, this process can be *extremely tedious or error-prone*, as they may subjectively favor certain nodes, failing to synthesize different viewpoints due to simpler mental models. In contrast, refer to Figs. 1, LLMs follow the instructions provided in Fig. 3 step by step and align them with their background knowledge. With the self-consistency prompt technique, LLMs generate more robust results, providing a highly cost-effective alternative to hiring multiple human experts for advice.

Discussion. LLMs’ primary value lies in scalability and availability, providing immediate, cost-effective guidance in real-time, especially for online causal discovery where rapid interventions are required. They excel in large-scale systems with many variables, where it’s infeasible for experts to assess all nodes. LLMs complement human oversight by filling gaps in availability, consistency, and knowledge while helping avoid expert biases. Additionally, LLMs quickly process metadata, saving experts time and providing a solid starting point, as seen in other AI-assisted tasks.

C.4. Detailed Baseline

1. **Random Choice:** A target node is randomly select from the set of all nodes at each step.
2. **Round Robin:** A target node is chosen randomly from the unvisited nodes at each step. Once all nodes are selected, the visitation counts are reset.

3. **Degree Prob Sample:** A target node is randomly chosen from all nodes, with selection probability normalized by each node’s out-degree.
4. **Human:** We ask five *master’s/Ph.D.-level* individuals, presenting them with the same information and process as provided to the LLMs.

Among the baselines, Degree Prob Sample can be considered as an *oracle* to LLM that adopts the out-degree of each node in the ground truth DAG. In addition, we also include the human baseline to better isolate and understand the unique contributions of LLMs.

C.5. Detailed Metrics

In this section, we present the details of 3 different metrics mentioned in the experiment part.

- The Structural Hamming Distance (SHD) (Tsamardinos et al., 2006): SHD is a frequently employed score comparing graph structures via their binary adjacency matrices. It represents the minimum sum of edge additions (A), deletions (D), and reversals (R) required to convert one adjacency matrix into that of the ground truth causal graph.

$$SHD = A + D + R \tag{1}$$

- Structural Intervention Distance (SID) (Peters & Bühlmann, 2015): This metric measures how closely two DAGs, \mathcal{G} and \mathcal{H} , align in terms of the causal effects they encode. SID is defined as the total count of intervention distributions (from node i to node j) that are inaccurately predicted by the candidate graph \mathcal{H} when compared against the reference graph \mathcal{G} . Consequently, SID reveals the impact of edge errors within \mathcal{H} on the resulting causal effect estimations.

$$SID = \#\{(i, j), i \neq j \mid \text{the intervention distribution from } i \text{ to } j \text{ is falsely estimated by } \mathcal{H} \text{ with respect to } \mathcal{G}\}. \tag{2}$$

- Balanced Scoring Function (BSF) (Constantinou, 2019): BSF offers an unbiased method for evaluating the performance of graph structure learning algorithms. It achieves this by normalizing the contributions of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) according to the prevalence of actual dependencies and independencies in the reference graph structure. The calculation is performed as follows:

$$BSF = \frac{1}{2} \left(\frac{TP}{a} + \frac{TN}{i} - \frac{FP}{i} - \frac{FN}{a} \right), \tag{3}$$

Here, a is the count of arcs in the ground truth graph. The term i corresponds to the number of absent arcs (independencies) in the true graph, calculated as $i = \frac{|N| \times (|N| - 1)}{2} - a$, where $|N|$ is the total count of nodes.

C.6. Implementation

In experiments presented in this section 4, we utilize ENCO (Lippe et al., 2022) as the backbone causal discovery algorithm. The observational dataset consists of $|\mathcal{D}_{obs}| = 5000$ samples, and we conduct $T = 33$ rounds of intervention sampling, with each round acquiring an interventional batch of $|\mathcal{D}_{int}^I| = 32$ samples, leading to a total of $N = 1056$ interventional samples. We set $T_{warmup} = 3$ and $T_{bootstrapped} = 2$ for LeGIT.

Table 4. Hyperparameters used for the ENCO framework.

parameter	value
Sparsity regularizer λ_{sparse}	4×10^{-3}
Distribution model	2 layers, hidden size 64, LeakyReLU($\alpha = 0.1$)
Batch size	128
Learning rate - model	5×10^{-3}
Weight decay - model	1×10^{-4}
Distribution fitting iterations F	1000
Graph fitting iterations G	100
Graph samples K	100
Epochs	30
Learning rate - γ	2×10^{-2}
Learning rate - θ	1×10^{-1}

For experiments using the ENCO framework, we used the exact parameters reported by (Lippe et al., 2022). These parameters are provided in Table 4 to ensure the completeness of our report. We used the GPT-4O API (OpenAI, 2024)(gpt-4o-2024-08-06 version) from azure platform for all LLM-based experiments except for special instructions.

C.7. More Details of Datasets

In this part, we will further introduce the 4 different domain Causal graph discovery dataset from bnlearn Repository (Elidan, 2025) and BMNA BN Repository (abn). For the description of each variable, we refer to the original papers of each dataset, the bnlearn Package Document (Scutari, 2010), and make a few changes (Long et al., 2023b). We show the ground truth and the out-degree node distributions as follows.

Fisram (Freshwater Fish Injurious Species Risk Assessment Model) shown as Fig.5 is to assess the potential invasiveness and harm of introduced freshwater fish species, aiding decisions on their importation. The model consists of 20 nodes and 23 edges, representing key species traits, environmental factors, and historical data used to assess potential ecological harm.

Child show as Fig.6 is used to model the diagnosis of pediatric health issues, particularly those that can occur in newborns or young children. It’s often employed in studies related to decision support systems, where probabilistic graphical models assist in medical diagnosis, with 20 nodes and 25 edges.

Insurance shown as Fig. 7 intended to simulate a situation in which an insurance company needs to assess various risks and make decisions regarding policies, claims, and customer behavior. It represents the interdependencies between multiple insurance factors. It has 27 nodes and 52 edges.

Alarm shown as Fig. 8 is known as the ALARM (A Logical Alarm Reduction Mechanism) network, and it was originally developed to model a patient monitoring system for anesthesia purposes. It helps in predicting physiological conditions of patients, detecting potential complications, and generating alerts when necessary, consists of 37 nodes and 46 edges.

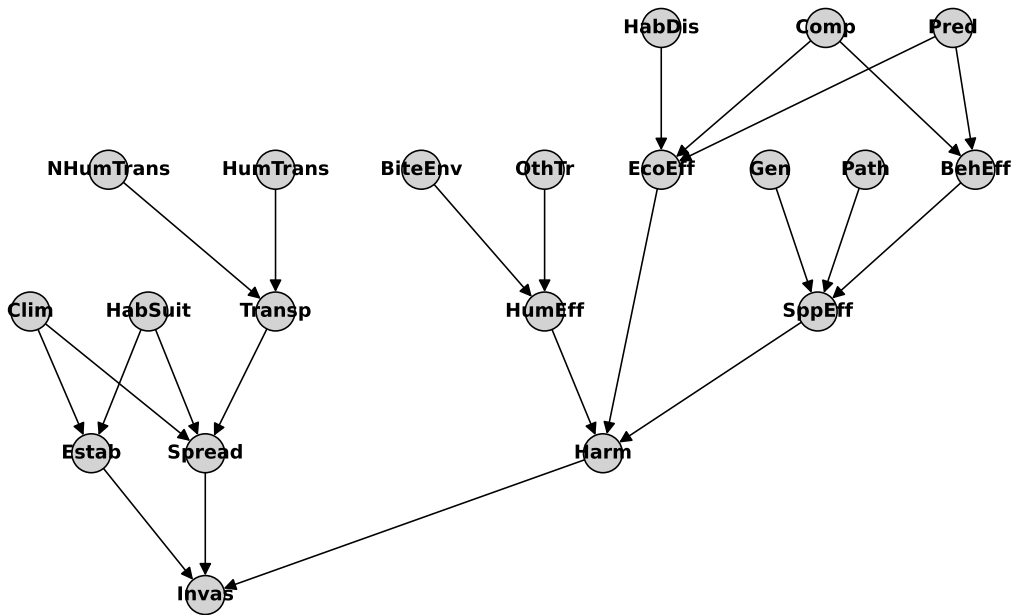


Figure 5. Ground truth Causal Graph for fisram data.

C.8. Final Causal Graph

In this section, we present the final causal graph after $T = 33$, total sample $N = 1056$ results with GIT, Human, and LeGIT.

C.9. Low-Data settings Visualization

In this section, we provide the SHD curve under low-data settings as Fig. 25 and the selected Node Frequency obtained by different strategies on Epoch 0-4 as Fig. 26.

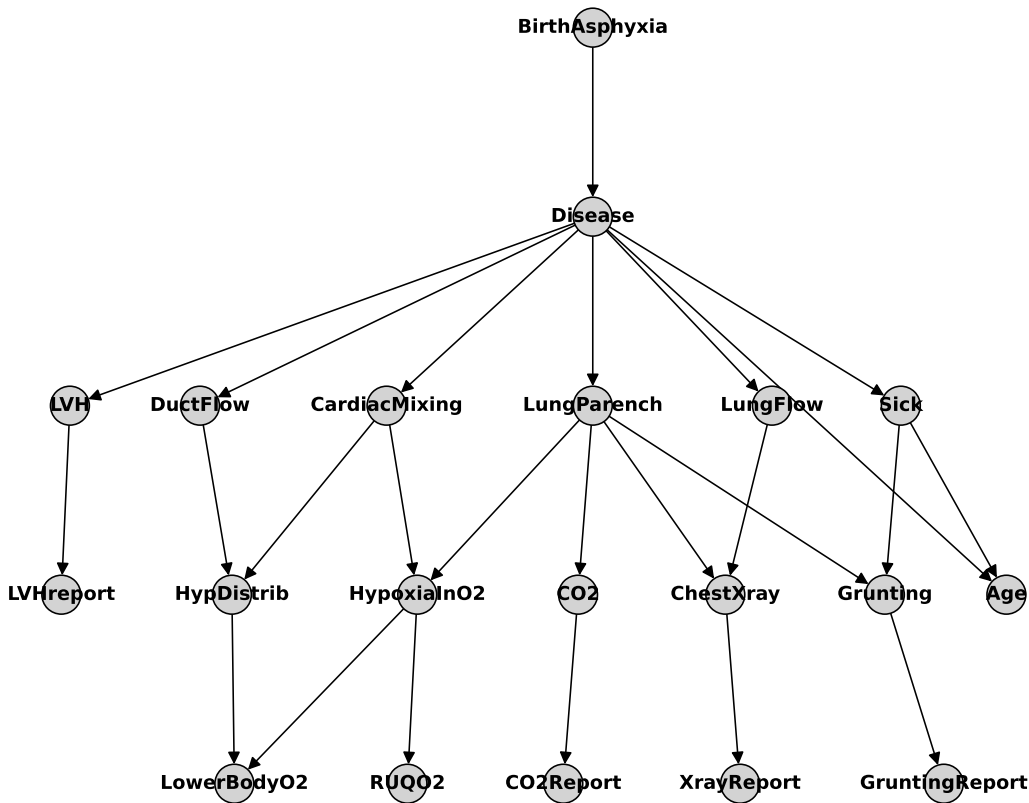


Figure 6. Ground truth Causal Graph for child data.

C.10. Open-source models experiments

We tested DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI, 2025), Deepseek-V3 (DeepSeek-AI, 2024) and LLama-3.1-405B (AI, 2024) in the Alarm and Insurance datasets. The SHD and MeanNHD ($NHD = \frac{1}{N_{node}^2} SHD$) results for the Alarm and Insurance datasets within 5 random seeds in 2 settings are shown in Table 5. With different LLMs, we can find that LeGIT still consistently shows strong performance with these LLMs, highlighting the robustness and adaptability of our proposed framework.

Table 5. SHD and MeanNHD of using different LLMs in LeGIT framework under 2 settings with 5 random seeds.

Methods	Normal Settings			Low Data Settings		
	Alarm	Insurance	MeanNHD ↓	Alarm	Insurance	MeanNHD ↓
GIT	19.60 ± 3.77	16.40 ± 3.14	0.0184	27.20 ± 4.71	22.40 ± 3.72	0.0253
LeGIT (GPT-4O)	17.40 ± 3.61	12.60 ± 0.80	0.0161	21.00 ± 2.37	18.20 ± 1.17	0.0202
LeGIT (DeepseekR1-14B)	20.00 ± 2.12	14.20 ± 2.71	0.0170	22.00 ± 0.82	24.60 ± 2.15	0.0249
LeGIT (DeepseekV3)	18.60 ± 3.44	14.60 ± 2.72	0.0168	22.40 ± 1.85	20.60 ± 1.02	0.0223
LeGIT (LLama-3.1-405B)	18.60 ± 2.33	15.80 ± 4.07	0.0176	28.00 ± 4.15	18.20 ± 4.26	0.0227

C.11. Resources

We utilized a system comprising two Intel Xeon Platinum 8358P processors with 2.6GHz, two NVIDIA A40 GPUs (48GB each) and 1 TB of memory. For the large language model (LLM) API, we leveraged the Azure platform.

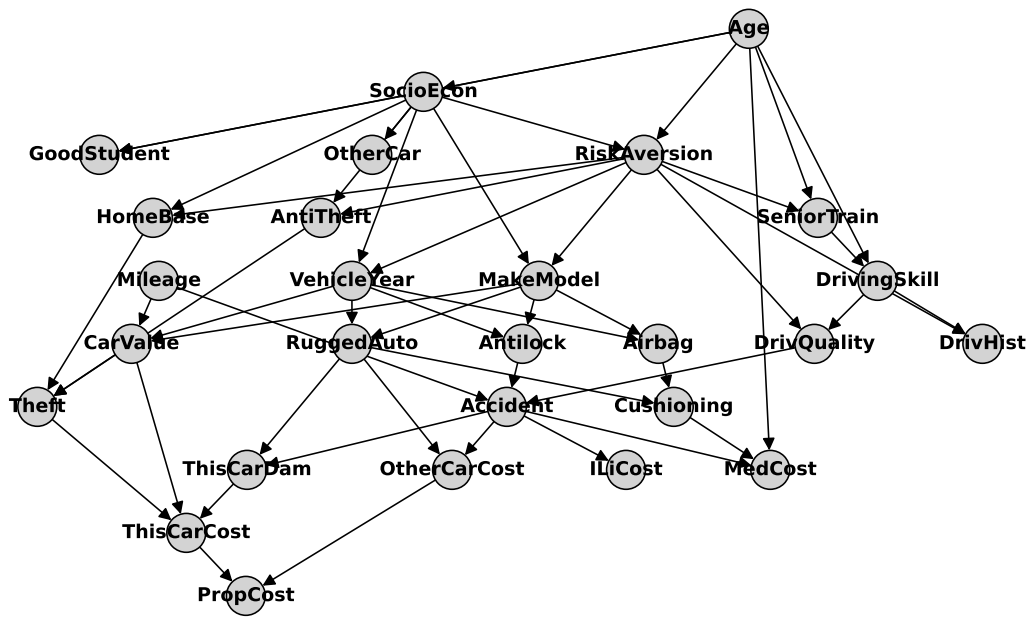


Figure 7. Ground truth Causal Graph for insurance data.

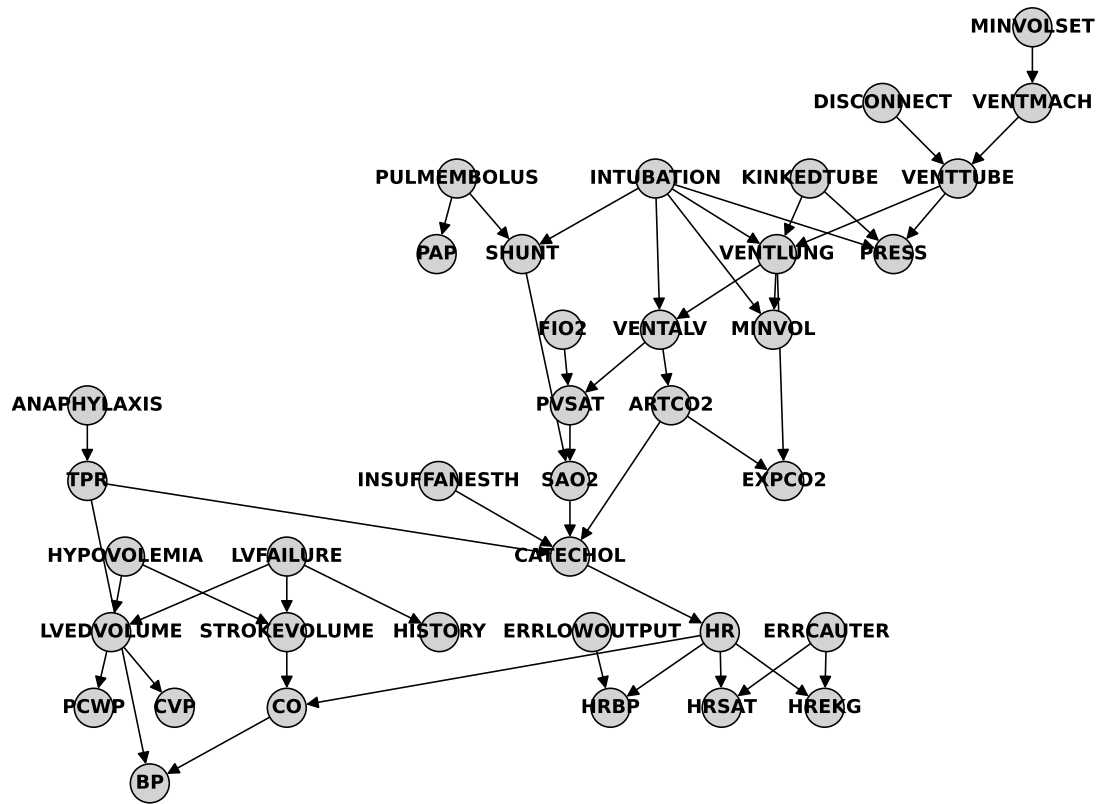


Figure 8. Ground truth Causal Graph for alarm data.

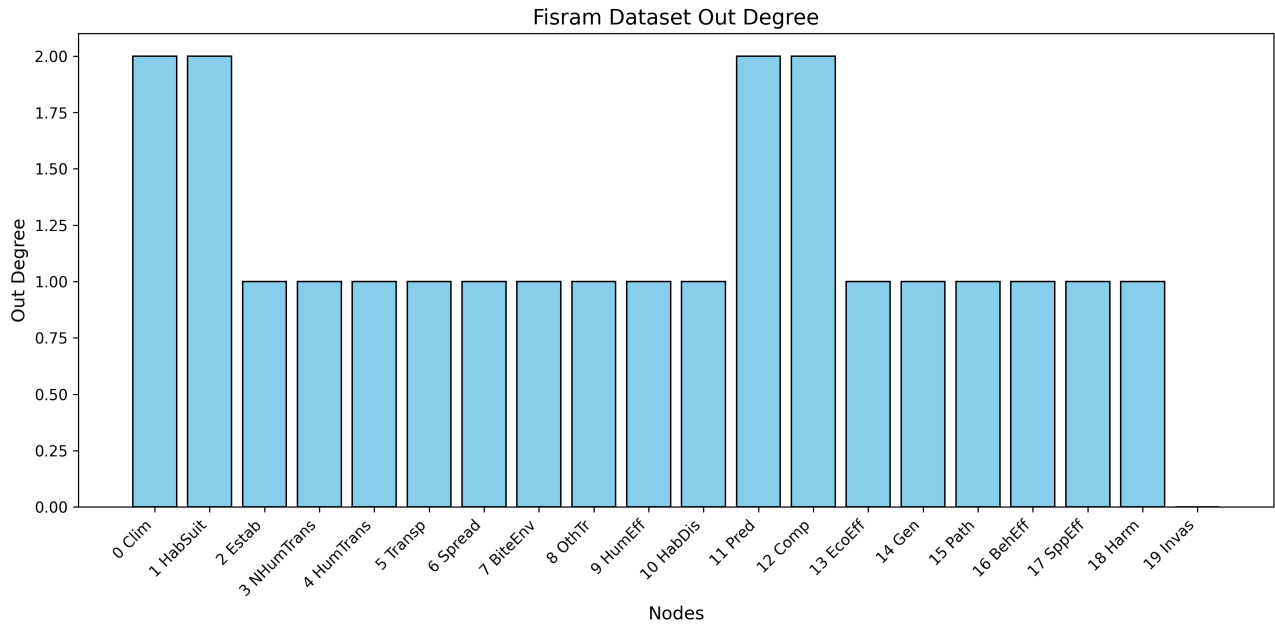


Figure 9. Out degree distribution of Fisram data.

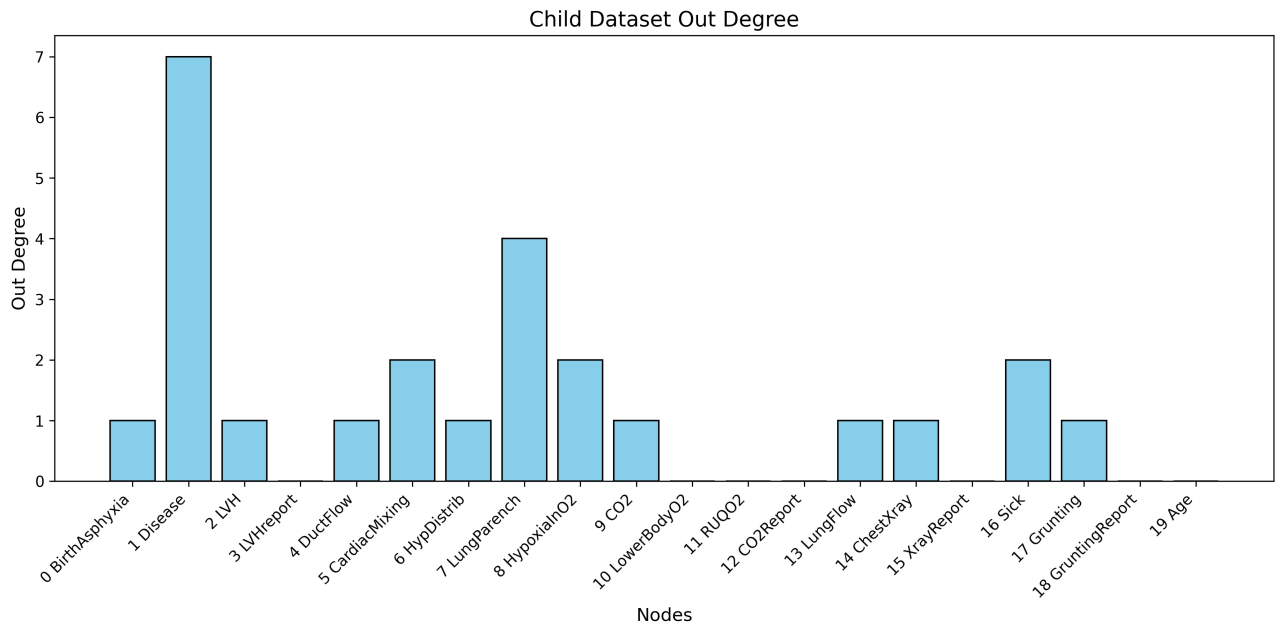


Figure 10. Out degree distribution of Child data.

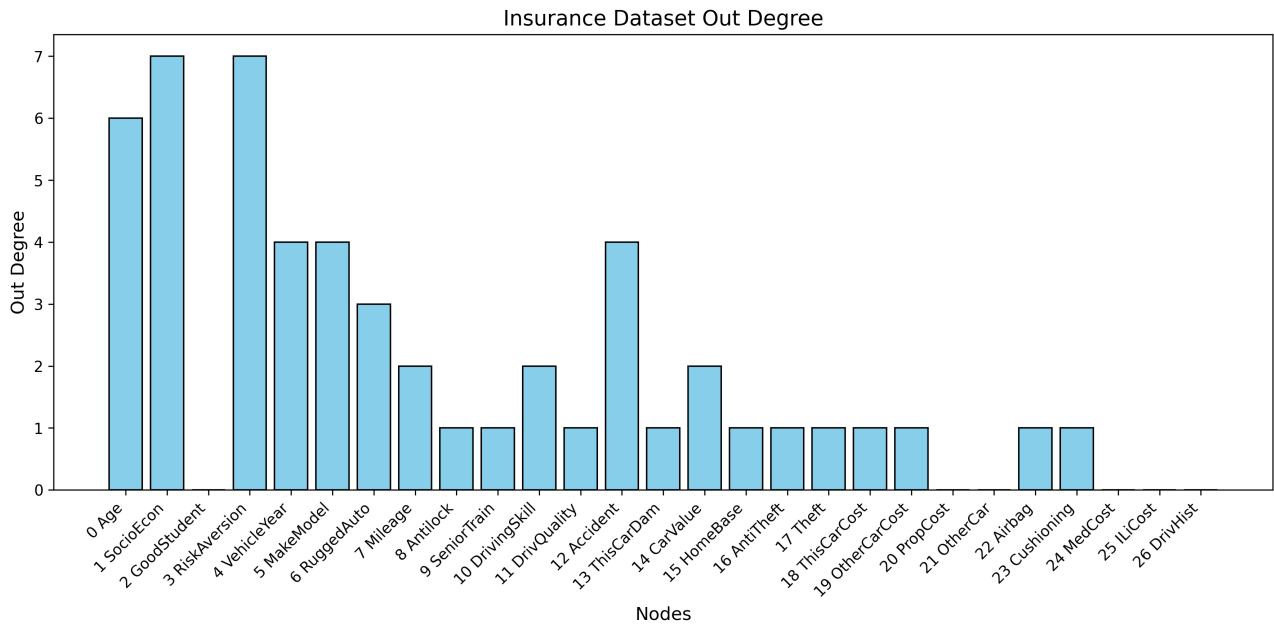


Figure 11. Out degree distribution of Insurance data.

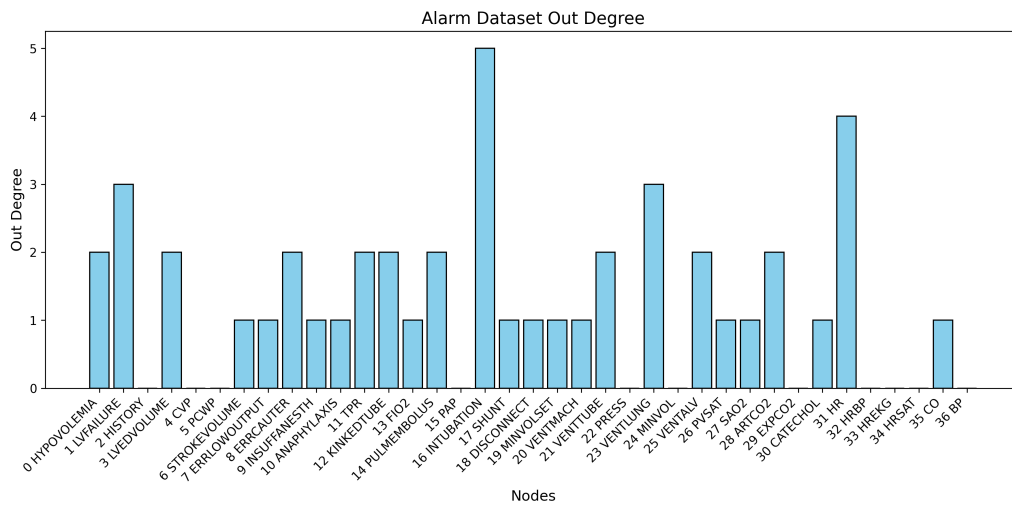


Figure 12. Out degree distribution of Alarm data.

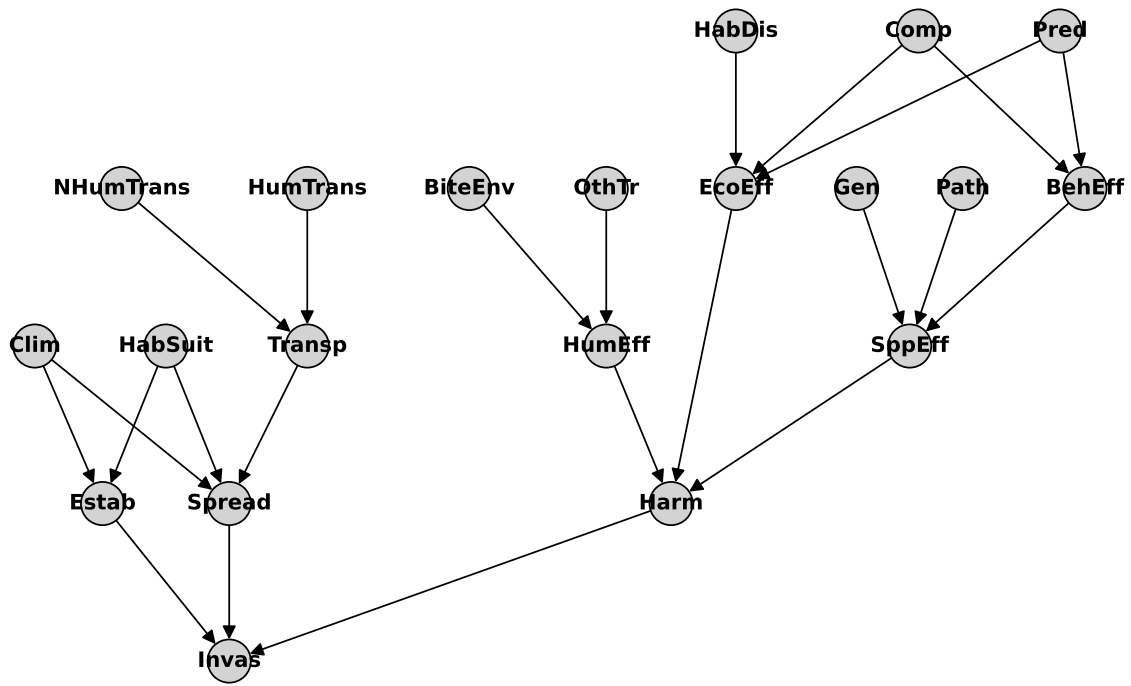


Figure 13. LeGIT final causal graph for Fisram dataset

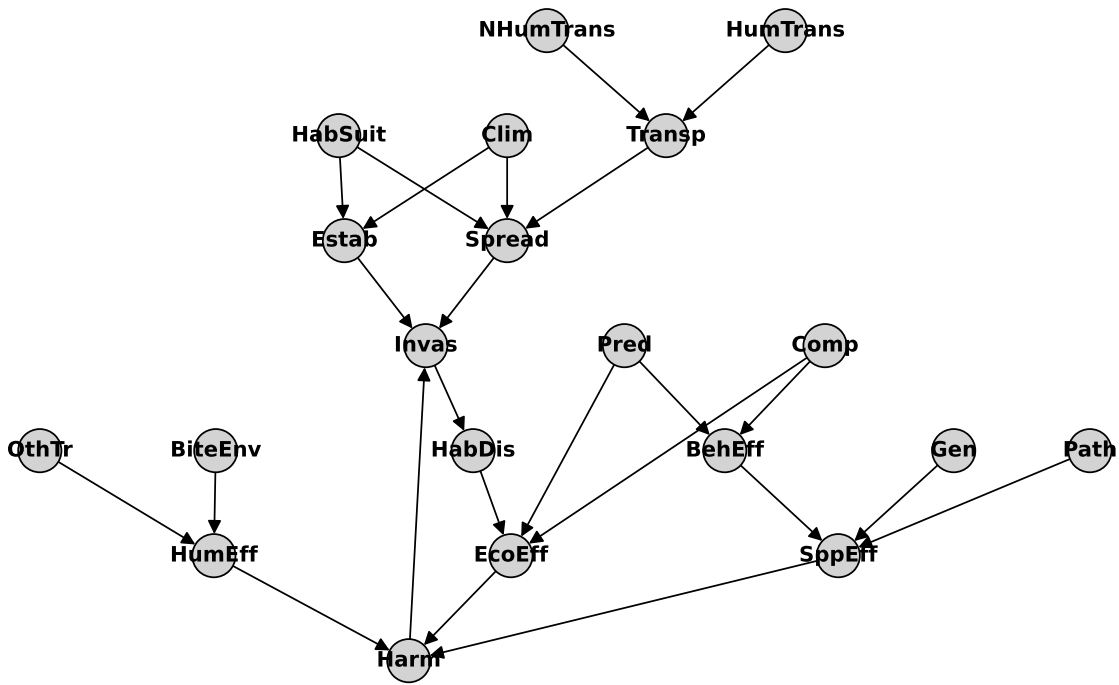


Figure 14. GIT's final causal graph for Fisram dataset

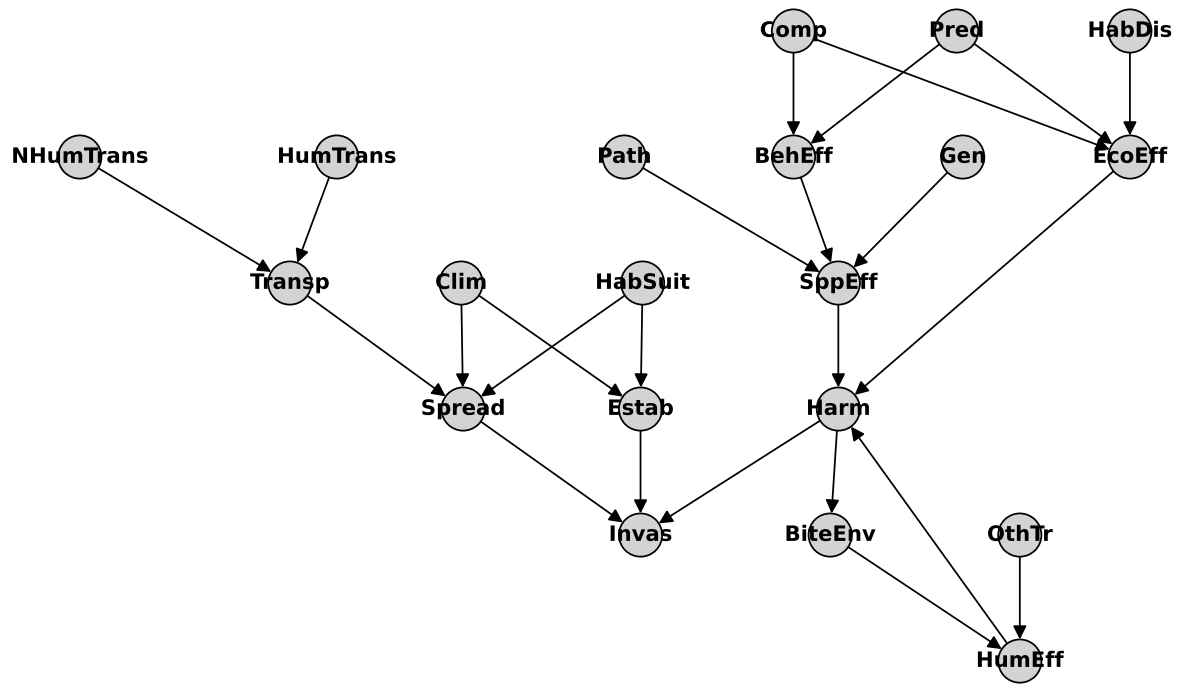


Figure 15. Huamn's final causal graph for Fisram dataset

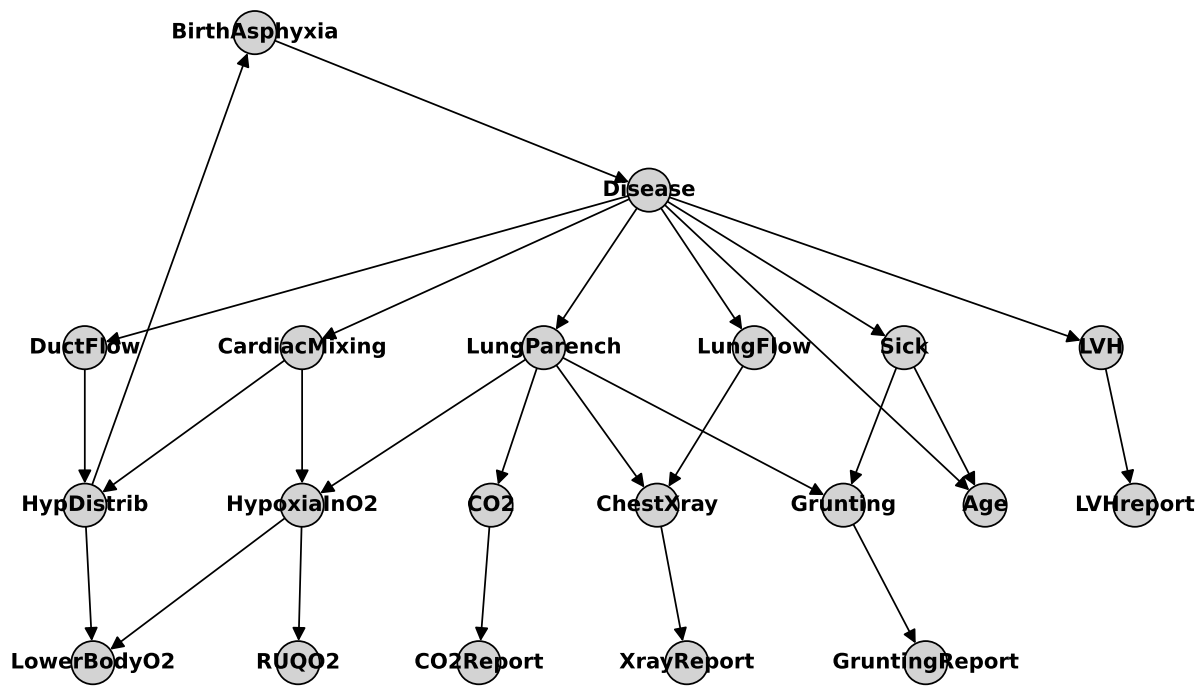


Figure 16. LeGIT final causal graph for Child dataset

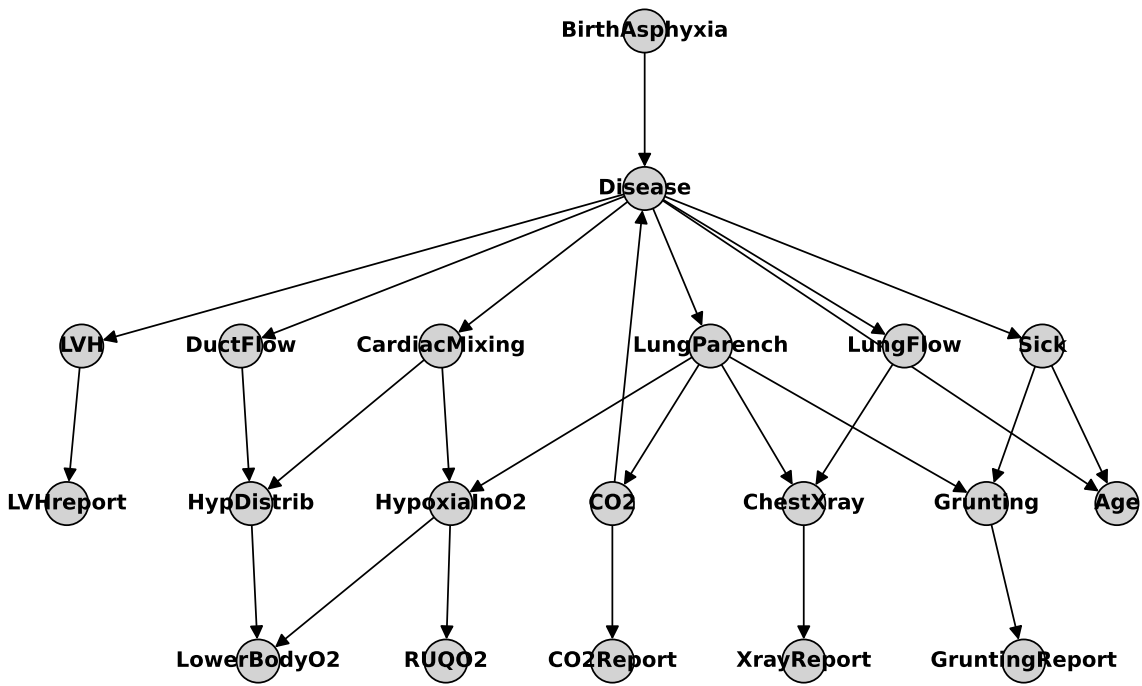


Figure 17. Human's final causal graph for child dataset

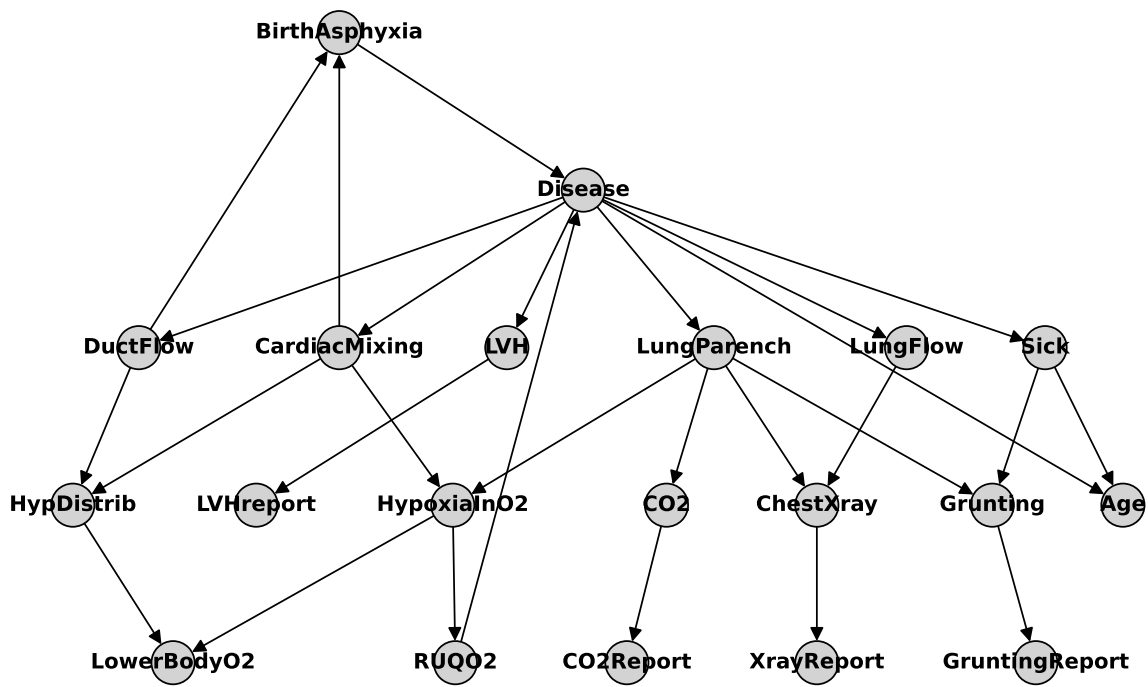


Figure 18. GIT's final causal graph for Child dataset

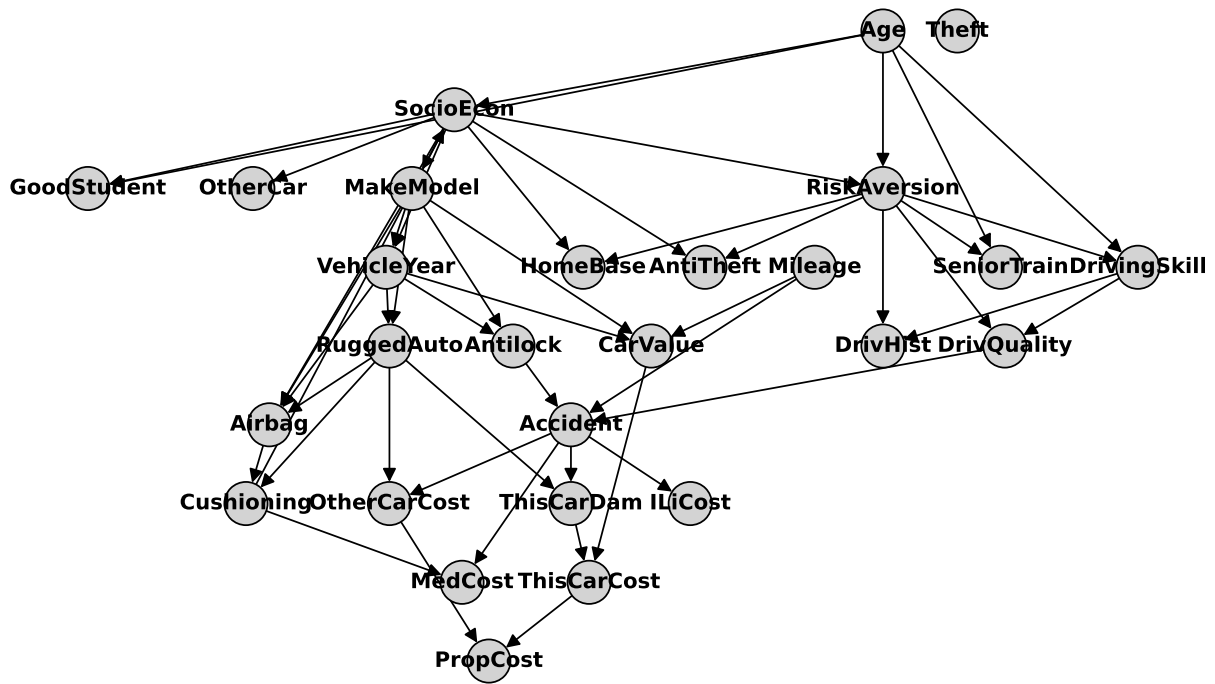


Figure 19. LeGIT final causal graph for Insurance dataset

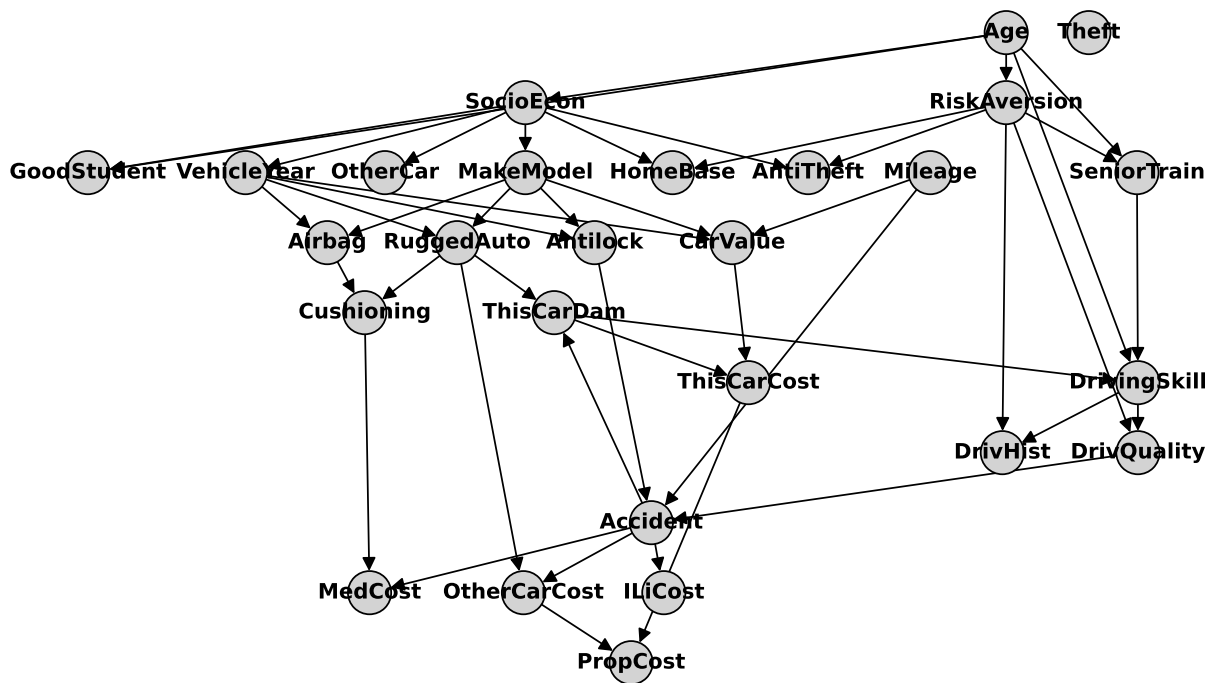


Figure 20. Human's final causal graph for Insurance dataset

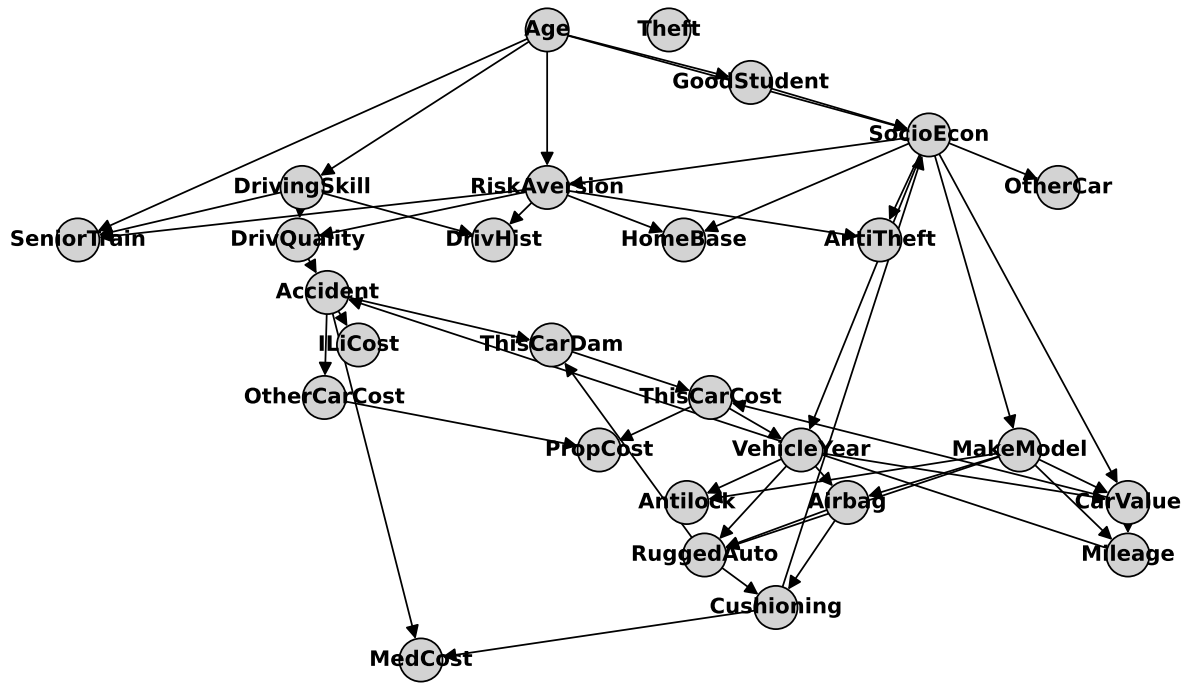


Figure 21. GIT's final causal graph for Insurance dataset

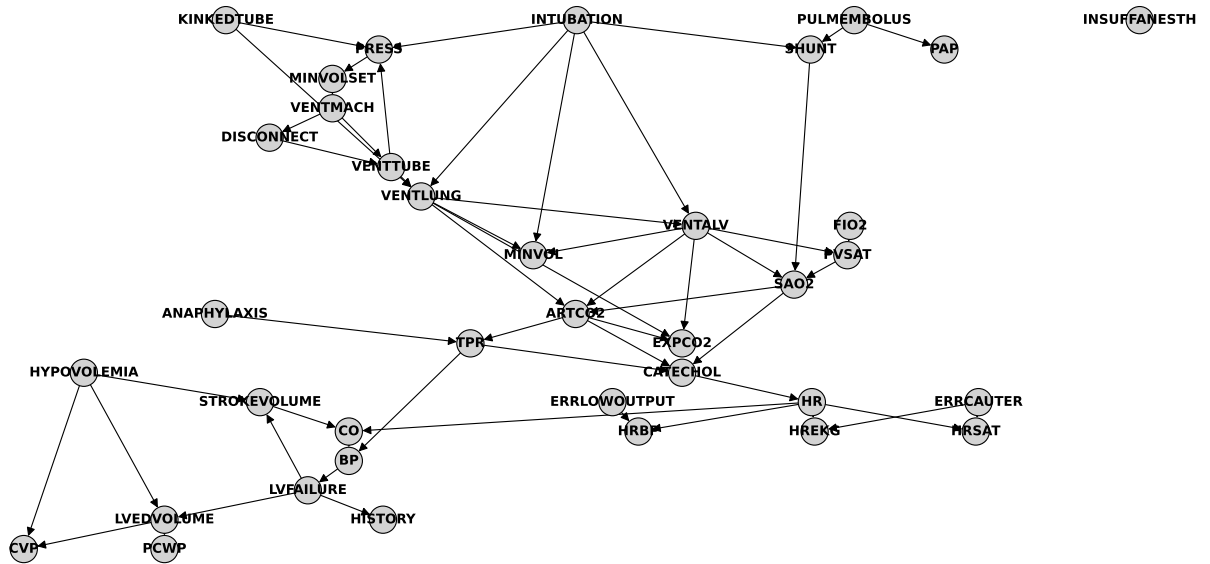


Figure 22. LeGIT final causal graph for Alarm dataset

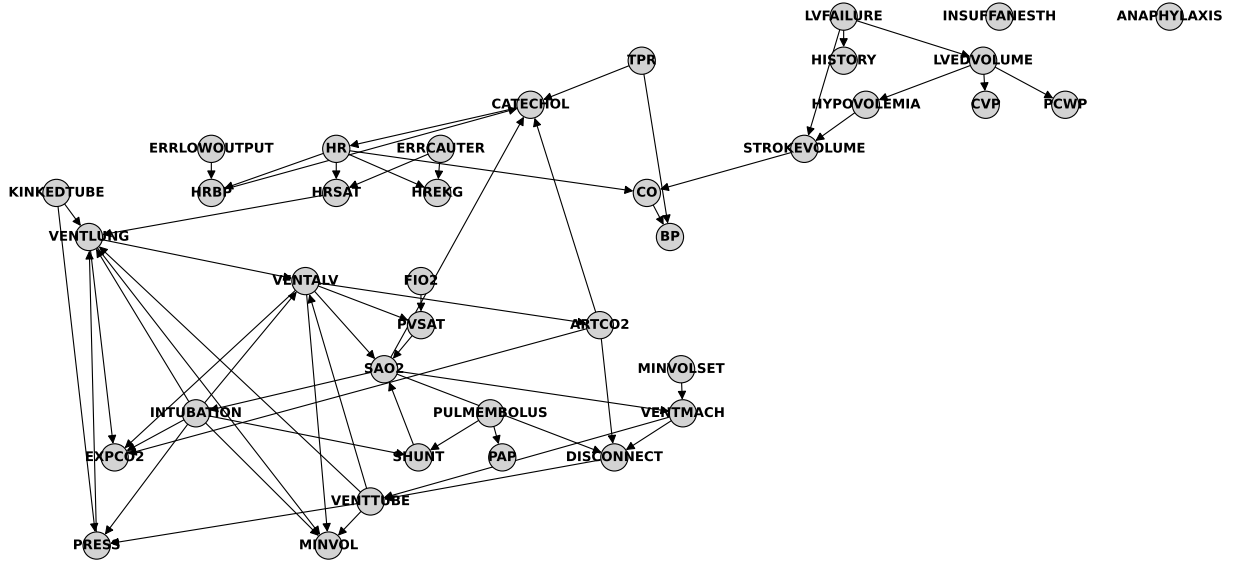


Figure 23. Human's final causal graph for Alarm dataset

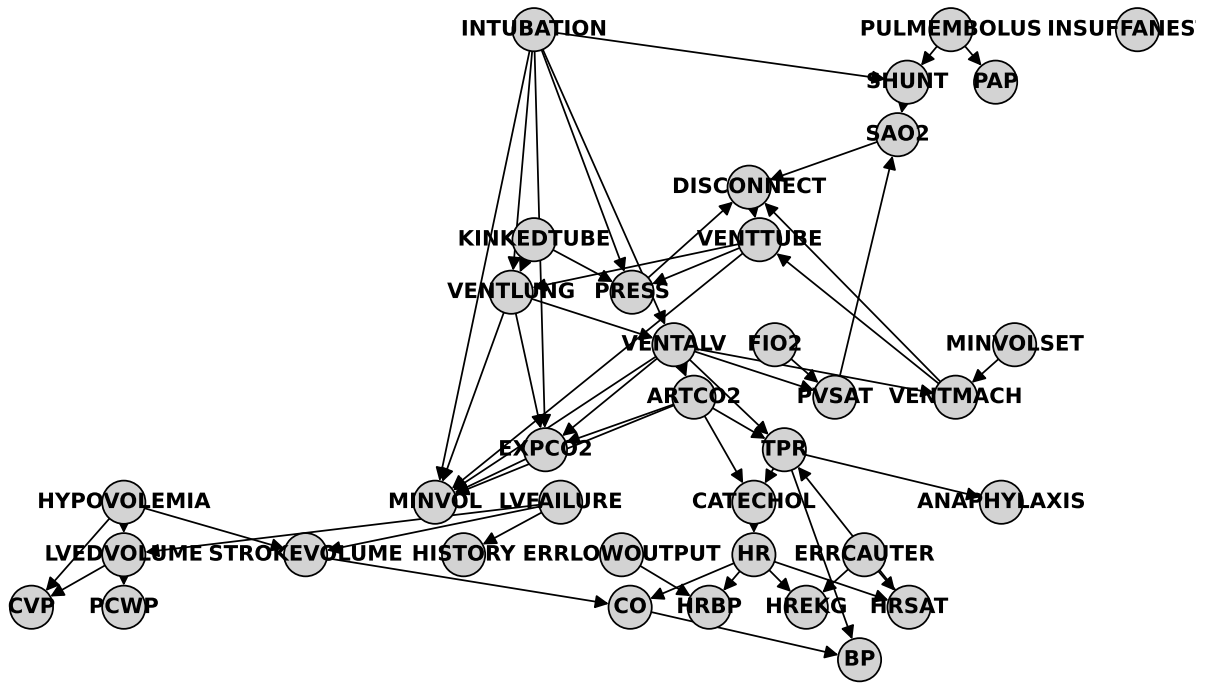


Figure 24. GIT's final causal graph for Alarm dataset

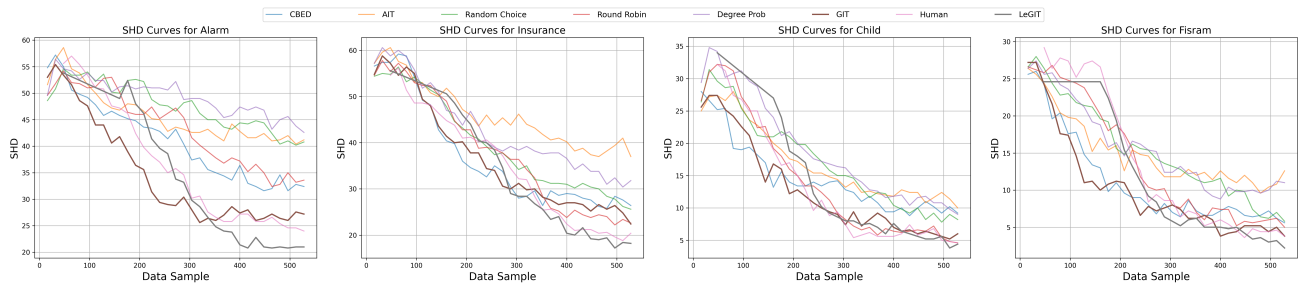


Figure 25. SHD metric for different methods (over 5 seeds) towards different intervention samples. ($T = 33$ rounds, $|D_{int}^I| = 32$, $N = 1056$)

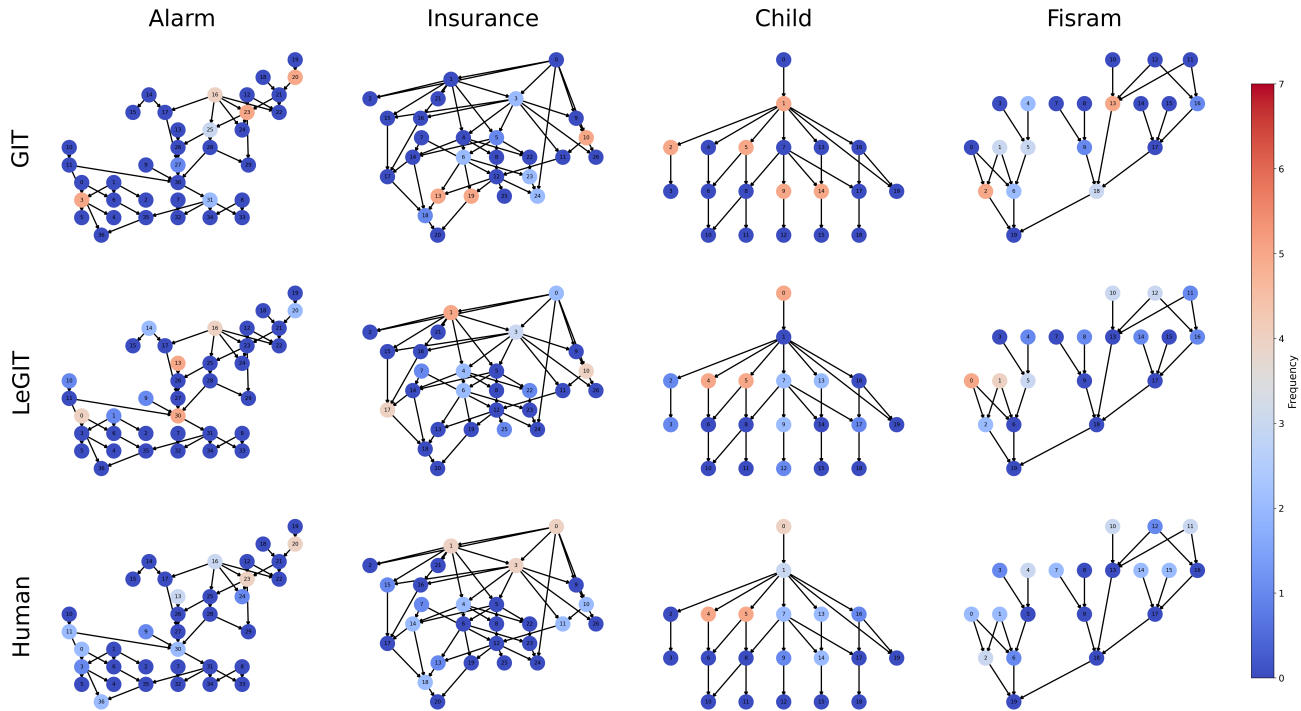


Figure 26. The selected Node Frequency obtained by different strategies on Epoch 0-4 under Table 3 setting.

D. Convergence of causal discovery with LeGIT

In this section, we provide a convergence argument for LeGIT, which combines a Large Language Model (LLM) warmup phase with a numerical-based intervention targeting strategy (e.g., GIT (Olko et al., 2023)).

D.1. Preliminaries and Notation

Structural Causal Models and Online Causal Discovery. We use the same definition of Structural Causal Models (SCMs), directed acyclic graphs (DAGs), and single-node interventions (hard interventions) described in, e.g., ENCO (Lippe et al., 2022). Suppose we have:

$$G^* = (V, E^*),$$

where $V = \{1, 2, \dots, n\}$ indexes the causal variables (X_1, \dots, X_n) , and we wish to recover G^* . In an online setting, at each round $t = 1, 2, \dots, T$, we choose an intervention target $i_t \in V$ and obtain a small batch of interventional samples from X under that intervention. This new interventional data is then used to update our current belief about the causal structure and functional parameters.

LeGIT and the Warmup Stage. LeGIT begins with a small number of *warmup* rounds, T_{warmup} , possibly augmented by an additional *bootstrap* stage $T_{\text{bootstrapped}}$. Across these initial stages, an LLM proposes intervention targets based on domain-specific descriptions or meta-information about the variables. Once the warmup phases are finished, the algorithm reverts to a purely numerical-based strategy for intervention selection—for example, GIT (Olko et al., 2023) or a Bayesian method (Brouillard et al., 2020). Let I_t be the random variable denoting the chosen intervention target at round t ; then

$$I_t = \begin{cases} \text{LLM-based selection, } t \leq 2 * (T_{\text{warmup}} + T_{\text{bootstrapped}}), \\ \text{numerical-based selection, } 2 * (T_{\text{warmup}} + T_{\text{bootstrapped}}) < t \leq T. \end{cases}$$

D.2. Convergence Proof

Throughout this section, we make the following standard assumptions:

- **A1 (Faithfulness).** The true distribution is *faithful* to a unique causal DAG G^* .
- **A2 (Sufficiency).** There are no hidden confounders, i.e., all relevant causal variables are observed.
- **A3 (Convergent Base-Algorithm).** After collecting a sufficient number of correct interventional data points, the base numerical method (e.g., GIT + ENCO) converges to G^* ; see original paper (Olko et al., 2023; Lippe et al., 2022) for formal statements.

Our main theorem shows that, under A1–A3, and given that the LLM-based warmup selects meaningful root-cause or high-influence variables in at least some fraction of the initial rounds, LeGIT converges to the correct graph G^* in the limit of acquiring more interventional data.

D.3. Key Lemma: Warmup Rounds Provide Informative Interventions

Let ϕ_{LLM} be the set of selected nodes from the LLM-based warmup. Although LLM selection can be imperfect, we show that with nonzero probability, ϕ_{LLM} contains enough influential/parent nodes to break symmetries or ambiguities in the MEC, thereby accelerating or guaranteeing eventual convergence.

Lemma D.1 (LLM Warmup is Sufficiently Informative). *Suppose that, during the warmup stage, the LLM selects a node v which is a direct cause (or ancestor) of at least one child c that is currently ambiguous or misoriented in the model. Intervening on v yields significant new information about the structure among $\{v, c\}$. If the warmup stage includes such interventions on enough distinct parents or high-degree nodes, the post-warmup structure has strictly fewer edges inconsistent with G^* on average.*

Proof. When the LLM intervenes on v , we obtain data from $P(X \mid do\{v\})$. Under A1–A2, the subsequent updates to the structural parameters will, with high probability, remove incorrect edges around v or correct orientation errors. Repeating

this for a sufficient set of v nodes with nontrivial out-degrees ensures that many orientation and adjacency ambiguities in G are resolved. A formal statement follows directly from standard identifiability arguments of single-node interventions. \square

D.4. Convergence Argument for the Combined Procedure

After $2(T_{\text{warmup}} + T_{\text{bootstrapped}})$ rounds, the base method (e.g., GIT) takes over and selects all subsequent targets $\{I_t\}_{t > 2(T_{\text{warmup}} + T_{\text{bootstrapped}})}$. The following proposition states that if the base method itself converges when given sufficiently many informative interventions (as assumed by A3), then the combination of warmup + base method must also converge.

Proposition D.2 (Convergence of LeGIT). *Assume the LLM-based warmup stage provides a non-empty set of interventions that reduce critical ambiguities in the causal structure (Lemma D.1). Let the base method be any procedure that is guaranteed to converge to G^* if it acquires sufficiently many samples from the relevant parts of the DAG (A3). Then, as $T \rightarrow \infty$, LeGIT converges to the correct causal DAG G^* with probability 1.*

Proof. By Lemma D.1, after the LLM warmup stage, the posterior space of graphs is already closer to the true DAG. That is, the number of structural ambiguities or misoriented edges around key high-influence nodes is reduced or completely resolved.

From round $t = 2(T_{\text{warmup}} + T_{\text{bootstrapped}}) + 1$ onward, the intervention targeting is dictated by the base method (e.g., GIT). Under assumption A3, we know that if GIT (plus the underlying gradient-based causal discovery algorithm like ENCO) is run on a system with sufficiently many informative interventions, it converges to G^* . Because the warmup stage of LeGIT has by design intervened on crucial nodes to reduce ambiguities, the base method from that point sees a significantly less confounded or ambiguous search space and, with high probability, chooses further interventions that refine the partial solution until it converges to G^* .

Thus, by the properties of the base method’s convergence proof (see (Olko et al., 2023)), the entire procedure (LLM warmup + GIT) converges to G^* given sufficient total rounds T . \square

Remark D.3 (Extension to Other Methods). *Although we have discussed GIT and ENCO as an illustrative example, any gradient-based or Bayesian-based method that ensures correct discovery given a suitable variety of interventions can replace GIT in Step (b) of LeGIT. Under the same conditions (A1–A3), the combined procedure likewise converges to the true DAG G^* .*

E. Examples of Prompts and Responses

For robust performance, we actually shuffle the order of variable descriptions following the self-consistency prompt skill. We provide the prompt templates and the description of the variables used in LeGIT below.

You are a helpful assistant and expert in Car Insurance system research. Assuming we can do interventions to all the variables, your job is to assist in designing the best intervention experiments among the following variables to help discover variables causal relations:

<variable name>: Variable descriptions

Assuming we can do interventions to all the variables, given the aforementioned variables and their descriptions, can you echo your knowledge about those variables, temporally analyze their relations, and then choose the best 5 intervention targets from all the variables which hopefully are the root causes of the other variables to start our analysis of their causal relations?

Let’s think and analyze step by step. Then, provide your final answer (variable names only) within the tags <Answer>...</Answer>, separated by “;”

Figure 27. Prompt template at warmup stage.

Fisram Warmup Prompt

You are a helpful assistant and expert in Freshwater Fish Injurious Species Risk Assessment Model system research. Here are some tips that you can pay attention to:

1. Assess whether there is a direct causal relationship, and consider potential confounding variables that might affect the relationship that could potentially not causal relationship.
2. Distinguish between correlations and causation; verify that correlations are not mistaken for causal relationships.
3. Ensure the correct temporal order of variables; confirm that the cause precedes the effect.

Assuming we can do interventions to all the variables, your job is to assist in designing the best intervention experiments among the following variables to help discover their causal relations:

<OthTr>: Non-bite/toxin traits posing human-health risks (e.g., zoonotic pathogens, physical injury from leaping species).

<Harm>: Actual or potential physical or behavioral injury to native species and/or humans, or damage to habitats.

<BehEff>: Combined effect of predation and competition on native species behavior and viability.

<EcoEff>: Overall impact of habitat disturbance, predation, and competition on ecosystem structure and function.

<Clim>: Sum of counts for climate similarity scores 6–10 divided by the sum of all climate scores, as calculated by the CLIMATCH or RAMP tools.

<HumEff>: Combined influence of bites/toxins and other detrimental traits on humans.

<Estab>: Actual or potential for self-sustaining wild populations based on climate and habitat inputs.

<NHumTrans>: Dispersal assistance by natural agents (wind, water, animals) beyond the species' own movement.

<HumTrans>: Intentional or unintentional movement by humans (e.g., trade, ballast water, recreational stocking).

<BiteEnv>: Direct adverse effects on human health via bites, stings, toxins, injections, ingestion, or absorption.

<Pred>: Capacity to prey on and negatively affect native species populations.

<Gen>: Capacity to affect native species' genetics via hybridization, GMO escape, or introgression.

<HabDis>: Capacity to modify or degrade habitat (erosion, eutrophication, sedimentation).

<Invas>: Final invasive-injurious outcome under the Lacey Act criteria, integrating Establishment, Spread, and Harm. <Comp>: Capacity to compete with native species for food, space, or habitat.

<SpEff>: Overall impact of predation, competition, and genetics on native species viability.

<Path>: Role in spreading infectious agents (bacteria, viruses, parasites, fungi) to native wildlife.

<HabSuit>: Degree to which available habitat in the potential introduction area matches the species' known habitats.

<Spread>: Actual or potential spatial expansion across ecosystems, driven by climate, habitat, and transport.

<Transp>: Combined human and non-human dispersal influence.

Assuming we can do interventions to all the variables, given the aforementioned variables and their descriptions, can you **echo your knowledge those variables**, **temporally analyze** their relations, and then **choose the best 4 intervention targets from all the variables** which hopefully are the root causes of the other variables to start our analysis of their causal relations?

Let's think and analyze step by step. Then, provide your final answer (variable names only) within the tags <Answer>...</Answer>, separated by ", ".

Child Warmup Prompt

You are a helpful assistant and expert in children's disease research. Here are some tips that you can pay attention to:

1. Assess whether there is a direct causal relationship, and consider potential confounding variables that might affect the relationship that could potentially not causal relationship.
2. Distinguish between correlations and causation; verify that correlations are not mistaken for causal relationships.
3. Ensure the correct temporal order of variables; confirm that the cause precedes the effect.

Assuming we can do interventions to all the variables, your job is to assist in designing the best intervention experiments among the following variables to help discover their causal relations:

<LungFlow>: low blood flow in the lungs

<ChestXray>: having a chest x-ray

<Disease>: infant methemoglobinemia

<Grunting>: grunting in infants

<Age>: age of infant at disease presentation

<XrayReport>: lung excessively filled with blood
<RUQO2>: level of oxygen in the right upper quadriceps muscle
<DuctFlow>: blood flow across the ductus arteriosus
<HypoxiaInO2>: hypoxia when breathing oxygen
<Sick>: presence of an illness
<CO2Report>: a document reporting high level of CO2 levels in blood
<LungParench>: the state of the blood vessels in the lungs
<LVH>: having left ventricular hypertrophy
<LowerBodyO2>: level of oxygen in the lower body
<BirthAsphyxia>: lack of oxygen to the blood during the infant's birth
<CO2>: level of CO2 in the body
<LVHreport>: report of having left ventri
<GruntingReport>: report of infant grunting
<CardiacMixing>: mixing of oxygenated and deoxygenated blood
<HypDistrib>: low oxygen areas equally distributed around the body

Assuming we can do interventions to all the variables, given the aforementioned variables and their descriptions, can you **echo your knowledge those variables**, **temporally analyze** their relations, and then **choose the best 4 intervention targets from all the variables** which hopefully are the root causes of the other variables to start our analysis of their causal relations?

Let's think and analyze step by step. Then, provide your final answer (variable names only) within the tags <Answer>...</Answer>, separated by ", ".

Insurance Warmup Prompt

You are a helpful assistant and expert in car insurance risks research. Here are some tips that you can pay attention to:

1. Assess whether there is a direct causal relationship, and consider potential confounding variables that might affect the relationship that could potentially not causal relationship.
2. Distinguish between correlations and causation; verify that correlations are not mistaken for causal relationships.
3. Ensure the correct temporal order of variables; confirm that the cause precedes the effect.

Assuming we can do interventions to all the variables, your job is to assist in designing the best intervention experiments among the following variables to help discover their causal relations:

<ThisCarDam>: damage to the car
<MakeModel>: owning a sports car
<OtherCarCost>: cost of the other cars
<PropCost>: ratio of the cost for the two cars
<AntiTheft>: car has anti-theft
<DrivQuality>: driving quality
<DrivHist>: driving history
<MedCost>: cost of medical treatment
<Mileage>: how much mileage is on the car
<Antilock>: car has anti-lock
<CarValue>: value of the car
<Accident>: severity of the accident
<OtherCar>: being involved with other cars in the accident
<SeniorTrain>: received additional driving training
<ILiCost>: inspection cost
<SocioEcon>: socioeconomic status
<Theft>: theft occurred in the car
<Age>: age
<RuggedAuto>: ruggedness of the car
<GoodStudent>: being a good student driver

<VehicleYear>: year of vehicle
<HomeBase>: neighbourhood type
<ThisCarCost>: costs for the insured car
<Cushioning>: quality of cushioning in car
<RiskAversion>: being risk averse
<DrivingSkill>: driving skill
<Airbag>: car has an airbad

Assuming we can do interventions to all the variables, given the aforementioned variables and their descriptions, can you **echo your knowledge those variables**, **temporally analyze** their relations, and then **choose the best 4 intervention targets from all the variables** which hopefully are the root causes of the other variables to start our analysis of their causal relations?

Let's think and analyze step by step. Then, provide your final answer (variable names only) within the tags <Answer>...</Answer>, separated by ", ".

Alarm Warmup Prompt

You are a helpful assistant and expert in alarm message system for patient monitoring system research. Here are some tips that you can pay attention to:

1. Assess whether there is a direct causal relationship, and consider potential confounding variables that might affect the relationship that could potentially not causal relationship.
2. Distinguish between correlations and causation; verify that correlations are not mistaken for causal relationships.
3. Ensure the correct temporal order of variables; confirm that the cause precedes the effect.

Assuming we can do interventions to all the variables, your job is to assist in designing the best intervention experiments among the following variables to help discover their causal relations:

<CATECHOL>: hormone made by the adrenal glands
<SAO2>: oxygen saturation of arterial blood
<VENTALV>: exchange of gas between the alveoli and the external environment
<ANAPHYLAXIS>: sever, life-threatening allergic reaction
<INSUFFANESTH>: whether there is insufficient anesthesia or not
<FIO2>: the concentration of oxygen in the gas mixture being inspired
<BP>: pressure of circulating blood against the walls of blood vessels
<PRESS>: breathing pressure
<VENTTUBE>: whether there is a breathing tube or not
<TPR>: amount of force exerted on circulating blood by vasculature of the body
<CO>: amount of blood pumped by the heart per minute
<PCWP>: pulmonary capillary wedge pressure
<ERRCAUTER>: whether there was an error during cautery or not
<KINKEDTUBE>: whether the chest tube is kinked or not
<PVSAT>: amount of oxygen bound to hemoglobin in the pulmonary artery
<INTUBATION>: process where a healthcare provider inserts a tube through a person's mouth or nose, then down into their trachea
<CVP>: measure of blood pressure in the vena cava
<HYPOVOLEMIA>: condition that occurs when your body loses fluid, like blood or water
<HRBP>: ratio of heart rate and blood pressure
<HREKG>: heart rate displayed on EKG monitor
<PAP>: blood pressure in the pulmonary artery
<EXPCO2>: expelled CO2
<ERRLOWOUTPUT>: error low output
<HISTORY>: previous medical history
<SHUNT>: hollow tube surgically placed in the brain (or occasionally in the spine) to help drain cerebrospinal fluid and redirect it to another location in the body where it can be reabsorbed
<VENTMACH>: the intensity level of a breathing machine

<VENTLUNG>: lung ventilation
<HRSAT>: measure of how much hemoglobin is currently bound to oxygen compared to how much hemoglobin remains unbound
<LVFAILURE>: occurs when there is dysfunction of the left ventricle causing insufficient delivery of blood to vital body organs
<DISCONNECT>: disconnection
<LVEDVOLUME>: amount of blood present in the left ventricle before contraction
<HR>: heart rate
<MINVOLSET>: the amount of time using a breathing machine
<PULMEMBOLUS>: sudden blockage in the pulmonary arteries, the blood vessels that send blood to your lungs
<STROKEVOLUME>: volume of blood pumped out of the left ventricle of the heart during each systolic cardiac contraction
<MINVOL>: amount of gas inhaled or exhaled from a person's lungs in one minute
<ARTCO2>: arterial carbon dioxide

Assuming we can do interventions to all the variables, given the aforementioned variables and their descriptions, can you **echo your knowledge** those variables, **temporally analyze** their relations, and then **choose the best 4** intervention targets from all the variables which hopefully are the root causes of the other variables to start our analysis of their causal relations?

Let's think and analyze step by step. Then, provide your final answer (variable names only) within the tags <Answer>...</Answer>, separated by ", ".

LeGIT response for Fisram dataset

Let's start by echoing the variables and their descriptions:

1. **Spread**: Refers to the actual or potential spatial expansion across ecosystems, driven by climate, habitat, and transport.
2. **Estab**: Indicates the actual or potential for self-sustaining wild populations based on climate and habitat inputs.
3. **BehEff**: Represents the combined effect of predation and competition on native species behavior and viability.
4. **Clim**: Measures climate similarity scores to assess potential habitat suitability.
5. **HumEff**: Describes the combined influence of bites/toxins and other detrimental traits on humans.
6. **HabDis**: Capacity to modify or degrade habitat (erosion, eutrophication, sedimentation).
7. **EcoEff**: Overall impact of habitat disturbance, predation, and competition on ecosystem structure and function.
8. **Gen**: Capacity to affect native species' genetics via hybridization, GMO escape, or introgression.
9. **Harm**: Actual or potential physical or behavioral injury to native species and/or humans, or damage to habitats.
10. **Invas**: Final invasive-injurious outcome under the Lacey Act criteria, integrating Establishment, Spread, and Harm.
11. **Transp**: Combined human and non-human dispersal influence.
12. **Comp**: Capacity to compete with native species for food, space, or habitat.
13. **Pred**: Capacity to prey on and negatively affect native species populations.
14. **HabSuit**: Degree to which available habitat in the potential introduction area matches the species' known habitats.
15. **Path**: Role in spreading infectious agents to native wildlife.
16. **SppEff**: Overall impact of predation, competition, and genetics on native species viability.
17. **OthTr**: Non-bite/toxin traits posing human-health risks.
18. **NHumTrans**: Dispersal assistance by natural agents beyond the species' own movement.
19. **HumTrans**: Intentional or unintentional movement by humans.
20. **BiteEnv**: Direct adverse effects on human health via bites, stings, toxins, etc.

Now let's analyze the temporal relationships and potential causal links:

1. **Clim** and **HabSuit** are likely early indicators of whether a species can establish (Estab) and spread (Spread), as they relate to environmental suitability.
2. **Transp**, **HumTrans**, and **NHumTrans** are likely influential in the spread of species (Spread), as they involve movement and dispersal.
3. **Estab** leads to **Spread** as a species that can establish will likely expand its range.
4. **Spread** and **Estab** can contribute to **Invas**, as they are integral parts of invasive potential.
5. **BehEff**, **Comp**, and **Pred** are likely results of established populations interacting with native species.
6. **HabDis**, **EcoEff**, and **Harm** might be consequences of established and spreading populations.
7. **Gen** might affect native species' genetics once a species is established and interacting.
8. **Path** might be influenced by **Spread**, as wider distribution can lead to more interactions and potential for disease spread.
9. **HumEff**, **OthTr**, and **BiteEnv** are related to human impacts and may be less directly connected to ecological establishment and spread.

Considering the temporal order and potential causation:

- **Clim** and **HabSuit** are foundational for establishment and spread.
- **Transp** is crucial for movement and dispersal.
- **Estab** is a precursor to many other impacts.

Therefore, the best intervention targets to start analyzing causal relations are those that influence establishment and spread directly, as they are likely root causes:

<Answer>Clim, HabSuit, Transp, Estab</Answer>

LeGIT response for Child dataset

To address the problem effectively, we will follow a structured approach to understand the potential causal relationships among the variables. We will start by echoing the knowledge of each variable, analyze their temporal relationships, and then identify the best intervention targets.

Echoing Knowledge of Variables

1. **RUQO2**: Level of oxygen in the right upper quadriceps muscle.
2. **LVHreport**: Report of having left ventricular hypertrophy.
3. **BirthAsphyxia**: Lack of oxygen to the blood during the infant's birth.
4. **DuctFlow**: Blood flow across the ductus arteriosus.
5. **Disease**: Infant methemoglobinemia.
6. **LungParench**: The state of the blood vessels in the lungs.
7. **CO2Report**: A document reporting high levels of CO2 in blood.
8. **LungFlow**: Low blood flow in the lungs.
9. **GruntingReport**: Report of infant grunting.
10. **Age**: Age of infant at disease presentation.
11. **LowerBodyO2**: Level of oxygen in the lower body.
12. **ChestXray**: Having a chest x-ray.
13. **CO2**: Level of CO2 in the body.
14. **HypDistrib**: Low oxygen areas equally distributed around the body.
15. **Grunting**: Grunting in infants.
16. **XrayReport**: Lung excessively filled with blood.
17. **LVH**: Having left ventricular hypertrophy.
18. **CardiacMixing**: Mixing of oxygenated and deoxygenated blood.
19. **HypoxiaInO2**: Hypoxia when breathing oxygen.
20. **Sick**: Presence of an illness.

Temporal Analysis

1. **BirthAsphyxia** is an early event that can lead to multiple downstream effects, such as hypoxia and potential cardiac issues.

2. **CardiacMixing** and **DuctFlow** are congenital or physiological conditions that can affect oxygenation and blood flow.
3. **HypoxiaInO2**, **LungFlow**, and **LowerBodyO2** are likely influenced by earlier conditions like **BirthAsphyxia**, **CardiacMixing**, and **DuctFlow**.
4. **CO2** and **CO2Report** are indicators of respiratory function and can be affected by lung and cardiac issues.
5. **LungParench**, **XrayReport**, and **ChestXray** are diagnostic or descriptive of lung conditions.
6. **Grunting** and **GruntingReport** are symptoms that may arise from respiratory distress.
7. **LVH** and **LVHreport** are indicative of cardiac stress or abnormalities, potentially secondary to other physiological issues.
8. **Disease** (infant methemoglobinemia) could be a result of hypoxic conditions and other metabolic issues.
9. **Sick** is a general indicator of illness and can be influenced by multiple preceding factors.

Choosing Intervention Targets

To identify root causes, we need to target variables that likely influence many others and are early in the causal chain:

1. **BirthAsphyxia**: Early event potentially causing hypoxia, cardiac stress, and other downstream effects.
2. **CardiacMixing**: Affects oxygenation and can lead to hypoxia and other cardiac issues.
3. **DuctFlow**: Influences blood flow and oxygenation, impacting many other variables.
4. **LungFlow**: Directly affects oxygenation and can lead to hypoxia and respiratory distress.

These variables are likely to be root causes or early influencers in the causal network, making them ideal intervention targets to explore causal relationships.

Final Answer

<Answer>BirthAsphyxia, CardiacMixing, DuctFlow, LungFlow</Answer>

LeGIT response for Insurance dataset

To analyze the causal relationships among these variables and select the best intervention targets, let's first categorize and understand the potential causal structure among them.

Step 1: Echoing Knowledge of Variables

1. **<ThisCar>**: Costs for the insured car
2. **<Mileage>**: Mileage on the car
3. **<DrivQuality>**: Driving quality
4. **<RiskAversion>**: Being risk-averse
5. **<CarValue>**: Value of the car
6. **<MakeModel>**: Owning a sports car
7. **<ThisCarDam>**: Damage to the car
8. **<Theft>**: Theft occurred in the car
9. **<ILiCost>**: Inspection cost
10. **<Age>**: Age
11. **<OtherCar>**: Involvement with other cars in an accident
12. **<SeniorTrain>**: Received additional driving training
13. **<SocioEcon>**: Socioeconomic status
14. **<VehicleYear>**: Year of vehicle
15. **<HomeBase>**: Neighborhood type
16. **<DrivHist>**: Driving history
17. **<Accident>**: Severity of the accident
18. **<Airbag>**: Car has an airbag
19. **<AntiTheft>**: Car has anti-theft
20. **<ThisCarCost>**: Costs for the insured car
21. **<MedCost>**: Cost of medical treatment
22. **<Antilock>**: Car has anti-lock
23. **<GoodStudent>**: Being a good student driver

24. **<OtherCarCost>**: Cost of the other cars
25. **<Cushioning>**: Quality of cushioning in the car
26. **<RuggedAuto>**: Ruggedness of the car
27. **<PropCost>**: Ratio of the cost for the two cars
28. **<DrivingSkill>**: Driving skill

Step 2: Temporal Analysis of Relations

Demographic and Personal Factors:

<Age>, **<RiskAversion>**, **<SocioEcon>**, **<GoodStudent>**: These are inherent traits or long-term conditions that can influence driving behavior and decisions.

Vehicle Characteristics:

<VehicleYear>, **<MakeModel>**, **<CarValue>**, **<Airbag>**, **<AntiTheft>**, **<Antilock>**, **<Cushioning>**, **<RuggedAuto>**: These are intrinsic to the car and can influence safety and risk but are typically decided before other events.

Driving Experience and Skills:

<DrivQuality>, **<DrivingSkill>**, **<DrivHist>**, **<SeniorTrain>**: These can evolve over time and influence accident likelihood.

Environmental Factors:

<HomeBase>: Affects exposure to theft and accidents.

Incident and Cost Factors:

<ThisCarDam>, **<Theft>**, **<Accident>**, **<ILiCost>**, **<ThisCarCost>**, **<MedCost>**, **<OtherCarCost>**, **<PropCost>**: These are outcomes or costs associated with incidents.

Step 3: Choosing Intervention Targets

To identify root causes, we should consider intervening on variables that are likely to influence many others and are not themselves primarily effects of other variables. The best interventions would be on variables that are upstream in the causal chain and can potentially change multiple downstream outcomes.

1. **<RiskAversion>**: Can influence driving behavior, choice of car, and safety measures.
2. **<SocioEcon>**: Affects car choice, home base, and potentially driving habits.
3. **<MakeModel>**: Influences car value, safety features, and possibly driving style.
4. **<DrivingSkill>**: Directly impacts driving quality, accident likelihood, and associated costs.

These variables are chosen because they are fundamental characteristics or decisions that precede and potentially influence many other variables in the system.

Final Answer

<Answer>RiskAversion, SocioEcon, MakeModel, DrivingSkill**</Answer>**

LeGIT response for Insurance dataset

To determine the best intervention targets, we need to consider the relationships and possible temporal orders among the variables. Here's a step-by-step analysis:

Step 1: Understanding the Variables

Respiratory and Ventilation Variables:

- **<VENTMACH>**: Intensity level of a breathing machine.
- **<MINVOLSET>**: Time using a breathing machine.
- **<VENTLUNG>**: Lung ventilation.
- **<PRESS>**: Breathing pressure.
- **<MINVOL>**: Minute volume; gas inhaled/exhaled per minute.
- **<VENTTUBE>**: Presence of a breathing tube.
- **<VENTALV>**: Gas exchange in alveoli.
- **<DISCONNECT>**: Disconnection.
- **<EXPCO2>**: Expelled CO2.
- **<ARTCO2>**: Arterial carbon dioxide.
- **<FIO2>**: Oxygen concentration in inspired gas.

Cardiovascular Variables:

- **<HREKG>**: Heart rate on EKG.
- **<PCWP>**: Pulmonary capillary wedge pressure.
- **<HRBP>**: Ratio of heart rate and blood pressure.
- **<BP>**: Blood pressure.
- **<CO>**: Cardiac output.
- **<HR>**: Heart rate.
- **<TPR>**: Total peripheral resistance.
- **<STROKEVOLUME>**: Blood volume per heart contraction.
- **<LVEDVOLUME>**: Blood in left ventricle before contraction.
- **<PAP>**: Pulmonary artery pressure.
- **<CVP>**: Central venous pressure.
- **<SAO2>**: Oxygen saturation of arterial blood.
- **<PVSAT>**: Oxygen saturation in pulmonary artery.
- **<HRSAT>**: Hemoglobin oxygen saturation.
- **<CATECHOL>**: Adrenal hormones.

3. **Medical Conditions and Procedures:**

- **<ANAPHYLAXIS>**: Severe allergic reaction.
- **<INTUBATION>**: Insertion of a breathing tube.
- **<INSUFFANESTH>**: Insufficient anesthesia.
- **<PULMEMBOLUS>**: Pulmonary embolism.
- **<HYPOVOLEMIA>**: Fluid loss condition.
- **<LVFAILURE>**: Left ventricle failure.
- **<SHUNT>**: Drainage tube for cerebrospinal fluid.
- **<KINKEDTUBE>**: Kinked chest tube.

4. **Errors and Anomalies:**

- **<ERRLOWOUTPUT>**: Low output error.
- **<ERRCAUTER>**: Cautery error.

5. **Other:** - **<HISTORY>**: Previous medical history.

Step 2: Temporal Analysis

- **<INTUBATION>** and **<VENTMACH>** are likely initial interventions that can influence many respiratory and cardiovascular variables.
- **<ANAPHYLAXIS>**, **<PULMEMBOLUS>**, and **<HYPOVOLEMIA>** are acute conditions that can cause significant changes in cardiovascular and respiratory variables.
- **<INSUFFANESTH>** can affect heart rate and blood pressure due to inadequate sedation.
- **<VENTLUNG>**, **<VENTALV>**, and **<MINVOL>** are outcomes of ventilation interventions.

Step 3: Choosing the Best Intervention Targets

To determine causality effectively, we want to target variables that are likely root causes or significant influencers:

1. **<INTUBATION>**: This is a direct intervention that can affect many respiratory variables.
2. **<VENTMACH>**: Controls the intensity of mechanical ventilation, influencing respiratory and possibly cardiovascular variables.
3. **<ANAPHYLAXIS>**: A condition that can have widespread effects on cardiovascular and respiratory systems.
4. **<HYPOVOLEMIA>**: Can significantly impact cardiovascular variables like blood pressure and cardiac output.

These interventions can provide insights into the causal relationships within the system.

Final Answer

<Answer>INTUBATION, VENTMACH, ANAPHYLAXIS, HYPOVOLEMIA**</Answer>**