Stress-Testing the Robustness of LLM-Based Causal Inference

Ankitkumar Patel*, Jigarkumar Patel†, Amit Kumar‡, Supreetha Sreeram§, and Venkatesh Kulkarni¶
*IntelliQuest Venture, USA, †University of Texas at Dallas, USA, ‡Texas A&M University–Corpus Christi, USA
§Pondicherry University, India, ¶IIT - Guwahati, India.
ankitkumar.patel179@gmail.com, jigarkumar.patel@utdallas.edu, akumar3@islander.tamucc.edu,
supreme172051@gmail.com, venkateshkulkarni2001@gmail.com

***Abstract*—Why do some of the world's most powerful Artificial Intelligence (AI) systems still struggle to reason about cause and effect? Causal discovery, the ability to discern the underlying drivers of a system, is fundamental to scientific inquiry and has traditionally relied on structured data and statistical methods. Recently, Large Language Models (LLMs) have been explored as an alternative paradigm for causal inference, leveraging their broad knowledge and reasoning capabilities. Yet, despite their linguistic fluency, LLMs often rely on surface-level pattern recognition rather than genuine causal logic. While it is known that LLMs stumble in this domain, we lack a precise map of where, why, and how significantly these failures occur. In this paper, we propose a systematic stress-test framework across six critical dimensions: model scale, causal graph complexity, prompting strategy, metadata quality, data quality, and sensitivity to uncertainty. Our findings reveal a significant "optimism bias" in LLMs; models consistently overestimate causal links, leading to low precision across the board. Furthermore, we demonstrate that traditional evaluation metrics can be deceptive: while reasoning-heavy models lead in performance, the highest scores often occur when uncertainty in inference is overlooked. Once uncertainty is integrated into the evaluation, performance drops, providing a more honest reflection of causal reliability. We also find that, while scaling model size improves outcomes, these gains are fragile; performance degrades sharply as the quality of metadata decreases. This work serves as a practical guide to the blind spots of LLMs, offering a clear-eyed assessment of when these tools can be integrated into a causal discovery pipeline and when they are likely to lead researchers astray.**

***Keywords*-Causal discovery; large language models (LLMs); uncertainty; evaluation metrics; sensitivity analysis.**

I. Introduction

Causal discovery, the task of inferring causal structure from observational data, has traditionally relied on statistical algorithms grounded in conditional independence testing and score optimization [1]. Recently, interest has grown in whether LLMs can serve as a new kind of causal reasoner by leveraging vast pretraining knowledge rather than numerical analysis.

However, early benchmarks such as CORR2CAUSE [2] revealed that LLMs (a) barely exceed random chance, (b) fail to generalize, and (c) collapse under simple surface-level perturbations, such as variable renaming. The authors argue that these models rely on lexical and positional cues rather than genuine causal reasoning. Building on this critique, Wu et al. [3] and the broader causal landscape analysis [4] showed that embedding LLM outputs as prior knowledge can undermine traditional causal discovery methods, with many reported gains attributable to engineered prompts rather than true inference. In contrast, Darvariu et al. [5] suggest that LLMs can serve as effective priors for causal graph discovery, while Ban et al. [6] demonstrate performance improvements when LLMs are used as heuristic guides within causal pipelines. Notably, the authors convinced that LLMs should play a role of a heuristic for search initialization rather than causal discovery.

Despite these insights, these previous works [2][3] share four significant methodological limitations: (a) evaluations are confined to small-scale networks, (b) metadata is treated as a binary condition (present or absent), (c) the influence of varying degrees of description richness on causal inference is not systematically characterized, and (d) the impact of various prompting strategies on causal discovery is not analyzed.

This paper addresses these gaps by proposing a systematic stress test framework on six critical dimensions; (1) model scale, (2) causal graph complexity, (3) prompting strategy, (4) metadata quality, (5) data quality, and (6) sensitivity to uncertainty. We also present a novel uncertainty-aware metric to evaluate causal inference. Our sensitive analysis across a subset of dimensions reveals that the failure modes identified by [2][3] do not scale linearly. Instead, specific failure patterns are qualitatively amplified at scale, and strategies effective for small graphs can become counterproductive as complexity grows. Notably, richer metadata improves performance on small networks. These findings provide a comprehensive map of the boundary conditions governing LLM reliability.

The rest of the paper is structured as follows. In Section II, we present the formal problem statement of causal discovery. The proposed framework for multi-dimensional sensitivity is outlined in Section III, and the proposed uncertainty-aware evaluation metric is defined in Section IV. Section V covers the performance analysis of state-of-art LLMs under the proposed framework. Finally, our findings and future work are summarized in Section VI.

II. Problem Statement

The primary goal of causal discovery is to recover the latent Directed Acyclic Graph (DAG) $G = (V, E)$ from observed variables $V = \{v_1, \ldots, v_n\}$. Accurately identifying these edges is essential for scientific explanation and intervention. While LLMs are increasingly used as heuristic priors in this domain, their reliability depends on interacting boundary conditions that remain poorly understood. We formalize this LLM-based inference as a function: $\hat{E} = \mathcal{F}(M, A, P, Q, D)$, where M denotes a model, A denotes a pair attributes, P denotes a prompt configuration, Q denotes metadata quality, and D denotes data. The fundamental challenge is to treat

$\mathcal{F}$ as a stable reasoner, ignoring its extreme sensitivity to perturbations in M, P, Q and D.

III. Multi-Dimensional Sensitivity Analysis

To provide a comprehensive map of LLM reliability, causal discovery can be analyzed across the following six domains.

A. Graph Scale and Topology Complexity

A causal graph's complexity is defined by its node count and edge density. As variables increase, the space of potential causal hypotheses grows combinatorially, creating two compounding challenges for LLMs. First, larger graphs result in crowded contexts, forcing the model to process overwhelming lists of relationships and descriptions. Second, finite context-length constraints risk implicit pruning, where the model deprioritized or truncates critical information. These factors limit the model toward surface heuristics rather than genuine reasoning. Consequently, LLM performance often exhibits sharp failures upon reaching capacity limits rather than degrading smoothly with topological complexity.

B. Model Scale and Pretraining Objectives

LLM effectiveness in causal discovery depends on non-linear interactions between model scale, training data, and pretraining objectives. This complexity makes it difficult to attribute performance gains solely to model size. Empirically, models optimized for reasoning subject to outperform standard LLMs on structured causal assessments. While next-token prediction models often rely on superficial lexical or frequency-based cues, true causal discovery requires specific inductive biases. Consequently, increasing model capacity only improves causal reasoning when paired with objectives that promote abstraction, counterfactual thinking, and logical consistency.

C. Prompting Strategy and Heuristic Bias

Prompting strategy centralizes LLM causal behavior, with methods like zero-shot, few-shot, and Chain-of-Thought (CoT) significantly altering performance. While CoT can elicit structured reasoning and uncover implicit assumptions through step-by-step explanation [7], its benefits are scale-dependent. As graph complexity increases, CoT prompts induce heavy context overhead, often leading to verbose but shallow reasoning. This can force models to prioritize narrative coherence over causal validity, amplifying heuristic biases. Consequently, prompting strategies that appear effective for small-scale problems may become counterproductive in large-scale problems.

D. Metadata Quality and Description Richness

Metadata refers to the semantic labels and textual descriptions of variables. While richer metadata can reduce ambiguity and provide the contextual cues necessary for causal reasoning, but on the other hand, it introduces contextual complexity which can overwhelm the reasoning capacity of LLMs. Consequently, the impact of metadata on causal accuracy depends on a non-linear trade-off between description precision, prompt structure, and the graph scale.

E. Observational Data Precision

The influence of observational data on LLM-based causal reasoning remains an open question, as interpretations are shaped by how data is represented and contextualized within a prompt rather than by numerical content alone. In practice, the granularity of observations is critical. While coarsely aggregated data may obscure causal signals, overly fine-grained representations can introduce noise and inflate prompt complexity. Consequently, LLM decision-making varies significantly based on data precision and formatting. This necessitates a careful balance between informational fidelity and contextual clarity in causal discovery tasks.

F. Impact of Evaluation Metrics

Conventional metrics like precision, recall, and F-score assume a binary framework, conflating uncertainty with error. This obscures the critical distinction between an incorrect assertion and a principled abstention, a limitation especially pronounced in LLMs, which lack classically calibrated probabilities. In our setting, we query LLMs with dual prompts for the existence ($P(e)$) and non-existence ($P(\neg e)$) of causal edges. Because LLMs rely on heuristic reasoning rather than coherent probabilistic inference, they frequently violate basic constraints (e.g., $P(e) + P(\neg e) = 1$). We also observe borderline cases where nearly equivalent probabilities reflect epistemic uncertainty rather than a meaningful causal preference. Treating these as hard decisions artificially inflates false positives and negatives. To better assess the trade-off between decisiveness and reliability, we propose a novel ternary decision framework ($True$, $False$, and $Uncertain$). By categorizing inconsistent or marginal predictions as "Uncertain," we can penalize confident errors more severely than informed abstentions. Our proposed metric is formally defined in the following section.

IV. Uncertainty-Aware Evaluation Theory

LLM-assisted causal discovery operates under inherent epistemic uncertainty, which is not adequately captured by traditional binary evaluation metrics such as precision, recall, and F-score. Unlike classical statistical methods that enforce hard decisions, LLMs naturally admit a third outcome, abstention, when confidence is insufficient. To account for this behavior, we adopt a ternary decision framework in which each ordered variable pair has a ground-truth state of *Edge* or *No-Edge*, while the model may predict *Edge*, *No-Edge*, or *Uncertain*. This yields a ternary confusion structure that distinguishes uncertainty over true edges (UE) from uncertainty over non-edges (UN), reflecting their differing implications for causal reasoning and intervention planning.

We incorporate this structure into uncertainty-adjusted precision and recall, defined as

$$R_{adj} = \frac{TP}{TP + FN + \beta \cdot UE}, \quad P_{adj} = \frac{TP}{TP + FP + \alpha \cdot UN},$$

where $\alpha, \beta \in [0, 1]$ control the penalty assigned to abstention. These parameters encode epistemic preferences between decisiveness and caution. The resulting F-Adj metric, defined as

TABLE I. REPRESENTATIVE WEIGHTING STRATEGIES AND THEIR EPISTEMIC INTERPRETATIONS.

(α, β)	Interpretation
$(0.5, 0.5)$	Balanced and symmetric treatment of uncertainty.
$(1.0, 1.0)$	Uncertainty treated as full failure; decisiveness is enforced.
$(0.8, 0.4)$	Abstention is preferred over guessing a causal edge.
$(0.4, 0.8)$	Guessing a causal edge is preferred over abstention.
$(0.2, 0.8)$	Aggressive discovery setting that prioritizes recall over caution.
$(0.8, 0.2)$	Conservative discovery setting that prioritizes precision and abstention.

the harmonic mean of P_{adj} and R_{adj}, generalizes the classical F-score by explicitly modeling uncertainty.

The weighting parameters (α, β) in the uncertainty-aware evaluation metric encode explicit epistemic policies that regulate the trade-off between decisiveness and caution in causal assessment. Rather than treating uncertainty as uniformly undesirable, these parameters allow evaluators to specify when a model should abstain versus when it should commit to a causal claim. Symmetric settings ($\alpha = \beta$) reflect balanced treatment of uncertainty, while asymmetric configurations express task-dependent priorities. For instance, $\alpha > \beta$ favors edge assertion by penalizing uncertainty over non-edges, whereas $\alpha < \beta$ prioritizes caution by discouraging uncertain edge claims. Such distinctions are critical in causal discovery, where the relative costs of false positives and false negatives vary across domains. By making these epistemic assumptions explicit, the framework aligns evaluation with downstream objectives and supports principled transitions between exploratory and confirmatory stages of a causal discovery pipeline. Table I summarizes representative weighting strategies and their corresponding epistemic goals.

V. PERFORMANCE EVALUATION

All experiments were conducted on the Asia network, a well-established benchmark from the Bayesian network literature [8]. The dataset models a small medical diagnostic system with eight binary variables related to pulmonary diseases and clinical symptoms. Despite its limited scale, the Asia network captures key causal motifs such as confounding and mediation, making it a canonical testbed for controlled causal discovery. The ground-truth structure is provided as a directed acyclic graph (DAG) encoding medically grounded causal relationships, including the effects of travel and smoking on disease outcomes and downstream symptoms, enabling precise evaluation of structural correctness and error types.

LLMs are evaluated under a zero-shot, instruction-based prompting setting. Specifically, we consider `GPT-4.1-mini` [9], `GPT-4.1` [10], `GPT-4o` [11], `DeepSeek-R1` [12], `GPT-5.2` [13], and `gpt-oss-20b` [14]. All models are accessed through a unified API-based evaluation pipeline and receive an identical task description along with the same rule-based instructions for identifying adjacency, spuriousness, and independence between attributes and metadata. This setup minimizes confounding effects from advanced prompt engineering, such as exemplar bias and context overload, and allows performance differences to be attributed more directly to model behavior and metadata quality. Full prompt templates and implementation details are provided in [15]; exploration of more advanced prompting strategies is left for future work.

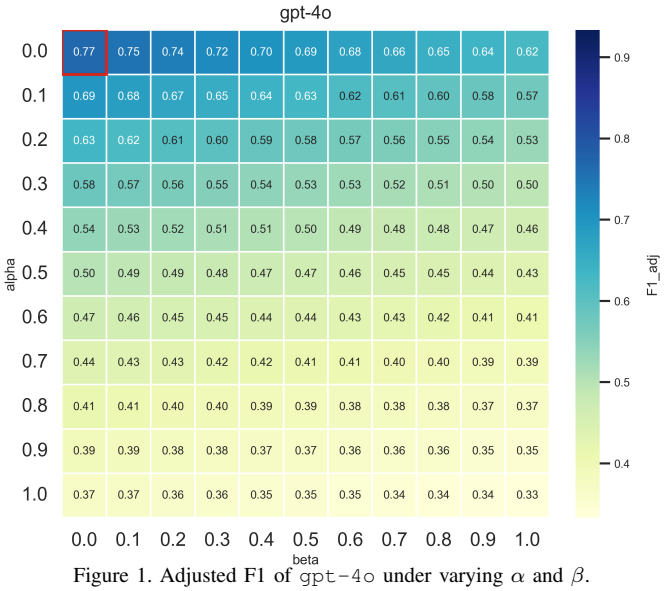

Figure 1. Adjusted F1 of `gpt-4o` under varying α and β.

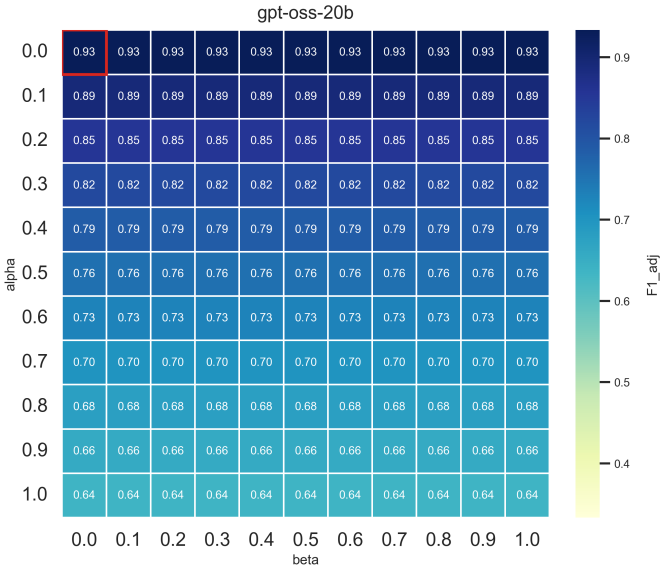

Figure 2. Adjusted F1 of `gpt-oss-20b` under varying α and β.

To examine the impacts of metadata systematically, we provided metadata with different level of richness, simulating increasing levels of noise and ambiguity. For Asia, metadata is provided at three informativeness levels (L1–L3). At L1, only variable names are provided, introducing maximum noise and forcing reliance on lexical cues. L2 augments names with variable roles and brief descriptions, reducing ambiguity without revealing explicit structure. L3 further explicit causal hints, latent variables, and confounding structure, making portions of the underlying graph approximately recoverable.

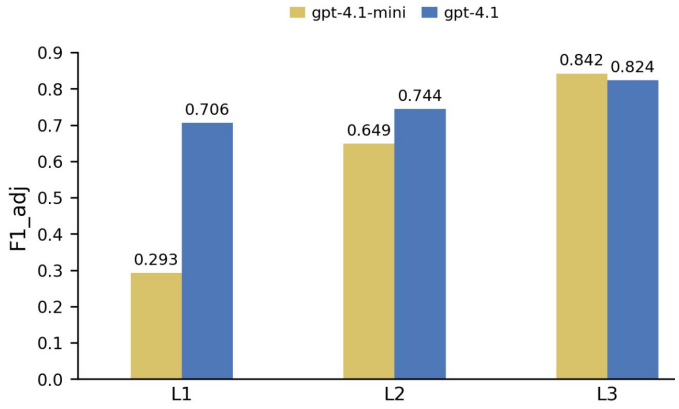

Figure 3. Adjusted F1 as a function of metadata quality on the Asia dataset.

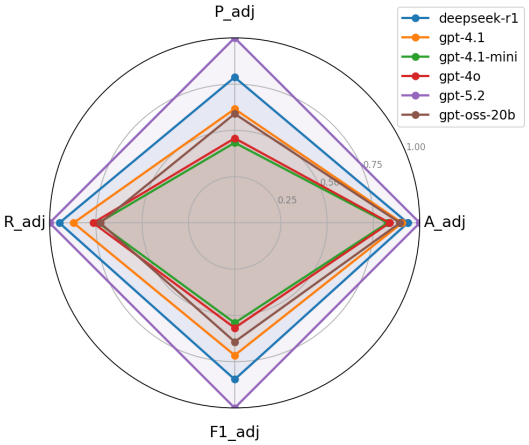

Figure 4. Comparison of evaluation metrics across models.

This controlled degradation framework enables the isolation of metadata-driven failure modes and distinguishes errors due to limited semantic grounding.

For each metadata variant, each model was queried twice using a simple vanilla prompting strategy: once with an *edge* prompt to estimate the likelihood of a directed causal relation between a variable pair, and once with a *no-edge* prompt to estimate the likelihood of no direct causal relation. In this setting, the LLM is provided only with the task description and metadata. This design choice minimizes confounding effects introduced by advanced prompt engineering, such as exemplar bias or context overload, and enables clearer attribution of performance differences to intrinsic model behavior and metadata quality. We used a fixed temperature of 0.0 and decision threshold of 0.7. For uncertainty-aware evaluation, the penalty parameters α and β were varied over $\{0.0, 0.1, \ldots, 1.0\}$, and predictions were aggregated into adjusted metrics, including `F1_adj`.

The heatmaps in Fig. 1 and Fig. 2 show how model performance ($F1_{adj}$) varies as a function of the uncertainty penalty parameters, α and β. In all models, the highest scores (max) are consistently observed in $(\alpha = 0.0, \beta = 0.0)$, where uncertainty is effectively ignored. As α and β increase, forcing the models to be penalized for ambiguous causal claims, the $F1_{adj}$ scores drop significantly. Among the models, GPT-5.2 and DeepSeek-R1 demonstrate near-perfect performance (1.00) maintaining high stability even as the penalty for uncertainty increases. GPT-4o and GPT-OSS-20B show high sensitivity to α and β. The score of GPT-4o is degraded by approximately 0.29 as decisiveness is enforced. GPT-OSS-20B exhibits the most dramatic drop, with its performance falling by 0.44 at maximum penalty. This confirms that models often appear more capable than they are by making "guesses" that a binary metric would treat as a hard decision.

In Fig. 3, the performance of gpt-4.1-mini and gpt-4.1 on the Asia dataset was evaluated using the $F1_{adj}$ metric ($\alpha = 0.5, \beta = 0.5$), which balances incorrect assertions against principled abstentions. The results reveal a non-linear relationship between metadata richness and accuracy. While both models improve as metadata moves from L1 to L3, gpt-4.1-mini shows the most dramatic gain, jumping from 0.293 to 0.842, corresponding to an improvement of approximately 187.4%. At low richness (L1), the larger gpt-4.1 significantly leads (0.706), indicating stronger internal reasoning when context is sparse. However, as metadata richness increases, the performance gap between the smaller and larger models narrows substantially. At L3, the models converge, and the smaller variant slightly leads (0.842 vs 0.824). In contrast, gpt-4.1 shows a more modest improvement of approximately 16.7% from L1 to L3, suggesting that richer metadata reduces the performance difference between smaller and larger models.

The radar chart in Fig. 4 provides a holistic view of model performance across four adjusted metrics: P_{adj}, R_{adj}, A_{adj}, and $F1_{adj}$ when $\alpha = 0.5$ and $\beta = 0.5$. Most models exhibit an "elongated" shape toward R_{adj}, indicating higher recall than precision. This provides empirical evidence for the "optimism bias", LLMs are inclined to over-identify causal links, leading to a high rate of false positives. GPT-5.2 occupies the outermost perimeter of the chart, showing nearly perfect scores (1.00) across all four metrics. DeepSeek-R1 follows closely, though it shows a slightly higher tendency toward recall over precision compared to the top frontier. Smaller models like GPT-4.1-mini and GPT-4o occupy a much smaller area of the radar, with significant dips in P_{adj}. This reflects, their tendency to rely on surface-level patterns and metadata hints rather than robust causal logic.

To improve readability and ground the quantitative results in concrete behavior, we examine representative model outputs. This qualitative inspection reveals clear and intuitive failure patterns that help explain the performance differences in Fig. 4. For instance, in the Asia network, several mid- and low-capacity models incorrectly infer a direct causal link from *smoking* to *dyspnoea*, ignoring the mediating disease variables despite metadata that explicitly describes the indirect relationship. In other cases, models assert bidirectional dependencies between clinically unrelated variables, often driven by superficial co-occurrence cues in the metadata. These errors become more frequent as metadata quality degrades or context

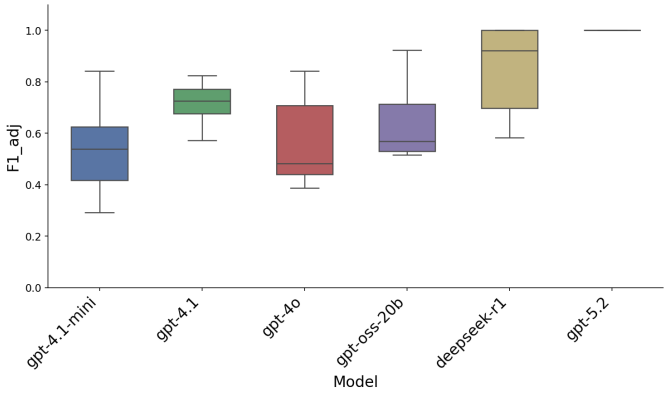

Figure 5. Distribution of adjusted F1 scores across models.

becomes crowded. In contrast, GPT-5.2 more often abstains in such ambiguous situations, which is reflected in its higher and more stable $F1_{adj}$ scores.

Figure 5 captures the variance and median performance of $F1_{adj}$ across different models in a box plot when the input context randomly varies between $L1$ to $L3$ levels. GPT-5.2, with its higher capacity and enhanced reasoning abilities, achieves substantially higher median $F1_{adj}$ scores and tighter performance distributions than smaller or non-reasoning models, whereas the remaining models exhibit wide inter-quartile ranges and multiple outliers, indicating inconsistent application of causal reasoning across varying metadata qualities.

GPT-5.2 delivers the highest and most consistent $F1_{adj}$ scores, however, it also comes at a substantially higher deployment cost which makes it impractical for all use cases. The box-plot shows that several smaller or cheaper models, such as GPT-4.1 and DeepSeek-R1, achieve reasonably stable performance under favorable conditions, suggesting they can be effective when graph size and metadata quality are controlled. The uncertainty-aware $F1_{adj}$ metric helps make these tradeoffs explicit, revealing when lower-cost models are reliable enough and when their variability poses unacceptable risk. This enables practitioners to reserve frontier models for high-stakes decisions while using less expensive models for exploratory analysis or resource-constrained settings, rather than defaulting to maximum scale by assumption.

While high-capacity models typically offer superior reliability and performance, they are not always the optimal choice when operating under stringent cost, latency, and scalability constraints. By explicitly accounting for the trade-offs between these operational metrics and raw model performance, practitioners can make informed decisions when adopting the appropriate models for specific applications. Furthermore, this balanced approach enables more efficient resource provisioning and reservation within generative AI cloud environments.

VI. Conclusion and Future Work

This paper presents a systematic stress test of LLMs for causal discovery, revealing a consistent optimism bias in which models overestimate causal relationships. To overcome the limitations of binary evaluation, we introduce the uncertainty-aware $F1_{adj}$ metric, which shows that many apparent performance gains disappear once epistemic uncertainty is accounted for. While reasoning-optimized models achieve stronger performance, smaller models like GPT-4.1-mini, narrowing the gap with larger models as metadata richness increases. Collectively, these findings provide a more realistic and reliable assessment of LLM capabilities in causal reasoning tasks. Building on these boundary conditions, future work will evaluate how LLM judgments shift when provided with raw observational data alongside variable descriptions. We plan to move beyond zero-shot prompting to study how advanced strategies, such as CoT and multi-agent debate, interact with graph scale. Finally, we will expand evaluations to complex systems exceeding 200 nodes to transform LLMs into robust components of high-stakes causal discovery pipelines.

References

[1] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in genetics*, vol. 10, p. 524, 2019.

[2] Z. Jin et al., "Can large language models infer causation from correlation?" In *The Twelfth International Conference on Learning Representations*, OpenReview.net, 2024. [Online]. Available: https://arxiv.org/pdf/2306.05836.

[3] X. Wu, K. Yu, J. Wu, and K. C. Tan, "LLM cannot discover causality, and should be restricted to non-decisional support in causal discovery," 2025, arXiv: 2506.00844 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2506.00844.

[4] Z. Zhang, R. Guo, Z. Wen, and Z. Zhang, "Large language models for causal discovery: Current landscape and future directions," 2024, arXiv: 2402.11068 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2402.11068.

[5] V.-A. Darvariu, S. Hailes, and M. Musolesi, "Large language models are effective priors for causal graph discovery," 2024, arXiv: 2405.13551 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2405.13551.

[6] T. Ban et al., "Integrating large language model for improved causal discovery," 2025, arXiv: 2306.16902 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2306.16902.

[7] J. Wei et al., "Chain of thought prompting elicits reasoning in large language models," 2022, arXiv: 2201.11903 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2201.11903.

[8] S. L. Lauritzen and D. J. Spiegelhalter, "Local computation with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society: Series B*, vol. 50, no. 2, pp. 157–224, 1988.

[9] OpenAI, *Gpt-4.1 mini model*, OpenAI API documentation. Accessed: 2026-04-12, 2025. [Online]. Available: https://developers.openai.com/api/docs/models/gpt-4.1-mini.

[10] OpenAI, *Gpt-4.1 model*, OpenAI API documentation. Accessed: 2026-04-12, 2025. [Online]. Available: https://developers.openai.com/api/docs/models/gpt-4.1.

[11] OpenAI, *Gpt-4o model*, OpenAI API documentation. Accessed: 2026-04-12, 2024. [Online]. Available: https://developers.openai.com/api/docs/models/gpt-4o.

[12] DeepSeek-AI et al., "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025. [Online]. Available: https://arxiv.org/abs/2501.12948.

[13] OpenAI, *Gpt-5.2 model*, OpenAI API documentation. Accessed: 2026-04-12, 2025. [Online]. Available: https://developers.openai.com/api/docs/models/gpt-5.2.

[14] OpenAI, *Gpt-oss-120b & gpt-oss-20b model card*, Official model card. Accessed: 2026-04-12, 2025. [Online]. Available: https://openai.com/index/gpt-oss-model-card/.

[15] A. Kumar, *Base prompt data for llm causal discovery*, https://github.com/aamitssharma07/uncertainty-aware-llm-causal-discovery-repro.git, GitHub repository, accessed: 2026-04-03, 2026.