Can LLMs Leverage Observational Data? Towards Data-Driven Causal Discovery with LLMs

Yuni Susanti[1,*], Michael Färber[2]

[1]Fujitsu Limited, Japan
[2]ScaDS.AI & TU Dresden, Germany

Abstract

Causal discovery traditionally relies on statistical methods applied to observational data, often requiring large datasets and assumptions about underlying causal structures. Recent advancements in Large Language Models (LLMs) have introduced new possibilities for causal discovery by providing domain expert knowledge. However, it remains unclear whether LLMs can effectively process observational data for causal discovery. In this work, we explore the potential of LLMs for data-driven causal discovery by integrating observational data for LLM-based reasoning. Specifically, we examine whether LLMs can effectively utilize observational data through two prompting strategies: *pairwise* prompting and *breadth first search* (BFS)-based prompting. In both approaches, we incorporate the observational data directly into the prompt to assess LLMs' ability to infer causal relationships from such data. Experiments on benchmark datasets show that incorporating observational data enhances causal discovery, boosting F1 scores by up to 0.11 point using both pairwise and BFS LLM-based prompting, while outperforming traditional statistical causal discovery baseline by up to 0.52 points. Our findings highlight the potential and limitations of LLMs for data-driven causal discovery, demonstrating their ability to move beyond textual metadata and effectively utilize observational data for more informed causal reasoning. Our studies lays the groundwork for future advancements toward fully LLM-driven causal discovery.

Keywords

large language models, causal discovery, prompt engineering

1. Introduction

Understanding cause-and-effect relationships is fundamental to scientific discovery and decision-making across various fields such as biomedical research, economics, and social sciences. Traditionally, causal discovery relies on statistical methods applied to observational data, often requiring large datasets and strong assumptions about causal structures. Despite such limitations, statistical-based methods such as constraint-based approaches (e.g., PC algorithms [1]) and score-based methods (e.g., GES [2]), are still widely used in causal discovery.

Recent advances in Large Language Models (LLMs) have opened new possibilities for causal discovery. LLMs have been primarily used as expert in *knowledge-based causal discovery*, leveraging metadata—such as variable names and textual descriptions—to infer causal relationships [3, 4]. However, this approach is limited by the quality and specificity of metadata, and the internal knowledge of the LLMs themselves making it prone to inconsistencies and domain-specific biases. With LLMs advancing in reasoning [5, 6, 7], especially in text-based inference, a natural question emerges:

Can LLMs leverage observational data for causal discovery?

Workshop on Causal Neuro-symbolic Artificial Intelligence, June 01–5, 2025, Portoroz, Slovenia
susanti.yuni@fujitsu.com (Y. Susanti); michael.faerber@tu-dresden.de (M. Färber)

Despite the importance of observational data in statistical causal discovery, existing LLM-based methods have yet to fully utilize it. To address this gap, we propose a **data-driven causal discovery approach** that integrates observational data into LLM-based causal reasoning. We introduce prompting strategies incorporating observational data into the causal discovery process. By systematically embedding observational data into the prompts, we explore whether LLMs can enhance causal discovery beyond metadata-based inference, without relying solely on the LLMs' pre-existing domain knowledge or textual contexts. Our experiments across multiple benchmark datasets show that incorporating observational data improve LLMs' performance, up to 0.15 points in F1 scores, and outperform statistical-based methods. These results suggest that LLMs demonstrate potential in utilizing observational data for causal discovery, marking progress toward a hybrid model that integrates statistical methods with natural language reasoning via LLMs to better interpret data patterns for causal insights.

2. Related Work

LLMs have recently been utilized as expert systems for causal discovery, primarily by reasoning over *metadata* of the variables rather than directly analyzing observational data. This approach, known as *knowledge-based causal discovery* [3, 4], leverages LLMs' ability to interpret domain-specific metadata—such as variable names and textual descriptions–to infer causal relationships. A widely adopted method in this paradigm is *pairwise* prompting [8, 3], where an LLM is systematically queried about the causal relationship between each pair of variables. This iterative process constructs a causal graph using LLM-derived insights, demonstrating promising results despite not incorporating observational data. Recent studies [3, 9, 10, 11, 12] show that LLMs effectively provide background knowledge for causal discovery and outperform traditional non-LLM approaches. Other research has evaluated LLMs' ability to identify causal relationships in text [13, 14, 15]. For instance, recent work by [4] introduced a method that integrates knowledge graph structures into LLM prompts to enhance causal relation extraction by smaller models.

A different line of research integrates LLMs with traditional causal discovery methods [16, 17, 18]. These approaches typically use LLMs to extract prior knowledge or serve as feedback agents to refine causal graphs. Some studies further examine how observational data can be used to improve LLMs' causal reasoning, such as by incorporating statistics calculated from observational data like Pearson correlation into the prompt [19]. Unlike previous work, our work focuses on leveraging observational data *directly* for LLM-based causal discovery. Rather than using LLMs solely as knowledge extraction tools or supplementary components for traditional methods, we investigate their ability to infer causal relationships by reasoning directly over structured observational data. This approach aims to push the boundaries of LLMs for data-driven causal discovery, demonstrating their potential as standalone reasoning agents.

3. Approach

3.1. Task Formulation

Given a set of observed variables $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$, the objective is to infer a **causal graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents directed causal relationships between variables, and V_i represents a node in the causal graph. We formulate our task as a *classification* task

where each pair of variables (V_i, V_j) must be classified as (1) V_i *causes* V_j, (2) V_j *causes* V_i, or (3) *neither–no causal relationship*. The causal discovery process is then conducted using an LLM by formulating structured natural language prompts $\mathcal{P}$ to elicit causal dependencies.

3.2. Data-Driven Causal Discovery with LLMs

Our approach to causal discovery with LLMs extends beyond existing knowledge-based method by integrating observational data $\mathcal{D}$ into the prompt. However, since $\mathcal{D}$ is often too large to fit within the prompt, we apply sampling function S to extract a representative subset $\mathcal{D}_s$:

$$\mathcal{D}_s = S(\mathcal{D}, k) \tag{1}$$

where $S(\cdot)$ is a sampling strategy (e.g., random, systematic, or cluster sampling), and k is the sample size constrained by the prompt length. The prompt $\mathcal{P}$ then may encompass both prior knowledge $\mathcal{K}$–including known causal edges or constraints– and the sampled data $\mathcal{D}_s$. LLM's ability to infer causal relationships from such structured data distributions serves as the foundation of our data-driven causal discovery approach. In the following, we elaborate on the details of our proposed approach for systematically incorporating observational data into the prompt, utilizing (1) *pairwise* and (2) *BFS* prompting methods.

Pairwise Prompting with Observational Data. Pairwise Prompting is a localized approach where the LLM is queried about causal relationships between individual variable pairs. Given a variable pair (V_i, V_j), the LLM is instructed to determine whether a causal relationship exists between them, considering sampled observational data. The prompt $\mathcal{P}(V_i, V_j, \mathcal{K}, \mathcal{D}_s)$ in pairwise prompting explicitly asks:

- **Existence:** Does V_i cause V_j?
- **Directionality:** If a causal relationship exists, is it $V_i \rightarrow V_j$ or $V_j \rightarrow V_i$?

The LLM then predicts the causal relationship by selecting from three options: V_i *causes* V_j, V_j *causes* V_i, or *neither—no causal relationship*, as illustrated in Figure 1 (*left*).

BFS Prompting with Observational Data. The pairwise prompting requires a quadratic number of queries, making it impractical for large graphs. To address this, [19] introduce a framework using *breadth-first search* (BFS) strategy, reducing the number of queries to a linear scale. Instead of analyzing pairs, the LLM explores causal relationship by traversing the graph using BFS technique. In this work, we apply BFS prompting by [19], consisting three stages:

1. **Initialization** – The LLM identifies variables that are not causally influenced by others.
2. **Expansion** – The LLM determines which variables are caused by the current node.
3. **Insertion** – The proposed variables are added to the BFS queue, and suggested edges are checked for cycles before being inserted.

Figure 1 (*right*) illustrates a BFS approach with observational data. Unlike the pairwise approach, the LLM directly responds with variables instead of selecting from given options.

Pairwise Prompting

Task: You are a helpful assistant to a lung disease expert. Choose the correct statement regarding the causal relationship between *bronchitis* and *dyspnoea*. As an additional information, here is a sample of observational data between the pair:

bronchitis	*dyspnoea*
1	1
0	1
1	0
⋮	⋮

Options:

A. *bronchitis* causes *dyspnoea*.

B. *dyspnoea* causes *bronchitis*.

C. There is no causal relationship between *bronchitis* and *dyspnoea*.

Answer: _____

BFS Prompting

Task: You are a helpful assistant to experts in lung disease research. Our goal is to construct a causal graph between the following variables: {*asia, smoke, bronc, dysp*, ...}
(Initialization) You will start with identifying the variable(s) that are unaffected by any other variables.
Answer=[*smoke, asia*]
(Expansion) Given *smoke, asia* is(are) not affected by any other variables. Select variables that are caused by *smoke*. Additionally, a sample of observational data for *smoke* and other variables are as follows:

smoke	*bronc*	*dysp*	...
1	1	0	...
1	0	1	...
1	1	0	...
⋮	⋮	⋮	⋮

Answer: _____

Figure 1: Prompt examples for Pairwise and BFS prompting [19] using observational data.

4. Evaluation

4.1. Evaluation Settings

Dataset. We conduct experiments on datasets from `BNLearn` [20], a collection of Bayesian network datasets widely used for testing causal discovery algorithms, as follows:

1. **ASIA [21]**: A network with 8 variables (e.g., *dyspnoea*, *bronchitis*, and *if a patient has recently traveled to Asia*), for lung diseases diagnosis based on medical observations.
2. **CANCER**: A network that models the factors influencing cancer development. It contains fewer variables than ASIA, but with more intricate dependency structures.
3. **SURVEY**: A dataset on how public transport usage varies across social groups, based on survey responses, with variables such as *age*, *occupation*, and *preferred means of transportation*.

Despite their modest size, we specifically selected them because they offer valuable insights for evaluating causal discovery in straightforward, well-defined relationships.

Model Comparison. We compare LLM-based causal discovery against statistical causal discovery methods, including: **(1) PC Algorithm** [1] and **(2) GES** [2]. For LLM-based causal discovery, we compare the two prompting strategies, with variations that include and exclude

	ASIA			CANCER			SURVEY		
	F1↑	NHD↓	Ratio↓	F1↑	NHD↓	Ratio↓	F1↑	NHD↓	Ratio↓
Statistical-based Methods									
PC [1]	0.50	0.24	0.50	0.33	0.44	0.67	0.50	0.44	0.50
GES [2]	0.38	0.28	0.63	0.33	0.44	0.67	0.25	0.10	0.75
LLM-based Methods									
Pairwise Prompting	0.47	0.29	0.53	0.60	0.22	0.39	0.45	0.31	0.55
+Pearson corr.	0.64	0.13	0.36	0.67	0.16	0.33	0.20	0.32	0.80
+Observational Data	0.58	0.16	0.42	0.66	0.18	0.35	0.53	0.19	0.47
BFS Prompting [19]	0.85	0.04	0.15	0.66	0.12	0.33	0.50	0.22	0.50
+Pearson corr.	0.88	0.03	0.12	0.72	0.12	0.27	0.45	0.31	0.55
+Observational Data	0.90	0.03	0.10	0.77	0.10	0.23	0.54	0.20	0.45

Table 1: Performance comparison on benchmark datasets. The best scores are marked in pink. For LLM-based approaches, we queried the model four times and reported the **average** scores.

observational data as an additional input: **(3) Pairwise Prompting, (4) Pairwise Prompting + Observational Data, (5) BFS Prompting, (6) BFS Prompting + Observational Data**. Additionally, we incorporate *pearson correlation* calculated from the observational data, following [19]: **(7) Pairwise Prompting + Pearson corr., (8) BFS Prompting + Pearson corr.**.

Experimental Setup. For each dataset, we conduct experiments across varying sample sizes in $\{100, 500, 1000\}$ for statistical-based methods, while keeping LLM-based methods fixed at k=100 observational data due to token length limitations. The samples were selected using various sampling strategy S —*random*, *cluster*, *systematic*, and *adaptive (K-means)* sampling methods. However, since the results showed no significant differences, we reported the scores from *simple random* sampling. We used `GPT` model `gpt-4-0125-preview` checkpoint, query it four times varying sampling temperatures in $\{0, 0.5, 0.7, 1.0\}$ and report the average results. We adapted the implementation code from the original BFS prompting paper [19] for our experiment, which includes implementation of PC [1] and GES [2] from `causal-learn` package [22].

4.2. Results and Discussion

Table 1 summarizes our experiment results. Since we frame causal discovery as a classification task, we compute classification metrics e.g., Precision, Recall, F1 score and report *normalized hamming distance* (NHD) and *ratio*, following [3, 19]. We discuss the key findings as follows:

LLM-based methods outperform statistical-based methods in most cases. Across all datasets, LLM-based methods including both Pairwise and BFS-based prompting show significant improvements over PC and GES in terms of F1 score. The improvement is especially significant in BFS-based prompting with observational data, achieving a 0.44-point increase (0.33 vs. 0.77 on CANCER) compared to PC method. Similarly, on ASIA, it delivers a 0.40-point gain (0.50 vs.0.90). When comparing LLM-based methods to GES, we observe a consistent F1 score improvement ranging from 0.29 to 0.52 across all datasets, highlighting the effectiveness of our approach.

While PC and GES are well-established for causal discovery, their performance heavily depends on sample size. In our experiments, we set a fixed number of 100 samples for LLM-based methods across all datasets, whereas statistical-based methods (PC and GES) are evaluated with varying sample sizes $\{100, 500, 1000\}$. Our results show that LLM-based methods, particularly when enriched with observational data, achieve strong performance even with a limited number of samples. In contrast, statistical-based methods may require as many as 1000 samples to reach comparable performance, as observed in the SURVEY dataset, where the PC method matches LLM-based methods at 1000 sample size. This underscores the robustness of LLM-based causal discovery, making it particularly valuable in data-limited scenarios.

Observational data improves LLM-based methods' performance. Across both LLM prompting methods, incorporating observational data consistently improves F1 scores while reducing NHD and Ratio values, demonstrating its effectiveness in improving causal discovery. In Pairwise Prompting, adding observational data results in a F1 score increase of up to 0.11 points (0.47 to 0.58 on ASIA). Similarly, in BFS-based Prompting, adding observational data leads to the best overall performance across all datasets, with F1 scores improving by up to 0.11 points (0.66 to 0.77 on CANCER). These results suggest that, despite being primarily trained using text-based data, LLMs demonstrate a potential to effectively leverage observational numerical data as contextual grounding for causal discovery and reasoning.

Additionally, we assess the impact of incorporating Pearson correlation derived from the same observational data, following [19]. The results demonstrate a consistent improvement over methods without any observational data. However, we find that our method of directly adding observational data yields better overall performance on average (0.075 vs. 0.035), especially with a more advanced prompting technique such as BFS prompting. This suggests that by directly incorporating observational data, LLMs make more informed causal inferences and reduce their reliance on surface-level textual patterns, further bridging the gap between data-driven and knowledge-driven approaches in causal discovery.

BFS Prompting consistently outperforms Pairwise Prompting. BFS Prompting achieves the highest F1 scores and lowest Ratio values in all datasets (values marked in pink in Table 1), demonstrating its superior ability to leverage observational data for causal discovery. Beyond being more efficient than its pairwise counterpart, BFS-based prompting demonstrates its superiority (up to 0.32 point, $0.58 vs. 0.90$ on ASIA) by offering a more contextual and structured approach to causal discovery. This suggests that leveraging global context awareness—multi-variable interactions rather than variable pairs in isolation—enhances causal inference. However, this prompting method includes the entire query history, which can lead to excessive prompt length and may be infeasible due to the LLM's token limitations.

5. Conclusion

In this work, we investigated the potential of Large Language Models (LLMs) for data-driven causal discovery by integrating observational data into their reasoning process. Through experiments on causal benchmark datasets, we assessed the extent to which LLMs can infer causal relationships from structured, observational data. Our results suggest that LLMs demonstrate

potential in utilizing observational data for causal discovery, marking progress toward a hybrid model that integrates statistical methods with natural language reasoning with LLMs.

Despite these promising results, the effectiveness of LLMs is still dataset-dependent, and reasoning stability can vary. Future work should further explore this hybrid approaches of LLM-based reasoning with statistical causal discovery, as well as refining prompting strategies and sampling selection approach. Additionally, it would be valuable to investigate performance across multiple LLMs and extending on larger dataset. By continuously improving LLMs' ability to process structured data, we move toward a more comprehensive framework that unifies statistical causal discovery with the reasoning capabilities of LLMs.

References

[1] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, The MIT Press, 2001. URL: https://doi.org/10.7551/mitpress/1754.001.0001.

[2] D. M. Chickering, Optimal structure identification with greedy search., J. Mach. Learn. Res. 3 (2002) 507–554. URL: http://dblp.uni-trier.de/db/journals/jmlr/jmlr3.html#Chickering02a.

[3] E. Kıcıman, R. Ness, A. Sharma, C. Tan, Causal reasoning and large language models: Opening a new frontier for causality, 2023.

[4] Y. Susanti, M. Färber, Knowledge graph structure as prompt: Improving small language models capabilities for knowledge-based causal discovery, in: G. Demartini, K. Hose, M. Acosta, M. Palmonari, G. Cheng, H. Skaf-Molli, N. Ferranti, D. Hernández, A. Hogan (Eds.), The Semantic Web – ISWC 2024, Springer Nature Switzerland, Cham, 2025, pp. 87–106.

[5] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, 2022. URL: https://arxiv.org/abs/2206.07682. arXiv:2206.07682.

[6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[7] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 22199–22213. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

[8] Y. Susanti, N. Holsmoelle, Prompting or fine-tuning? exploring large language models for causal graph validation, 2025. URL: https://arxiv.org/abs/2406.16899. arXiv:2406.16899.

[9] R. Tu, C. Ma, C. Zhang, Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis, 2023. arXiv:2301.13819.

[10] M. Willig, M. Zečević, D. S. Dhami, K. Kersting, Can foundation models talk causality?, 2022. arXiv:2206.10591.

[11] C. Zhang, S. Bauer, P. Bennett, J. Gao, W. Gong, A. Hilmkil, J. Jennings, C. Ma, T. Minka,

N. Pawlowski, J. Vaughan, Understanding causality with large language models: Feasibility and opportunities, 2023. arXiv:2304.05524.

[12] J. Gao, X. Ding, B. Qin, T. Liu, Is ChatGPT a good causal reasoner? a comprehensive evaluation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 11111–11126. URL: https://aclanthology.org/2023.findings-emnlp.743. doi:10.18653/v1/2023.findings-emnlp.743.

[13] V. Khetan, M. I. Rizvi, J. Huber, P. Bartusiak, B. Sacaleanu, A. Fano, MIMICause: Representation and automatic extraction of causal relation types from clinical notes, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 764–773. URL: https://aclanthology.org/2022.findings-acl.63. doi:10.18653/v1/2022.findings-acl.63.

[14] Y. Susanti, K. Uchino, Causal-evidence graph for causal relation classification, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 714–722. URL: https://doi.org/10.1145/3605098.3635894. doi:10.1145/3605098.3635894.

[15] P. Chatwal, A. Agarwal, A. Mittal, Enhancing causal relationship detection using prompt engineering and large language models, in: C.-C. Chen, A. Moreno-Sandoval, J. Huang, Q. Xie, S. Ananiadou, H.-H. Chen (Eds.), Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal), Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 248–252. URL: https://aclanthology.org/2025.finnlp-1.26/.

[16] M. Takayama, T. Okuda, T. Pham, T. Ikenoue, S. Fukuma, S. Shimizu, A. Sannai, Integrating large language models in causal discovery: A statistical causal approach, 2024. URL: https://arxiv.org/abs/2402.01454. arXiv:2402.01454.

[17] T. Ban, L. Chen, X. Wang, H. Chen, From query tools to causal architects: Harnessing large language models for advanced causal discovery from data, 2023. URL: https://arxiv.org/abs/2306.16902. arXiv:2306.16902.

[18] A. Abdulaal, adamos hadjivasiliou, N. Montana-Brown, T. He, A. Ijishakin, I. Drobnjak, D. C. Castro, D. C. Alexander, Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning, 2024. URL: https://openreview.net/forum?id=pAoqRlTBtY.

[19] T. Jiralerspong, X. Chen, Y. More, V. Shah, Y. Bengio, Efficient causal graph discovery using large language models, 2024. URL: https://arxiv.org/abs/2402.01207. arXiv:2402.01207.

[20] M. Scutari, Learning bayesian networks with the bnlearn r package, Journal of Statistical Software 35 (2010) 1–22. URL: https://www.jstatsoft.org/v35/i03/. doi:10.18637/jss.v035.i03.

[21] S. L. Lauritzen, D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society: Series B (Methodological) 50 (2018) 157–194. doi:10.1111/j.2517-6161.1988.tb01721.x.

[22] Y. Zheng, B. Huang, W. Chen, J. Ramsey, M. Gong, R. Cai, S. Shimizu, P. Spirtes, K. Zhang, Causal-learn: Causal discovery in python, Journal of Machine Learning Research 25 (2024) 1–8.