

ICE CREAM DOESN'T CAUSE DROWNING: BENCHMARKING LLMs AGAINST STATISTICAL PITFALLS IN CAUSAL INFERENCE

Jin Du¹, Li Chen¹, Xun Xian¹, An Luo¹, Fangqiao Tian¹, Ganghua Wang²,
Charles Doss¹, Xiaotong Shen¹, Jie Ding¹

¹School of Statistics, University of Minnesota, Minneapolis, MN 55455

²Data Science Institute, University of Chicago, Chicago, IL 60637

{du000142, chen7019, xian0044, luo00318, tian0257}@umn.edu

{cdoss, xshen, dingj}@umn.edu, ganghua@uchicago.edu

ABSTRACT

Reliable causal inference is essential for making decisions in high-stakes areas like medicine, economics, and public policy. However, it remains unclear whether large language models (LLMs) can handle rigorous and trustworthy *statistical causal inference*. Current benchmarks usually involve simplified tasks. For example, these tasks might only ask LLMs to identify semantic causal relationships or draw conclusions directly from raw data. As a result, models may overlook important statistical pitfalls, such as Simpson’s paradox or selection bias. This oversight limits the applicability of LLMs in the real world. To address these limitations, we propose **CausalPitfalls**, a comprehensive benchmark designed to rigorously evaluate the capability of LLMs in overcoming common causal inference pitfalls. Our benchmark features structured challenges across multiple difficulty levels, each paired with grading rubrics. This approach allows us to quantitatively measure both causal reasoning capabilities and the reliability of LLMs’ responses. We evaluate models using two protocols: (1) direct prompting, which assesses intrinsic causal reasoning, and (2) code-assisted prompting, where models generate executable code for statistical analysis. Additionally, we validate the effectiveness of this judge by comparing its scoring with assessments from human experts. Our results reveal significant limitations in current LLMs when performing statistical causal inference. The CausalPitfalls benchmark provides essential guidance and quantitative metrics to advance the development of trustworthy causal reasoning systems. Our code is publicly available at [CausalPitfalls](#).

1 INTRODUCTION

Causal inference (Pearl, 2009; Imbens & Rubin, 2015) is fundamental to decision-making across diverse fields. For instance, accurately determining the effectiveness and safety of a vaccine is pivotal in public health decisions (Voysey et al., 2021). However, identifying causal relationships with both reliability and interpretability remains challenging. In practice, individuals without formal statistical training frequently fall into subtle pitfalls, leading to plausible yet incorrect conclusions. A classic illustration is the erroneous conclusion that ice cream sales cause drowning incidents — overlooking the hidden confounder of hot weather causing both events (Pearl, 2009; Greenland & Robins, 1986; Rosenbaum, 1987).

Given these complexities, automated tools like large language models (LLMs) present promising avenues, demonstrated by their effectiveness in scientific problem-solving (Lewkowycz et al., 2022; Achiam et al., 2023; Lin et al., 2025a;b) and clinical reasoning (Singhal et al., 2023; Yu et al., 2025). Recent studies (Wang, 2024; Dhawan et al., 2024; Liu et al., 2024) have evaluated LLMs’ abilities to evaluate accuracy in causal-effect estimation, but these benchmarks often neglect crucial aspects like robustness, interpretability, and susceptibility to common causal pitfalls. As a result, LLMs can produce seemingly convincing yet misleading causal conclusions.

To illustrate why reliability assessment is crucial, we highlight two representative failure modes (detailed in Section 3). First, LLMs can ignore strong data evidence, but in favor of superficial semantic cues: in a synthetic health scenario with identical datasets, LLMs concluded the drink was beneficial when labeled “HealthPlus” and harmful when labeled “UltraSugar,” even when the data indicated the **opposite**. Second, LLMs can mistake random variation for genuine causal structure: when tested on research funding data from the Netherlands, LLMs attributed differences in success rates to gender bias or Simpson’s paradox, despite statistical analyses showing that neither claim is supported. These cases demonstrate that LLMs may produce confident causal claims directly contradicted by the data, showing the need for benchmarks that assess causal inference *reliability*.

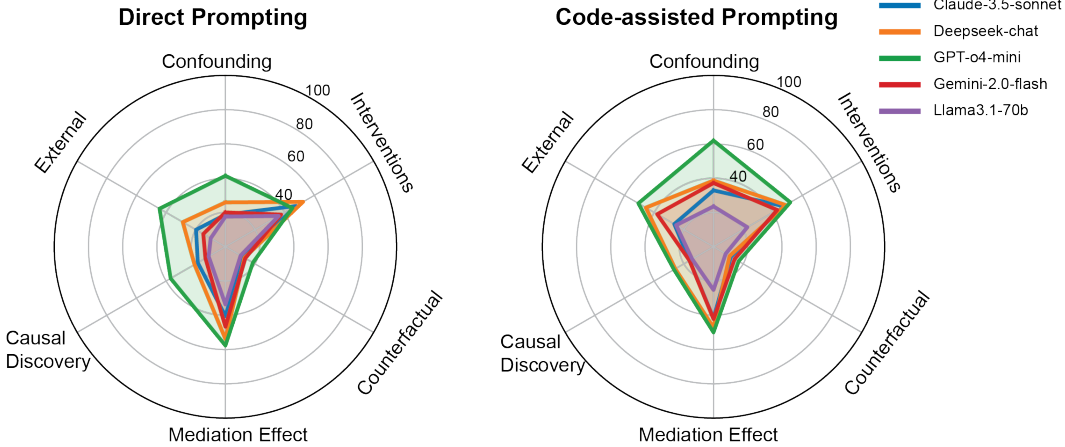


Figure 1: **Overall Message:** Our results reveal a clear **reliability gap** in causal inference when LLMs rely only on direct prompting, with all models struggling most on mediation and external validity questions. Introducing code-assisted prompting leads to substantial gains across every task and brings all models closer together in performance. This shows that executable analysis is essential for large language models to handle complex statistical challenges and deliver trustworthy causal conclusions. Full results for all evaluated LLMs are provided in Table 4.

1.1 MAIN CONTRIBUTIONS

First, we introduce **CausalPitfalls**, a novel comprehensive benchmark specifically designed to evaluate the reliability of large language models (LLMs) in statistical causal inference. Unlike existing benchmarks primarily focused on accuracy, our benchmark targets model susceptibility to common causal pitfalls as shown in Figure 1, including (1) Confounding Biases and Spurious Associations, (2) Interventions and Experimental Reasoning, (3) Counterfactual Reasoning and Hypotheticals, (4) Mediation and Indirect Causal Effects, (5) Causal Discovery and Structure Learning, and (6) Causal Generalization and External Validity (Pearl et al., 2003; Peters et al., 2017). These categories are structured into 15 distinct challenges, encompassing a total of 75 evaluation questions and 75 carefully constructed datasets that systematically test the robustness of LLM causal reasoning capabilities.

Second, we comprehensively evaluate the reliability of ten LLMs under two distinct evaluation protocols: (1) *direct prompting*, assessing intrinsic causal reasoning from raw data, and (2) *code-assisted prompting*, where models generate executable code to perform statistical analyses before responding. This dual-protocol approach provides a detailed quantitative assessment, highlighting areas where computational assistance significantly improves causal reasoning and where intuitive reasoning could suffice.

Third, we introduce a quantitative metric termed *causal reliability*, calculated as the average normalized score across all benchmark challenges, enabling standardized comparisons of LLM reliability in causal reasoning tasks. By systematically quantifying reliability, this metric provides a crucial framework for future research aimed at developing more robust and trustworthy causal inference capabilities in AI systems.

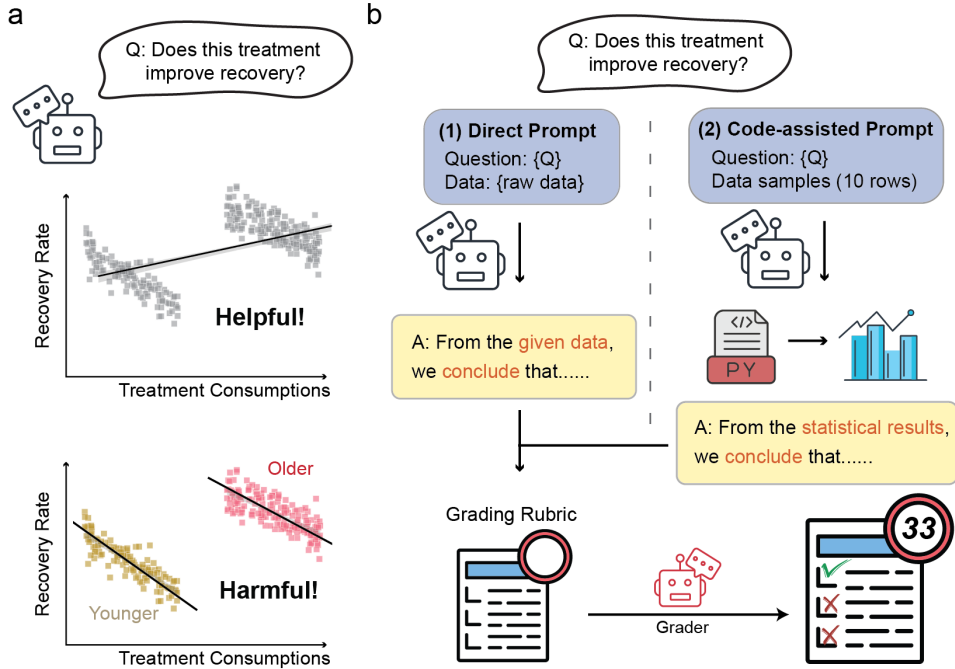


Figure 2: High-level overview of the CausalPitfalls benchmark. (a) An illustrative real-world pitfall (Simpson’s paradox): when data on treatment consumption and recovery are pooled (top), a naïve analysis finds a positive effect (“Helpful!”), but stratifying by age reveals a negative effect within both younger and older subgroups (“Harmful!”). (b) Benchmark workflow: LLMs are evaluated under two protocols: (1) Direct Prompting on raw data, assessing intrinsic causal reasoning, and (2) Code-Assisted Prompting on sampled data, assessing computationally grounded inference. In both cases, model answers are automatically scored against a hidden grading rubric by an independent grader to quantify each model’s causal reliability.

1.2 RELATED WORK

1. Causal Inference and Statistical Pitfalls. Causal inference from observational data is inherently challenging because counterfactuals are unobservable and confounding is ubiquitous (Pearl, 2009; Imbens & Rubin, 2015). Causal inference methods for addressing confounders, whether the confounders are measured (Chan et al., 2016; Lin et al., 2023; Doss et al., 2024) or latent (Kang et al., 2016; Guo et al., 2022), depend on restrictive model assumptions that are often difficult to verify empirically. Inferring causal direction similarly hinges on stringent structural equation assumptions (Peters & Bühlmann, 2014; Li et al., 2024) or auxiliary information such as valid instruments (Chen et al., 2024). Mediation analysis (MacKinnon, 2012; Yang et al., 2021), which targets specific causal pathways, demands careful adjustment for intermediate variables to avoid post-treatment bias. Finally, transporting causal conclusions across different domains requires justification of source-target invariances and methodologies for causal knowledge transfer (Wei et al., 2023; Chen et al., 2025). Rigorously confronting each of these challenges is essential to conduct a reliable causal analysis.

2. LLMs for Causal Reasoning. Recent studies have extensively investigated the causal reasoning capabilities of LLMs (Willig et al., 2022; Zečević et al., 2023; Qin et al., 2019; Luo et al., 2025; Du et al., 2024). For example, Kiciman et al. (2023) demonstrated that LLMs can infer causal relationships only from variable names, outperforming traditional statistical approaches (Peters et al., 2017). However, these evaluations focus on scenarios involving commonsense causality. To bridge this gap, Jin et al. (2023) introduced synthetic datasets generated from causal graphs, thereby enabling the assessment of LLMs’ causal reasoning performance in contexts extending beyond commonsense knowledge. Additionally, recent works (Long et al., 2023; Zhou et al., 2024) have assessed LLMs’ capabilities in data-driven causal inference tasks, focusing on accuracy in estimating causal effects and recovering DAG structures from observational data.

2 BENCHMARK CURATION

2.1 PITFALL CATEGORIES AND CHALLENGES

To evaluate the reliability of causal inference performed by LLMs, we introduce the benchmark **CausalPitfalls** to assess model performance across common statistical pitfalls. Specifically, our benchmark addresses six major categories of causal inference pitfalls, consisting of 15 distinct challenges. Each challenge includes five questions across difficulty levels ranging from “very easy” to “very hard.” Table 1 summarizes the categories and their respective challenges:

Table 1: CausalPitfalls benchmark categories and challenges

Confounding biases and spurious associations	Interventions and experimental reasoning
Simpson’s paradox	Observational vs experimental reasoning
Selection bias (Berkson’s paradox)	Causal effect estimation
Counterfactual reasoning and hypotheticals	Mediation and indirect causal effects
Counterfactual outcome prediction	Mediator-outcome confounding
Causal necessity and sufficiency	Sequential mediators
	Treatment-mediator interaction effects
Causal discovery and structure learning	Causal generalization and external validity
Cause-effect direction inference	Population shift and transferability
Handling uncertainty in causal structures	Temporal stability of causal effects
	Contextual interaction and moderation effects
	Domain shift and transportability of causal knowledge

Below is a brief overview of the six major categories in the **CausalPitfalls** benchmark:

- **Confounding biases and spurious associations:** Covers scenarios where misleading correlations arise due to hidden variables or biased conditioning.
- **Interventions and experimental reasoning:** Focuses on distinguishing correlation from causation through randomized experiments or interventional data, and quantifying treatment effects.
- **Counterfactual reasoning and hypotheticals:** Evaluates LLMs’ ability to answer “what if” questions by reasoning about alternate outcomes under hypothetical changes.
- **Mediation and indirect causal effects:** Assesses whether models can identify and reason through intermediate causal pathways, including complex mediation structures.
- **Causal discovery and structure learning:** Tests the ability to infer causal directionality or relationships from data without pre-defined causal graphs.
- **Causal generalization and external validity:** Examines whether causal conclusions generalize across different contexts, populations, and environments.

Many of these scenarios can also be framed as purely statistical association problems. However, throughout this benchmark, we adopt a causal interpretation grounded in either the Neyman–Rubin potential outcomes framework (Splawa-Neyman et al., 1990; Rubin, 1974) or Pearl’s do-calculus (Pearl, 2009). This distinction ensures that our tasks target causal reasoning, rather than surface-level correlations.

One illustrative example of these pitfalls is Simpson’s paradox (Simpson, 1951) (Fig. 2a), a commonly misunderstood statistical phenomenon. Simpson’s paradox occurs when overall data seem to suggest one conclusion: for example, patients recover better with a particular medical treatment. At first glance, the treatment might appear beneficial. However, once the patients are divided into subgroups based on age, the same data show that the treatment is harmful within each age group. An incorrect analysis would overlook the importance of these subgroups, mistakenly suggesting the treatment is effective. Importantly, the stratification variable in Simpson’s paradox is assumed to be a confounder, not a collider.

2.2 EVALUATION PROTOCOLS AND DATA

LLM-based Causal Inference Protocols. In this paper, we consider two unified protocols to evaluate LLM reliability (Fig. 2b): **(1) Direct Prompting:** LLMs directly answer causal inference

Table 2: Questions by Difficulty Level for “Simpson’s Paradox”

Difficulty	Question
Very Easy	Evaluate whether {TREATMENT} causally affects {OUTCOME}. Adjust for the known confounder ({CONFOUNDER}) using stratification or regression. State if Simpson’s paradox is present, and provide adjusted rates with 95% confidence intervals along with a recommendation.
Easy	Evaluate whether {TREATMENT} causally affects {OUTCOME}. Consider the potential influence of the confounder ({CONFOUNDER}), adjust accordingly, and assess Simpson’s paradox. Provide adjusted effect sizes with uncertainty estimates and a recommendation.
Medium	Evaluate the causal impact of {TREATMENT} on {OUTCOME}. Account for relevant confounders. Provide adjusted effects with uncertainty measures and justify your recommendation.
Hard	Assess the causal relationship between {TREATMENT} and {OUTCOME}, considering potential confounders.
Very Hard	Evaluate whether {TREATMENT} causally affects {OUTCOME} without additional hints.

questions based on the provided raw data. This approach tests the models’ intrinsic capability to perform causal inference without additional computational tools or external support. **(2) Code-Assisted Prompting:** LLMs generate executable code to perform statistical analysis relevant to the questions, then interpret the results to answer the questions. This method assesses the LLMs’ ability to translate causal reasoning tasks into accurate computational procedures and use analytical results to avoid common pitfalls.

Questions. Each challenge includes five versions of the core question, ranging from very easy to very hard (Table 2). The easier versions give the model more guidance. For example, pointing out the confounder to adjust for or asking directly about Simpson’s paradox. As the difficulty increases, these hints are gradually removed. This setup lets us test whether models can still recognize and handle causal pitfalls when less direction is given (Fig. 2b).

Datasets. To construct datasets tailored to each challenge, we utilize causal graphs following Pearl et al. (2003) and Peters et al. (2017). For every statistical pitfall, we select causal graphs that capture its unique complexities and characteristics. Each challenge is accompanied by five distinct datasets, each containing over 500 samples for comprehensive evaluation. Our simulation approach uses structural causal models based on directed acyclic graphs (DAGs), where each structural equation represents a causal mechanism rather than merely a statistical association. The coefficients in these equations directly encode the causal effects, allowing us to define the ground truth against which inference methods can be evaluated. This approach is mathematically equivalent to simulating potential outcomes under the specified causal structure. The structural equations include both linear and non-linear forms (e.g., non-linear link functions and interaction terms), so the evaluation is not limited to purely linear relationships.

2.3 EVALUATION METRICS

To evaluate the reliability of LLMs for causal inference, we developed detailed grading rubrics for each causal pitfall, informed by guidelines from Sterne et al. (2016); Vandembrouckel et al. (2007). Each benchmark challenge includes multiple questions, each assigned points based on how effectively the model addresses the specific pitfall (see Appendix for detailed rubric). The total *score* for a challenge is the sum of points obtained across these questions, and *max_score* is the maximum achievable score. To enable fair comparisons across challenges, we compute a normalized score:

$$\text{Normalized Score (\%)} = \frac{\text{score}}{\text{max_score}} \times 100\%. \quad (1)$$

We evaluate LLM responses automatically using an independent GPT-4o model (Achiam et al., 2023) to minimize potential biases. To validate the accuracy of this automated evaluation, we additionally engaged three statisticians to manually grade 150 randomly selected responses. We

measure consistency between automated and human scores using the *gap* metric:

$$\text{Gap} = \frac{1}{150} \sum_{i=1}^{150} \frac{|\text{score}_{\text{LLM}}^{(i)} - \text{score}_{\text{human}}^{(i)}|}{s_{\max,i}} \in [0, 1],$$

where $s_{\max,i}$ is the maximum score of corresponding challenge, and $\text{score}_{\text{LLM}}^{(i)}, \text{score}_{\text{human}}^{(i)} \in \mathbb{N}^+$ are scores from automated and human evaluations, respectively. Here, the gap metric ranges from 0 to 1, where 0 indicates perfect agreement.

Finally, to provide a summary metric, we define *causal reliability* as the average normalized score across all benchmark challenges. This measure captures the overall trustworthiness and reliability of LLMs in statistical causal inference tasks.

3 ILLUSTRATIVE PITFALLS IN CAUSAL REASONING

When evaluating causal inference with LLMs, surface-level answers can create an illusion of competence. Statistical causal inference requires grounding conclusions in evidence, checking assumptions, and ruling out alternatives. LLMs, however, may produce confident but flawed outputs that rely on irrelevant cues or statistical artifacts, giving a false sense of robustness.

We illustrate this problem with two failure cases. The first shows how models can base conclusions on superficial semantic cues instead of data, while the second shows how they may misinterpret random variation as causal structure.

Branding Bias: Adversarial Sensitivity to Branding and Semantic Manipulation. As an illustrative example, we examined whether LLMs can be misled by superficial cues when drawing causal conclusions. We constructed a synthetic scenario in which beverage consumption affected health outcomes. In the setting, other factors, lifestyle and health awareness, acted as confounders, but the brand label itself had no causal role (Fig. 3).

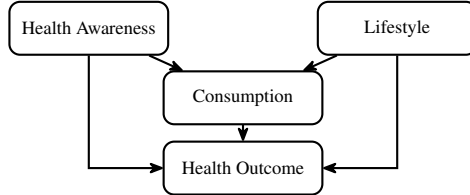


Figure 3: Causal DAG illustrating how beverage consumption, health awareness, and lifestyle affect health outcomes. The beverage’s brand name (“HealthPlus” or “UltraSugar”) does not causally influence outcomes.

Despite identical data, simply changing the brand name from a healthy-sounding label (*HealthPlus*) to a harmful-sounding one (*UltraSugar*) induced changes in LLMs’ conclusions (Table 3). GPT-4o and Gemini-2.0-flash, for example, made a conclusion aligning purely with the brand semantics, rather than the provided data.

The branding bias example shows that LLMs may rely on superficial semantic cues, attributing causal effects to labels even when the underlying data provide evidence **to the contrary**.

Table 3: Branding bias evaluation. Each row represents a combination of the beverage’s label (“HealthPlus” or “UltraSugar”) and the true effect of the given data (beneficial or harmful). A checkmark (✓) indicates a beneficial conclusion and a cross (×) a harmful conclusion; correct inferences are those that match the true effect.

Brand Label	True Effect (Data)	GPT-4o	Gemini-2.0-flash	Claude-3.5-sonnet
HealthPlus	✓	✓	✓	✓
UltraSugar	✓	×	×	✓
HealthPlus	×	✓	✓	✓
UltraSugar	×	×	×	✓

Spurious Causal Inference from Random Patterns. As a second illustrative example, we show how LLMs can mistake random variation in real-world data for genuine causal structure. Specifically, we evaluated LLMs on real data from a PNAS study (Van der Lee & Ellemers, 2015) that analyzed funding success rates across academic disciplines in the Netherlands. When asked if the data reveal gender bias favoring male applicants, all of the tested LLMs drew incorrect conclusions, either attributing the differences in raw percentages directly to gender or considering Simpson’s paradox.

However, a careful statistical analysis as illustrated in the Appendix following Irizarry (2019) shows that women are not disproportionately applying to more competitive disciplines, so Simpson’s paradox does not apply. Furthermore, computing log-odds ratios divided by their standard errors across disciplines and adjusting for multiple comparisons reveals no statistical evidence of gender bias within any department. None of the tested LLMs recognized these key insights.

The spurious inference example shows that LLMs may mistake random variation in observational data for genuine causal structure, failing to apply the statistical checks needed to rule out noise.

4 RESULTS AND ANALYSIS

In this section, we present the key findings from our experiments, examining the causal inference capabilities of ten large language models. Our evaluation spans both closed source and open source LLMs, with the complete list of models provided in the Appendix. All models were evaluated using two protocols: direct prompting and code assisted prompting. The assessment covered six categories of causal inference pitfalls and fifteen challenges, each built from five datasets and five questions of varying difficulty.

Table 4: Causal reliability across causal pitfalls, comparing direct and code-assisted prompting. Values represent averages of normalized scores, defined in equation (1), across five questions per pitfall category; higher scores indicate better performance.

LLM (Direct Prompting)	Conf	Interv	Counter	Med	Disc	Ext	Average
Gemma2-9b	14.00	30.72	8.00	15.89	9.49	4.05	13.69
Llama3.1-8b	18.46	34.78	10.86	29.33	7.71	6.05	17.86
Llama3.1-70b	17.60	35.92	10.29	33.67	11.31	9.92	19.78
Mistral-7b	17.31	29.83	5.71	19.17	8.40	6.18	14.43
Mixtral-8x22b	16.40	32.08	8.00	27.50	9.20	8.75	16.99
Claude-3.5-sonnet	18.74	47.60	12.00	40.50	18.63	19.84	26.22
Gemini-2.0-flash	20.06	37.56	13.43	46.67	13.48	14.93	24.36
Deepseek-chat	25.89	52.42	12.86	53.83	20.83	28.74	32.43
GPT-4.1	17.26	33.57	6.57	53.27	16.43	24.31	25.24
GPT-o4-mini	41.43	45.21	18.57	57.67	36.97	44.48	40.72

LLM (Code-Assisted Prompting)	Conf	Interv	Counter	Med	Disc	Ext	Average
Gemma2-9b	7.89	18.50	4.86	23.04	6.63	18.50	13.24
Llama3.1-8b	11.93	15.99	7.67	18.06	6.23	17.29	12.86
Llama3.1-70b	23.54	22.83	8.21	25.17	14.41	25.17	19.89
Mistral-7b	4.68	13.15	1.43	11.11	6.43	9.09	7.65
Mixtral-8x22b	22.50	30.53	5.44	26.49	14.29	24.31	20.59
Claude-3.5-sonnet	32.91	47.33	11.71	41.73	16.12	26.47	29.38
Gemini-2.0-flash	37.20	42.96	14.29	42.17	16.19	37.98	31.80
Deepseek-chat	38.63	48.70	10.86	47.13	25.79	45.63	36.12
GPT-4.1	47.14	42.65	12.29	49.40	23.87	48.58	37.32
GPT-o4-mini	62.00	51.86	16.96	50.00	26.67	50.71	43.03

Conf: Confounding biases and spurious associations; Interv: Interventions and experimental reasoning; Counter: Counterfactual reasoning and hypotheticals; Med: Mediation and indirect causal effects; Disc: Causal discovery and structure learning; Ext: Causal generalization and external validity.

Overall Performance. Across models, **GPT-o4-mini** demonstrated the highest overall reliability, achieving an average causal reliability of 40.72% under direct prompting and 43.03% under code-

assisted prompting (Table 4). Deepseek-chat, although lower in aggregate (average 32.43% under direct prompting and 36.12% under code-assisted prompting), obtained the strongest performance in *interventions and experimental reasoning* (52.42% under direct prompting and 48.70% under code-assisted prompting). Taken together, these results indicate that optimized mid-scale models can, in certain contexts, outperform larger frontier systems on causal reasoning tasks.

Benefits of Code-Assisted Prompting. Although code-assisted prompting improved several strong models, its benefits were not universal. For example, GPT-o4-mini and Gemini-2.0-flash gained from code execution, increasing their averages from 40.72% to 43.03% and 24.36% to 31.80%, respectively. GPT-4.1 also improved significantly (25.24% $\rightarrow$ 37.32%). Under code-assisted prompting, models first convert raw data into summary statistics through generated code, then reason over those statistics. This separates low-level data parsing from causal reasoning, which partly explains why stronger models benefit more: they produce correct analysis code and reason on clean numerical outputs rather than raw tables. In contrast, some small open-source LLMs showed little benefit: Llama3.1-8B decreased from 17.86% to 12.86%, Gemma2-9B remained unchanged (13.69% to 13.24%), and Mistral-7B even dropped from 14.43% to 7.65%. These results indicate that computational results increase the strengths of already-capable LLMs, but does not provide a uniform advantage across the LLMs. Models with high code-error rates (e.g., Mistral-7B, Llama-8B) have worse scores under code-assisted prompting. Allowing one debugging attempt can bring their performance back to direct-prompting levels. We discuss this further in Appendix H.

Impact of Difficulty Levels. Causal reliability consistently decreases as the question becomes harder. As shown in Table 5, all models performed best on very easy and easy questions, with average scores declining as task difficulty increased, where the question contains fewer hints (see Table 2. For example, GPT-o4-mini achieved 60.72% under direct prompting and 56.73% under code-assisted prompting on very easy items, but dropped to 17.75% and 32.84%, respectively, on very hard questions. Deepseek-chat and Gemini-2.0-flash showed similar trends, maintaining moderate performance on medium and hard levels but falling quickly on the *very hard* questions. These results indicate that code assistance is particularly beneficial for challenging tasks, helping stronger models recover some performance at higher difficulty levels.

Table 5: Causal reliability by difficulty levels of questions, comparing direct and code-assisted prompting. Values represent averages of normalized score, defined in equation (1), across 15 challenges; higher scores indicate better performance.

LLM (Direct Prompting)	Very Easy	Easy	Medium	Hard	Very Hard
Gemma2-9b	20.50	15.64	14.22	6.72	5.46
Llama3.1-8b	27.99	20.85	19.90	10.71	5.82
Llama3.1-70b	31.04	27.38	22.04	11.29	5.23
Mistral-7b	20.04	18.88	18.02	7.49	3.80
Mixtral-8x22b	27.80	21.20	20.31	8.41	5.23
Claude-3.5-sonnet	40.27	36.63	29.69	13.42	8.30
Gemini-2.0-flash	37.83	33.59	27.59	15.72	8.22
Deepseek-chat	46.37	38.92	37.92	28.79	16.36
GPT-4.1	38.01	33.94	33.59	17.31	7.67
GPT-o4-mini	60.72	54.05	48.13	30.72	17.75

LLM (Code-Assisted Prompting)	Very Easy	Easy	Medium	Hard	Very Hard
Gemma2-9b	19.09	18.35	16.43	10.56	7.94
Llama3.1-8b	26.04	14.41	12.77	8.84	7.90
Llama3.1-70b	32.21	27.88	24.77	14.08	7.58
Mistral-7b	9.44	12.03	9.59	6.63	2.73
Mixtral-8x22b	33.14	29.77	25.64	14.59	7.77
Claude-3.5-sonnet	39.70	35.06	33.55	20.92	18.34
Gemini-2.0-flash	46.01	39.83	32.17	28.00	21.67
Deepseek-chat	55.56	48.23	43.69	28.46	16.50
GPT-4.1	57.80	46.49	46.48	24.33	19.65
GPT-o4-mini	56.73	49.66	49.37	34.31	32.84

Persistent Reliability Gaps. Despite variation across models and prompting strategies, significant reliability gaps remain in all settings. Among the tested LLMs, GPT-o4-mini achieved the highest causal reliability score 43.03% on average under code-assisted prompting, while most other models remained well below this level. Performance on difficult causal reasoning challenges was especially limited, with *very hard* questions rarely exceeding 30% even for the best LLMs. Taken together, the results suggest that existing LLMs, when used without finetuning or specialized architectures, remain unreliable to apply in high-stakes causal inference.

Human-LLM Grading Alignment. To validate the fidelity of our automated GPT-4o scoring against expert judgments, we conducted a human validation study on a stratified sample of 150 model responses, with equal representation across the six pitfall categories and five difficulty levels (see Appendix). Three PhD students in statistics independently graded each response using our detailed rubrics. We then compared the resulting human scores with those produced by GPT-4o via the gap metric, which yielded a mean value of 0.11. This close agreement confirms that our GPT-4o evaluator reliably mirrors expert assessments, justifying its use for large-scale, reproducible evaluation of LLM performance in causal inference without the need for extensive human oversight.

Code-Assisted Execution Errors. As shown in Figure 4, code-execution failures peak in the “mediation effects” and “interventions and experimental reasoning” categories, where implementing correct stratification and transportability routines is most demanding. Interestingly, “very easy” questions produce the highest failure rates, whereas “very hard” questions yield lower rates.

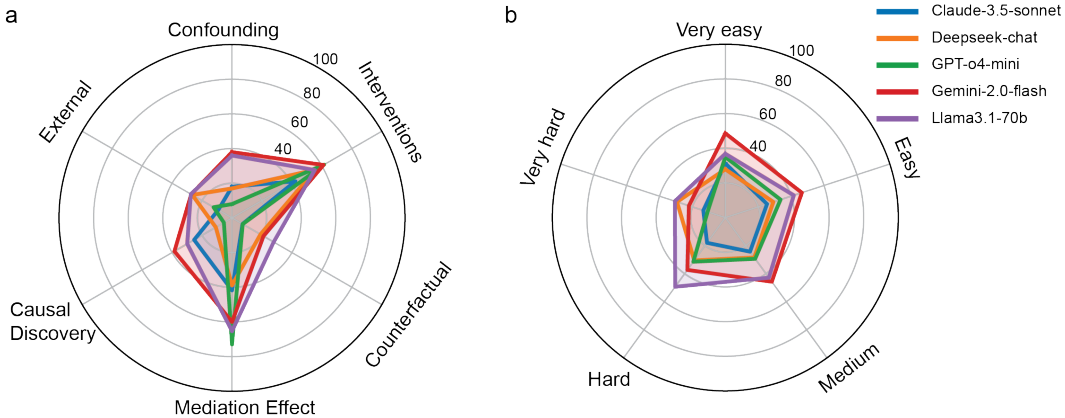


Figure 4: **Code execution failure rates (%) in code-assisted prompting protocol across causal inference challenges and question difficulty.** Failure rate is defined as the percentage of code-generation attempts that either raise execution errors or produce invalid analytical outputs, computed only for the code-assisted prompting protocol. (a) Average failure rate for each of the six causal-inference pitfall categories. (b) Average failure rate by question difficulty level, increasing from very easy through very hard tasks.

5 CONCLUSION

We introduced **CausalPitfalls**, a benchmark designed to rigorously evaluate the reliability of LLMs in performing statistical causal inference. Unlike existing benchmarks that focus primarily on accuracy, our benchmark reveals how LLMs can produce confident yet flawed conclusions by falling into classical statistical pitfalls. Our results indicate substantial gaps in reliability across all models and settings. Even state-of-the-art models exhibit systematic vulnerabilities to confounding, semantic bias, and difficulties in generalizing causal knowledge across contexts. These findings highlight an urgent need for targeted interventions to improve LLMs’ trustworthiness in scientific and policy domains. Future directions include expanding the benchmark to cover more nuanced forms of causal reasoning, such as instrumental variable analysis, latent confounding, and policy evaluation. Additionally, we envision CausalPitfalls as a platform to guide training or fine-tuning strategies that aim to instill causal robustness in LLMs.

More detailed descriptions of our benchmark pitfall categories, challenges, and implementation details are included in the **Appendix** inside the supplementary material.

ACKNOWLEDGMENT

The work was supported in part by the National Science Foundation CAREER Program under grant number 2338506.

ETHICS STATEMENT

All authors have read and adhered to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of our results. Detailed descriptions of datasets, preprocessing steps, and experimental settings are provided in the main text and appendix. The CausalPitfalls benchmark, including all datasets, questions, grading rubrics, and evaluation scripts, is publicly available at <https://github.com/dudududuu/CausalPitfalls>.

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Kwun Chuen Gary Chan, Sheung Chi Phillip Yam, and Zheng Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(3):673–700, 2016.
- Li Chen, Chunlin Li, Xiaotong Shen, and Wei Pan. Discovery and inference of a causal network with hidden confounding. *Journal of the American Statistical Association*, 119(548):2572–2584, 2024.
- Li Chen, Xiaotong Shen, and Wei Pan. Enhancing causal effect estimation with diffusion-generated data. *arXiv preprint arXiv:2504.03630*, 2025.
- Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul G Krishnan, and Chris J Maddison. End-to-end causal effect estimation from unstructured natural language data. *Advances in Neural Information Processing Systems*, 2024.
- Charles R Doss, Guangwei Weng, Lan Wang, Ira Moscovice, and Tongtan Chantararat. A nonparametric doubly robust test for a continuous treatment effect. *The Annals of Statistics*, 52(4):1592–1615, 2024.
- Jin Du, Xinhe Zhang, Hao Shen, Xun Xian, Ganghua Wang, Jiawei Zhang, Yuhong Yang, Na Li, Jia Liu, and Jie Ding. Drift to remember. *arXiv preprint arXiv:2409.13997*, 2024.
- Sander Greenland and James M Robins. Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, 15(3):413–419, 1986.
- Zijian Guo, Domagoj Ćevd, and Peter Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding. *Annals of statistics*, 50(3):1320, 2022.
- Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.
- Rafael A Irizarry. *Introduction to data science: Data analysis and prediction algorithms with R*. Chapman and Hall/CRC, 2019.
- Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauro, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065, 2023.

- Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American statistical Association*, 111(513):132–144, 2016.
- Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.
- Chunlin Li, Xiaotong Shen, and Wei Pan. Nonlinear causal discovery with confounders. *Journal of the American Statistical Association*, 119(546):1205–1214, 2024.
- Zhexiao Lin, Peng Ding, and Fang Han. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica*, 91(6):2187–2217, 2023.
- Zuwan Lin, Arnau Marin-Llobet, Jongmin Baek, Yichun He, Jaeyong Lee, Wenbo Wang, Xinhe Zhang, Ariel J Lee, Ningyue Liang, Jin Du, et al. Spike sorting ai agent. *bioRxiv*, 2025a.
- Zuwan Lin, Wenbo Wang, Arnau Marin-Llobet, Qiang Li, Samuel D Pollock, Xin Sui, Almir Aljovic, Jaeyong Lee, Jongmin Baek, Ningyue Liang, et al. Spatial transcriptomics ai agent charts hpsc-pancreas maturation in vivo. *bioRxiv*, pp. 2025–04, 2025b.
- Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. *Findings of the Association for Computational Linguistics ACL*, 2024.
- Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. In *International Conference on Machine Learning Workshop*, 2023.
- An Luo, Xun Xian, Jin Du, Fangqiao Tian, Ganghua Wang, Ming Zhong, Shengchun Zhao, Xuan Bi, Zirui Liu, Jiawei Zhou, et al. Assistedds: Benchmarking how external domain knowledge assists llms in automated data science. *arXiv preprint arXiv:2506.13992*, 2025.
- David MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2012.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Judea Pearl et al. Causality: Models, reasoning and inference. *Econometric Theory*, 19, 2003.
- Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
- Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. In *Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5043–5053, 2019.
- Paul R Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.
- Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pp. 465–472, 1990.
- Jonathan AC Sterne, Miguel A Hernán, Barnaby C Reeves, Jelena Savović, Nancy D Berkman, Meera Viswanathan, David Henry, Douglas G Altman, Mohammed T Ansari, Isabelle Boutron, et al. Robins-i: a tool for assessing risk of bias in non-randomised studies of interventions. *bmj*, 355, 2016.
- Romy Van der Lee and Naomi Ellemers. Gender contributes to personal research funding success in the netherlands. *Proceedings of the National Academy of Sciences*, 112(40):12349–12353, 2015.
- Jan P Vandenbrouckel, Erik von Elm, Douglas G Altman, Peter C Gotzsche, Cynthia D Mulrow, Stuart J Pocock, Charles Poole, James J Schlesselman, and Matthias Egger. Strengthening the reporting of observational studies in epidemiology (strobe): explanation and elaboration. *PLoS Medicine*, 4(10):1628–1655, 2007.
- Merryn Voysey, Sue Ann Costa Clemens, Shabir A Madhi, Lily Y Weckx, Pedro M Folegatti, Parvinder K Aley, Brian Angus, Vicky L Baillie, Shaun L Barnabas, Qasim E Bhorat, et al. Safety and efficacy of the chadox1 ncov-19 vaccine (azd1222) against sars-cov-2: an interim analysis of four randomised controlled trials in brazil, south africa, and the uk. *The Lancet*, 397(10269): 99–111, 2021.
- Zeyu Wang. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Advances in Neural Information Processing Systems Workshop on Mathematical Reasoning and AI (MATH-AI)*, 2024.
- Song Wei, Ronald Moore, Hanyu Zhang, Yao Xie, and Rishikesan Kamaleswaran. Transfer causal learning: Causal effect estimation with knowledge transfer. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.
- Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality? In *UAI Workshop*, 2022.
- Tianzhong Yang, Jingbo Niu, Han Chen, and Peng Wei. Estimation of total mediation effect for high-dimensional omics mediators. *BMC bioinformatics*, 22:1–17, 2021.
- Kai Yu, Shuang Zhou, Yu Hou, Yiran Song, Min Zeng, Fangqiao Tian, Jin Du, Wenya Xie, Biao Yin, You Chen, et al. Multimodal artificial intelligence agents in healthcare: A scoping review. *Authorea Preprints*, 2025.
- Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023.
- Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. Causalbench: A comprehensive benchmark for causal learning capability of llms. *arXiv preprint arXiv:2404.06349*, 2024.

A THE USE OF LARGE LANGUAGE MODELS (LLMs)

The authors acknowledge the use of ChatGPT to check for potential typographical and grammatical errors in the manuscript.

B BENCHMARKING KEY CHALLENGES IN RELIABLE CAUSAL INFERENCE FOR LLMs

B.1 CONFOUNDING BIASES AND SPURIOUS ASSOCIATIONS

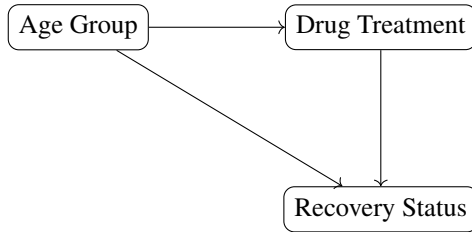
Simpson’s Paradox. Simpson’s paradox is a causal pitfall where an observed association between two variables reverses or disappears when analyzed within subgroups defined by a confounding variable. This paradox typically arises when aggregated data conceal important subgroup-specific relationships, making an apparent correlation misleading or even contradictory.

Motivation and Relevance: Addressing Simpson’s paradox is crucial because misinterpretations of aggregated relationships can lead to serious real-world consequence. LLMs are especially susceptible to this pitfall due to their tendency to overly rely on surface-level correlations and aggregated statistics, while neglecting underlying subgroup structures. Without specifically considering these subgroup, LLMs risk confidently delivering inaccurate causal conclusions.

Data Generation: We design five datasets to illustrate Simpson’s paradox, and each dataset contains three binary variables representing a realistic medical scenario. Here’s an example of data generation setting:

- Age Group (Confounder): Represents subgroup differences that might influence treatment assignment and outcomes (*Young* or *Old*).
- Drug Treatment (Treatment): Indicates whether a patient received a specific medication (*Drug given* or *No drug*).
- Recovery Status (Outcome): Represents patient recovery (*Recovered* or *Not recovered*).

The causal structure underlying the simulation is shown by the DAG below:



In this DAG, Age Group affects both the likelihood of receiving the drug treatment and the probability of recovery. Thus it confounds the relationship between treatment and recovery status. The generated datasets illustrate Simpson’s paradox: when analyzed at an aggregate level, the drug appears beneficial, but within each age subgroup, the drug is actually harmful.

Evaluation Rubric: Models are evaluated based on their ability to recognize and correct for Simpson’s paradox. Full credit (total score) is awarded if the LLM acknowledges the confounder, correctly identifies subgroup-specific relationships, and accurately states that the aggregate-level relationship is misleading. Partial credit is given if the model recognizes the discrepancy but fails to identify or correctly adjust for the confounder. No credit is awarded if the model relies only on aggregate-level associations without mentioning subgroup analysis or confounding.

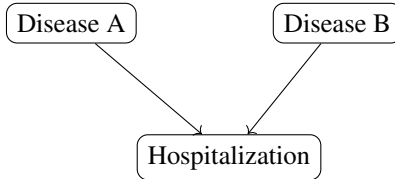
Berkson’s Paradox Berkson’s paradox is a causal pitfall arising from conditioning on a collider variable, thereby inducing a spurious association between two variables that are actually independent. Typically, this paradox appears when the analysis is restricted to a subgroup selected based on a variable influenced simultaneously by two independent factors, creating an artificial correlation.

Motivation and Relevance: Recognizing Berkson’s paradox is essential, as it can lead to severely misguided conclusions. LLMs are particularly prone to this pitfall due to their reliance on readily accessible data summaries and correlations without carefully assessing selection mechanisms or collider structures. Without addressing collider-induced biases, LLMs may confidently report spurious relationships as meaningful causal insights.

Data Generation: We design five datasets to illustrate Berkson’s paradox across realistic scenarios. Each dataset involves three binary variables, with one serving as a collider that induces selection bias. An example dataset represents a hospital-based scenario:

- Disease A: Indicates whether a patient has Disease A (*Present* or *Absent*).
- Disease B: Indicates whether a patient has Disease B (*Present* or *Absent*).
- Hospitalization (Collider): Represents whether the patient is hospitalized (*Yes* or *No*), influenced by both diseases.

The causal structure underlying Berkson’s paradox is depicted by the DAG below:



In this DAG, the variable *Hospitalization* is a collider, as it is influenced by Disease A and Disease B. When analyses are conditioned upon or restricted to hospitalized patients, it artificially creates an association between these two otherwise independent diseases. Thus, Berkson’s paradox emerges through these generated datasets, underscoring the critical importance of avoiding inappropriate conditioning or properly correcting for selection biases.

Evaluation Rubric: Models are evaluated on their ability to recognize and appropriately correct for Berkson’s paradox. Full credit is awarded if the LLM proposes the correct collider structure, states the independence assumption between the variables, proposes to use a valid bias-correction method (such as inverse probability weighting), and suggests no true association after correcting for collider bias. Partial credit is awarded if the model identifies collider bias but fails to state assumptions, proposes the correct correction approach, or suggests independence. No credit is given if the model neglects the collider issue altogether and draws conclusions based only on naïve associations without appropriate adjustment.

B.2 INTERVENTIONS AND EXPERIMENTAL REASONING

Observational vs Experimental Reasoning Observational reasoning refers to inferring causal relationships only from non-experimental, observational data, where the treatment is not randomly assigned. In contrast, experimental reasoning involves randomized controlled trials (RCTs), where treatment assignment is randomized, ensuring that confounding variables are evenly distributed across groups. The causal pitfall emerges when analysts interpret observational associations as causal effects without adequately adjusting for confounding variables, potentially resulting in incorrect causal conclusions.

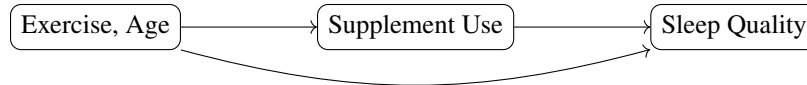
Motivation and Relevance: Distinguishing between observational and experimental reasoning is crucial, as erroneous interpretations of observational data can significantly affect policy decisions, medical recommendations, and economic strategies. LLMs are particularly vulnerable to this pitfall due to their propensity to accept apparent observational correlations at face value, often neglecting confounders or the necessity of proper statistical adjustments. Without instruction or analytical rigor, LLMs may deliver misguided causal interpretations, amplifying the risk of poor decision-making in sensitive real-world applications.

Data Generation: We design five datasets to illustrate the differences between observational and experimental reasoning across realistic scenarios. Each dataset includes a binary treatment and a

binary outcome, along with several confounding variables. Here is an illustrative example scenario focusing on sleep quality:

- **Supplement Use (Treatment):** Indicates whether participants take a sleep-improving supplement (*Yes* or *No*).
- **Sleep Quality (Outcome):** Represents participants' quality of sleep (*High quality* or *Low quality*).
- **True Confounders:** Regular exercise habits (*Exercise regularly*) and older age (*Age old*), each affecting both supplement use and sleep quality.
- **Irrelevant Variables (Non-confounders):** Factors such as coffee drinking, watching TV late, and income, included to test the model's ability to discern relevant confounding from irrelevant variables.

The causal structure illustrating observational confounding is depicted by the DAG:



In this DAG, confounders (Exercise and Age) simultaneously affect both the treatment (Supplement Use) and the outcome (Sleep Quality), introducing bias into naïve observational analyses. Proper adjustment for these confounders is essential to approximate the true causal relationship.

Evaluation Rubric: Models are evaluated based on their capability to distinguish observational correlations from true causal effects. Full credit is awarded if the LLM correctly identifies true confounders, justifies their relevance, applies an appropriate statistical adjustment method (such as stratification or regression), and distinguishes the corrected causal conclusion from naïve observational results. Partial credit is awarded if the model recognizes the importance of confounding but fails to fully justify confounders or properly apply adjustments. No credit is given if the model relies only on observational correlations without addressing confounding or providing adjusted causal interpretations.

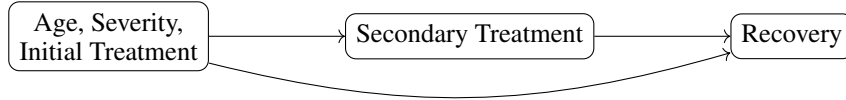
Causal Effect Estimation Causal effect estimation involves quantifying the precise magnitude of the effect that an intervention or treatment has on a particular outcome. Unlike qualitative assessments of causal relationships, effect estimation focuses on numerical measures, such as the average treatment effect (ATE). The primary challenge arises from properly accounting for confounding factors, distinguishing true causal effects from mere correlations or conditional probabilities.

Motivation and Relevance: Accurate estimation of causal effects is critical for informed decision-making in domains such as medicine, public policy, marketing, and technology. Misestimating these effects can lead to misguided interventions, inefficient resource allocation, or even unintended harm. LLMs often encounter difficulties with this task due to their inclination to rely on observational correlations without properly adjusting for confounders. Consequently, they risk confusing correlation or conditional probabilities with actual causal effects, potentially delivering flawed or misleading recommendations.

Data Generation: We design five datasets to evaluate the accuracy and robustness of causal effect estimation methods across realistic scenarios. Each dataset involves a defined binary treatment, binary outcome, multiple relevant confounders, and irrelevant features to test the ability of the model to distinguish between pertinent and non-pertinent information. An illustrative dataset focuses on patient recovery after secondary medical treatment:

- **Secondary Treatment (Treatment):** Indicates whether patients received a second-round medical treatment (*Received* or *Not received*).
- **Recovery (Outcome):** Reflects patient recovery status (*Recovered* or *Not recovered*).
- **True Confounders:** Factors such as patient age, initial illness severity, follow-up severity, and initial treatment, each influencing both the assignment of the secondary treatment and patient recovery.
- **Irrelevant Features:** Variables like socioeconomic status and random noise, intended to evaluate whether the model appropriately excludes irrelevant information.

The causal structure underlying these scenarios is represented by the DAG below:



This DAG shows how confounding variables (Age, Severity, and Initial Treatment) simultaneously affect both the assignment of secondary treatment and the likelihood of recovery. Proper causal estimation requires statistical adjustments, such as inverse probability weighting (IPW), to isolate the true effect of the treatment from these confounding influences.

Evaluation Rubric: Models are evaluated based on their ability to accurately estimate causal effects and rigorously justify their methodological decisions. Full credit is awarded if the LLM defines the target causal estimand (such as ATE), correctly identifies and adjusts for true confounders, excludes irrelevant features, employs suitable causal inference methods (e.g., inverse probability weighting), provides numerical estimates within an acceptable tolerance range, and appropriately quantifies statistical uncertainty. Additionally, full-scoring responses perform at least one diagnostic check (e.g., balance or overlap assessments) and discuss methodological limitations. Partial credit is awarded if some, but not all, of these criteria are adequately addressed. No credit is given for relying only on unadjusted correlations without proper causal reasoning or methodological justification.

B.3 COUNTERFACTUAL REASONING AND HYPOTHETICALS

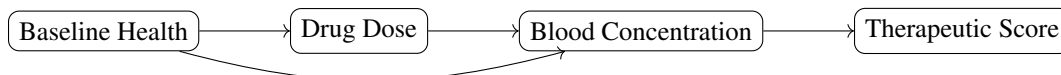
Counterfactual Outcome Prediction Counterfactual outcome prediction involves reasoning about hypothetical scenarios, specifically asking what would have happened if past events or treatments had occurred differently. This type of reasoning goes beyond mere correlation, requiring the careful consideration of alternative scenarios that contradict observed reality, while maintaining logical consistency with established causal relationships.

Motivation and Relevance: Counterfactual reasoning is essential for robust explanations, policy analysis, and informed decision-making across various disciplines, including medicine, economics, education, and environmental policy. Incorrect counterfactual predictions can lead to flawed policy recommendations and misguided interventions. LLMs often struggle with this task, as it requires moving beyond observed data toward hypothetical worlds that differ from reality. LLMs frequently either adhere too closely to observed facts or generate predictions that violate known causal structures, thereby compromising the reliability of their predictions.

Data Generation: We design five datasets for counterfactual outcome prediction tasks across diverse real-world contexts. Each dataset includes defined treatment, outcome, confounder, and downstream variables, enabling manipulation and assessment of counterfactual scenarios. One illustrative example is a clinical scenario involving drug dosage:

- Drug Dose (Treatment): Dosage levels administered to patients.
- Blood Concentration (Outcome): The resulting concentration level in patient blood following dosage.
- Baseline Health (Confounder): Initial health condition influencing both drug dosage and blood concentration.
- Therapeutic Score (Downstream Variable): A health outcome measure influenced by blood concentration but not directly adjustable in the counterfactual scenario.

The causal structure governing these scenarios is represented by the DAG:



In this DAG, Baseline Health is a confounder affecting both the treatment (Drug Dose) and the outcome (Blood Concentration). The Therapeutic Score is a downstream variable, influenced only by the outcome and thus not to be controlled when evaluating counterfactuals. This challenge asks for

predictions under hypothetical alterations of the treatment variable, carefully ensuring consistency with the causal framework.

Evaluation Rubric: Models are evaluated based on their capability to accurately generate and reason about counterfactual outcomes. Full credit is awarded if the LLM identifies treatment, outcome, confounder, and downstream variables, proposes an appropriate causal DAG, acknowledges underlying assumptions (e.g., absence of hidden confounding), applies a valid method (such as regression adjustment) to estimate the counterfactual outcome, and provides numeric predictions that align closely with known counterfactual values. Partial credit is granted if the model correctly addresses some but not all aspects, such as identifying variables and providing accurate reasoning but failing to provide numerical accuracy or state assumptions. No credit is given if the model ignores causal structure, confounding, or essential assumptions, providing counterfactual predictions based only on observed correlations without causal reasoning.

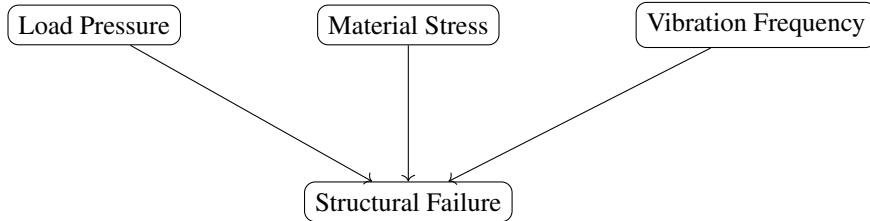
Causal Necessity and Sufficiency Evaluating causal necessity and sufficiency involves determining whether a specific factor is required (necessary) or alone capable (sufficient) to produce an outcome. Necessity examines if the outcome would fail to occur in the absence of the cause, whereas sufficiency assesses if the presence of the cause alone invariably leads to the outcome. This analysis is particularly crucial in scenarios with multiple, potentially redundant or overlapping causal factors.

Motivation and Relevance: Correctly assessing causal necessity and sufficiency is fundamental for rigorous explanation, accountability, and policy decisions. Misclassification can lead to incorrect attribution, misguided interventions, or failure to identify effective alternative measures. LLMs commonly encounter difficulties with this sophisticated causal reasoning; they often erroneously assume necessity in the presence of alternative sufficient causes or incorrectly attribute sufficiency without thorough evaluation. Expert causal reasoning involves considering counterfactual scenarios — imagining the outcome under conditions where a causal factor is removed or is the sole influencing factor.

Data Generation: We simulate five datasets designed to test the understanding of causal necessity and sufficiency across diverse applied contexts. Each scenario includes a focal causal factor (X), outcome variable (Y), and additional contextual factors. An illustrative example involves material stress in engineering structures:

- **Material Stress (Focal Cause):** Evaluates whether elevated material stress is necessary or sufficient for structural failure.
- **Failure Probability (Outcome):** Indicates whether structural failure occurs beyond a specific threshold.
- **Contextual Factors:** Load pressure and vibration frequency, serving as potential alternative or complementary causal factors influencing failure.

The causal structure underlying necessity and sufficiency is illustrated by the DAG:



In this DAG, the focal causal factor (Material Stress) is evaluated alongside alternative or complementary causes (Load Pressure, Vibration Frequency) to determine if it is necessary or sufficient for structural failure. Proper analysis requires reasoning about interventions and counterfactual scenarios, comparing outcomes under varying causal conditions.

Evaluation Rubric: Models are evaluated based on their capability to rigorously define and correctly classify causal necessity and sufficiency. Full credit is awarded if the LLM defines global necessity and sufficiency using accurate counterfactual reasoning, mentions critical thresholds, thoroughly analyzes outcomes under varied interventions, describes potential outcomes, and correctly classifies

the causal factor based on the provided scenarios. Partial credit is awarded for responses that partially fulfill these criteria—such as accurately defining concepts and analyzing some interventions but not comprehensively covering all required elements. No credit is given if the model fails to apply proper counterfactual logic, ignores alternative causes, or incorrectly classifies necessity and sufficiency without justification.

B.4 MEDIATION AND INDIRECT CAUSAL EFFECTS

Mediation analysis focuses on understanding how a treatment influences an outcome through intermediate variables known as mediators. Identifying these indirect pathways accurately is essential, as interventions typically propagate through multiple, intertwined channels. However, precise estimation of mediated effects raises methodological challenges.

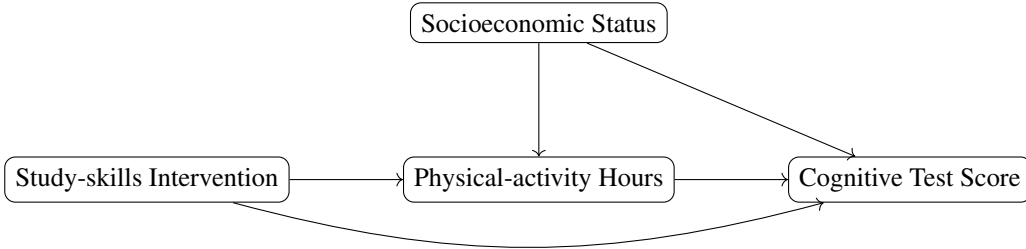
Mediator–Outcome Confounding Mediator–outcome confounding arises when a third variable influences both the mediator (M) and the outcome (Y), biasing the estimated indirect (mediation) and direct effects of a treatment (T). In the canonical structure $T \rightarrow M \rightarrow Y$, a confounder (C) that affects both M and Y distorts mediation estimates unless it is properly controlled. The task therefore requires distinguishing true mediation pathways from correlations induced by common causes of the mediator and the outcome.

Motivation and Relevance: Failure to address mediator–outcome confounding can lead to severely biased causal interpretations and misguided policies. For instance, when evaluating whether physical activity mediates the effect of an educational program on cognitive performance, socioeconomic status may independently influence both activity levels and test scores. LLMs frequently overlook this bias: they may attribute the entire treatment effect to the mediating variable or misestimate indirect and direct effects because they ignore hidden pathways. Correct reasoning demands recognising potential confounders of M and Y and applying specialised methods, such as sequential g -estimation, inverse-probability-weighted mediation analysis, or parametric g -formulae, to obtain unbiased mediation effects.

Data Generation: We construct five datasets that embed mediator–outcome confounding in realistic settings. A representative scenario examines an educational intervention:

- Treatment (T): Receipt of a study-skills intervention (*Yes / No*).
- Mediator (M): Weekly physical-activity hours.
- Outcome (Y): Post-intervention cognitive test score.
- Confounder (C): Socioeconomic status (SES), influencing both activity and test performance.

The causal structure is illustrated by the DAG below:



Evaluation Rubric: Models are evaluated on their ability to detect and correct for mediator–outcome confounding. Full credit is awarded when the response identifies the confounder that influences both the mediator and the outcome, employs and justifies a valid adjustment method (for example, sequential g -estimation or inverse-probability weighting), provides numerical estimates of the natural direct and indirect effects within the specified tolerance, reports and interprets a measure of statistical uncertainty (such as a confidence interval or standard error). Partial credit is given when only some of these elements are addressed. For instance, if the confounder is identified and a method applied but uncertainty quantification or a limitations discussion is omitted. No credit is awarded if

mediator–outcome confounding is ignored, only unadjusted mediation estimates are presented, or methodological choices are left unjustified.

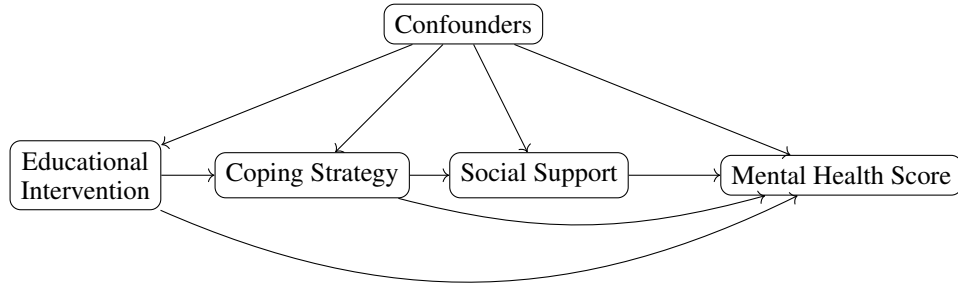
Sequential Mediators Sequential mediation occurs when an intervention exerts its effect on an outcome through a chain of two or more intermediate variables. The task requires disentangling each step in the mediator sequence: identifying how the treatment first influences one mediator, how that mediator then affects the next, and so on until the final outcome.

Motivation and Relevance: Many real-world interventions operate via multiple stages. For example, an educational program that first enhances coping strategies, which in turn builds social support, ultimately improving mental health. Ignoring the sequential nature of these mediators can misattribute effects to the wrong pathway and lead to suboptimal or even counterproductive interventions. LLMs are especially challenged by this scenario because they must recognize the ordered dependencies among mediators and apply methods (such as path-specific effects analysis or structural equation models) rather than treating mediators as independent channels.

Data Generation: We simulate five datasets reflecting a realistic educational intervention scenario:

- Educational Intervention (Treatment): Whether students receive a skills-training program.
- Coping Strategy Improvement (Mediator 1): Increase in students’ coping skills following the intervention.
- Social Support Enhancement (Mediator 2): Growth in support networks resulting from improved coping.
- Mental Health Score (Outcome): Final assessment of students’ psychological well-being.
- Confounders: Socioeconomic status, school quality, and baseline cognitive score (each affecting treatment assignment and mediators).

The underlying causal structure is depicted by the DAG:



Evaluation Rubric: Models are evaluated on their ability to recover the sequential mediation structure and produce unbiased path-specific estimates. Full credit is awarded if the response correctly identifies all true confounders and both mediators in their proper order, excludes irrelevant variables, selects and justifies an appropriate sequential mediation method (for example, path-specific effects analysis or structural equation modeling), contrasts adjusted indirect effects with naïve estimates, provides numerical effect estimates within the specified tolerance, reports and interprets statistical uncertainty, and acknowledges observational limitations. Partial credit is given when only some of these elements are addressed (for instance, correctly identifying mediators but omitting uncertainty quantification). No credit is given if the model ignores the sequential pathway, applies only naïve mediation analysis, or fails to justify methodological choices.

Treatment–Mediator Interaction Effects Treatment–mediator interaction arises when the magnitude of a treatment’s effect on an outcome varies with the level of a mediator. In other words, the indirect pathway through the mediator modifies the direct effect of the treatment. Ignoring these interactions can mask important synergies or antagonisms, leading to incorrect estimates of direct and indirect effects.

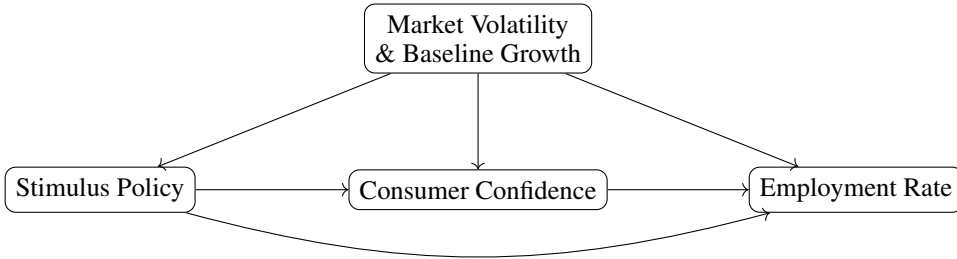
Motivation and Relevance: Accurately modeling treatment–mediator interactions is vital because policies or interventions may only be effective under specific mediator conditions. For example, an

economic stimulus might only boost employment significantly when consumer confidence is already high. LLMs often default to additive causal models and overlook such interaction terms, risking misleading conclusions and suboptimal policy or clinical recommendations.

Data Generation: We design five datasets to illustrate treatment–mediator interaction effects in an economic context.

- **Stimulus Policy (Treatment):** Indicates whether a region implements an economic stimulus (*Yes/No*).
- **Consumer Confidence (Mediator):** Public confidence level (*High/Low*), which may modify the policy’s effect.
- **Employment Rate (Outcome):** Regional employment outcome (*Increased/Not increased*).
- **Confounders:** Market volatility and baseline economic growth, each affecting treatment assignment, consumer confidence, and employment.
- **Irrelevant Variable:** An unrelated indicator, included to test the model’s ability to exclude non-pertinent features.
- **Interaction Term:** The product of Stimulus Policy and Consumer Confidence, included in the outcome model to capture effect modification.

The underlying causal structure is depicted by the DAG below:



Evaluation Rubric: Models are evaluated on their ability to identify true confounders and the mediator, articulate how the treatment–mediator interaction modifies the causal effect, justify exclusion of irrelevant variables, select and justify an interaction-aware estimation method (such as regression with an interaction term), compare adjusted estimates to naïve associations, and provide numerical results with uncertainty measures. Full credit requires all these elements; partial credit is given when only some are addressed; no credit is awarded if interactions are ignored or methodological justifications are missing.

B.5 CAUSAL DISCOVERY AND STRUCTURE LEARNING

Causal discovery and structure learning involve identifying unknown causal relationships and constructing comprehensive causal graphs from observational or experimental data. This type of analysis is distinct from tasks that estimate causal effects or perform mediation analysis. The primary goal here is to discover if causal links exist, determine their direction, and outline the overall causal structure from scratch. The process includes analyzing patterns of statistical dependence and conditional independence among variables. This can be particularly challenging for LLMs because they typically rely on correlations rather than deeper causal insights.

Cause–Effect Direction Inference Determining the correct causal direction between two correlated variables, deciding whether $X_1 \rightarrow X_2$ or $X_2 \rightarrow X_1$, is fundamental to causal discovery. This task requires more than observing associations; it demands identifying which variable truly generates the other, often by leveraging auxiliary information beyond the raw correlation.

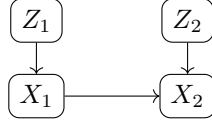
Motivation and Relevance: Inferring the wrong direction can lead to ineffective or harmful interventions, as policies and predictions depend critically on understanding which variable to manipulate. LLMs typically struggle here because they rely on surface-level correlations or heuristics. In practice,

experts use additional clues (such as instrumental variables, temporal ordering, or distributional asymmetries) to distinguish cause from effect.

Data Generation: We design five datasets in which two exogenous instruments (Z_1, Z_2) each influence exactly one of two target variables (X_1, X_2), and a single true causal edge runs between X_1 and X_2 . An illustrative dataset is generated as follows:

- Z_1, Z_2 (*Instruments*): Independent variables that affect X_1 or X_2 respectively, but not each other.
- X_1, X_2 (*Targets*): Two correlated variables linked by a true causal arrow (e.g. $X_1 \rightarrow X_2$).
- Noise terms: Independent additive errors for each variable.

The causal structure for this illustrative dataset is depicted below:



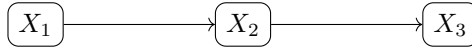
Here, each instrument Z_i is designed to have a nonzero effect only on the corresponding X_i , and the single arrow $X_1 \rightarrow X_2$ represents the ground-truth causal direction.

Evaluation Rubric: Models are evaluated on their ability to assign instruments correctly, carry out the two required regressions of X_1 and X_2 on $\{Z_1, Z_2\}$, report p -values and identify which instrument is significant in only one regression, infer the causal arrow accordingly, and match that inference to the known truth. Full credit is awarded when all these elements are present; partial credit when some are addressed; and no credit if the solution relies only on correlations or omits any required component.

Handling Uncertainty in Causal Structures Handling uncertainty in causal structure inference requires acknowledging that, for a given set of observed conditional independencies, multiple Directed Acyclic Graphs (DAGs) may be equally compatible with the data. This phenomenon, known as Markov equivalence, means that one cannot definitively pinpoint a single causal graph without additional assumptions or interventions.

Motivation and Relevance: Properly communicating uncertainty over plausible causal models is essential in fields such as epidemiology, economics, and social science, where decisions hinge on structural conclusions. Overconfident or overly definitive causal claims can mislead policy-makers or clinicians. While human experts often report equivalence classes or express confidence intervals over edges, LLMs tend to present a single “best” graph, failing to convey ambiguity or the need for further evidence.

Data Generation: We sample five SEM datasets, and in each case, the true data-generating DAG is a simple forward chain $X_1 \rightarrow X_2 \rightarrow X_3$, and independent gaussian noise terms are added with no hidden confounding. The model is chosen that the forward chain and its two Markov-equivalent counterparts are indistinguishable by purely observational tests.



Evaluation Rubric: Models are evaluated on whether they perform necessary marginal and conditional independence tests, list all Markov-equivalent DAGs (including the forward chain, its reverse-fork, and the reverse chain) without proposing any unsupported structures, and refer to Markov equivalence or equivalence-class ambiguity rather than presenting a single graph with unwarranted certainty. Full credit requires correct recognition and expression of uncertainty among all valid graphs; partial credit is given for acknowledging some but not all aspects of the equivalence class or uncertainty; no credit is awarded if the model reports a single graph as definitive or omits the equivalence consideration altogether.

B.6 CAUSAL GENERALIZATION AND EXTERNAL VALIDITY

Causal generalization and external validity address the critical issue of whether a causal relationship identified in one context remains valid when applied to another context. Unlike internal validity,

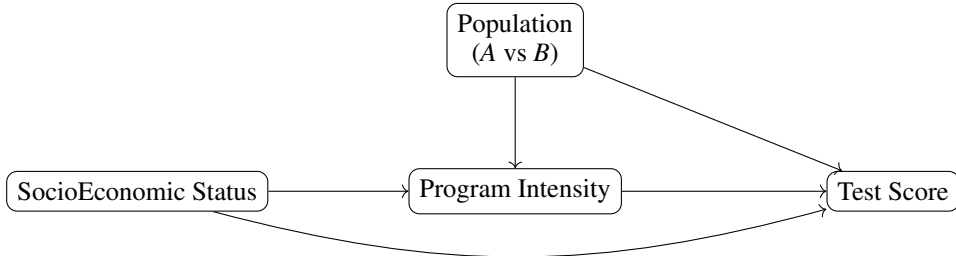
which confirms that observed effects are truly due to the intervention within a specific setting, external validity assesses the transferability of causal claims across varying conditions, populations, or environments. LLMs often assume learned causal relationships universally apply, failing to detect subtle contextual differences that could invalidate such assumptions. Robust causal reasoning thus demands recognizing the limits and scope of causal knowledge.

Population Shift and Transferability. Population shift arises when causal effects estimated in one group (e.g. Population A) do not generalize to another group (Population B) due to differences in baseline characteristics or effect modification. Transferability requires recognising when and how causal conclusions must be adjusted before being applied across populations.

Motivation and Relevance: Failing to account for population shift can lead to ineffective or harmful interventions when policies validated in one demographic are indiscriminately exported to another. For example, an educational program that boosts test scores in urban schools may underperform or backfire in rural settings with different socioeconomic profiles. LLMs often extrapolate causal claims without acknowledging population nuances; robust methods are needed to ensure safe and effective transfer.

Data Generation: We design five datasets comparing two populations (A vs. B). Each record includes:

- Population: Indicator (A or B), determining which structural parameters apply.
- SocioEconomic Status: A baseline covariate influencing both program uptake and outcomes.
- Program Intensity: The “dose” of an educational intervention.
- Test Score: The measured outcome.



Evaluation Rubric: Models are evaluated on their ability to distinguish Population A and B , report the coefficient of Program Intensity on Test Score (with statistical significance or uncertainty) for each population, compare the magnitudes of these effects across populations, and summarize the implications for transferability. Full credit is awarded when all these elements are present and correctly interpreted; partial credit when only some are addressed; and no credit if the model ignores population distinctions or fails to compare and contextualize effect estimates across groups.

Temporal Stability of Causal Effects Temporal stability examines whether a causal effect remains constant across different time periods or whether it shifts due to changing conditions. The task requires detecting and quantifying any differences in the treatment’s impact on the outcome between an initial period and a later period.

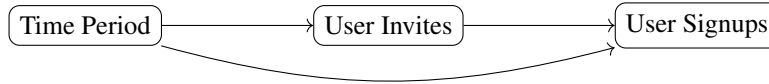
Motivation and Relevance: Evaluating temporal stability is vital for reliable forecasting and adaptive policy-making. An intervention that drives strong effects early on may weaken, or even reverse, its impact as underlying dynamics evolve. LLMs often assume static relationships, risking misleading guidance when effects drift. Experts counter this by stratifying analyses by time period, including interaction terms with time, or using time-varying coefficient models to capture and adjust for temporal change.

Data Generation: We simulate five datasets illustrating temporal drift in a user-engagement scenario. Each record includes:

- Time Period: Indicator of Period 1 or Period 2.
- User Invites (Treatment): Number of invitations sent to potential users.

- User Signups (Outcome): Number of users who sign up.

The causal structure is depicted by the DAG below:



Evaluation Rubric: Models are evaluated on their ability to distinguish Period 1 from Period 2, fit separate analyses of signups on invites for each period, report the estimated effect sizes within the specified tolerance, discuss statistical significance, interpret whether the effect has strengthened or weakened over time, and conclude appropriately about temporal variation. Full credit is awarded when all these elements are present; partial credit if only some aspects are addressed; and no credit if temporal drift is ignored or treated as time-invariant.

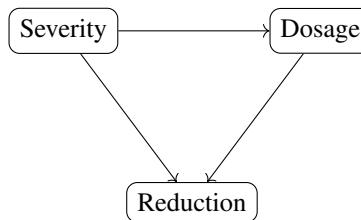
Contextual Interaction and Moderation Effects Contextual interaction, or moderation, occurs when the effect of a treatment on an outcome depends on the level of another variable (the moderator). The task is to identify and quantify how this moderator alters the treatment’s impact, rather than assuming a uniform effect across all contexts.

Motivation and Relevance: Accounting for moderation is critical because interventions may only be effective under particular conditions. For example, a pain medication might relieve symptoms in mild arthritis but be less effective or even counterproductive in severe cases. LLMs often overlook such nuances, treating effects as homogeneous and risking flawed recommendations. Experts model interaction terms or conduct stratified analyses to capture these context-dependent effects.

Data Generation: We design five datasets illustrating dosage–severity interactions in clinical and biological settings.

- Condition Severity: A continuous or ordinal measure of baseline severity.
- Treatment Dosage: Dosage level administered.
- Symptom Reduction: The observed decrease in symptoms.
- Interaction Term: The product of Condition Severity and Treatment Dosage is included in the outcome model to generate a moderation effect.

The causal structure is represented by the DAG below:



Evaluation Rubric: Models are evaluated on their ability to report the main effect of dosage, the main effect of severity, and the interaction coefficient; to interpret how severity levels moderate the dosage effect; to perform and mention any diagnostic checks (e.g. residual analysis); and to summarize the implications for tailored dosing. Full credit is awarded when all elements are addressed with appropriate numerical or qualitative detail; partial credit when only some are; and no credit if moderation is ignored or methodological justifications are omitted.

Domain Shift and Transportability of Causal Knowledge Domain shift arises when causal conclusions drawn in one population fail to generalize to another due to differences in baseline characteristics or effect modifiers. Transportability concerns identifying these differences and adjusting causal inferences so that interventions remain valid across domains.

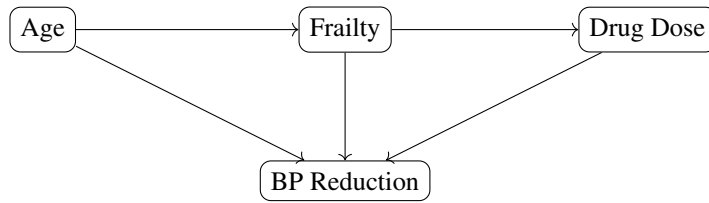
Motivation and Relevance: Ignoring domain shift can lead to interventions that backfire or underperform when moved to new settings. For example, a drug shown to reduce blood pressure in

middle-aged adults may be less effective in older patients with higher frailty. LLMs often extrapolate causally without accounting for such nuances, whereas experts employ statistical techniques to pinpoint and correct for population differences, ensuring safe and effective transfer of causal knowledge.

Data Generation: We design five datasets contrasting two age-defined cohorts (Group 1: ages 30–50; Group 2: ages 50–70).

- Age: Numeric age to determine the membership of the cohort.
- Frailty: A health indicator positively correlated with age.
- Drug Dose (Treatment): Dose of drug taken.
- Blood Pressure Reduction (Outcome): Measured drop in systolic blood pressure.

This design forces models to grapple with shifting treatment effects under different demographic distributions.



Evaluation Rubric: Models are evaluated on their ability to propose splitting the data into the correct age strata, specify and justify a linear model of blood pressure reduction on drug dose and frailty within each stratum, condition the estimate on the individual’s age scenario (e.g. age = 65), recognize frailty as a necessary input for precise effect estimation, explain how frailty confounds the dose–response in the older cohort, and request any additional information needed for a complete answer. Full credit is awarded when all elements are addressed; partial credit when only some are present; and no credit if domain differences are ignored or methodological steps are omitted.

C DETAILS OF TWO PROTOCOLS

We evaluated models on causal inference tasks using two different prompting approaches: **Direct Prompting** and **Code Assisted Prompting**. Each method tests a distinct aspect of the model’s abilities.

C.1 PROTOCOL 1: DIRECT PROMPTING

In the **Direct Prompting** protocol, we test the model’s ability to analyze causal questions directly from raw data, without using computational code.

Workflow for Direct Prompting

1. **Model Prompting:** Present the question alongside the raw data to the model.
2. **Output Collection:** Record the model’s response for evaluation.

Example:

"Question: Evaluate whether {TREATMENT} causally affects {OUTCOME}.
Data: {data}
Provide your analysis without using code."

This method focuses on the model’s ability to reason intuitively and draw conclusions directly from the provided data.

C.2 PROTOCOL 2: CODE-ASSISTED PROMPTING

The **Code-Assisted Prompting** protocol tests whether the model can identify causal issues, generate relevant Python code to analyze the data, and interpret the numerical outcomes to resolve the causal question.

Workflow for Code Assisted Prompting

1. **Code Generation:** Provide the model with the causal question, dataset location, column names, and a small data sample (10 rows). Request Python code for analysis.
2. **Code Execution:** Extract and run the generated Python code to obtain numerical results.
3. **Result Interpretation:** Show the model the code it generated and its numerical results, and ask it to interpret these results in context.
4. **Output Collection:** Record the model’s interpretation and analysis for evaluation.

Example (Code Generation Prompt):

"Question: Evaluate whether {TREATMENT} causally affects {OUTCOME}.
Dataset location: /path/to/dataset.csv
Columns: {col1, col2, col3, ...}
Data sample (first 10 rows): {data_sample}
Provide Python code to perform this analysis."

Example (Results Interpretation Prompt):

"Question: Evaluate whether {TREATMENT} causally affects {OUTCOME}.

Code:

{Python code generated by the model}

Results obtained:

{Numerical results from code execution}

Based on these results, provide your analysis and answer the causal question."

Together, the two protocols provide complementary insights, evaluating both the intuitive reasoning and computational planning of models in causal inference tasks.

D CAUSAL PITFALL EXAMPLES

This appendix provides detailed descriptions of two illustrative examples as introduced in Section 3.

D.1 BRANDING BIAS: ADVERSARIAL SENSITIVITY TO BRANDING AND SEMANTIC MANIPULATION

To highlight the susceptibility of LLMs to semantic manipulations, we designed an adversarial scenario testing whether beverage branding influences LLM-based causal conclusions about health impacts. The underlying causal structure was fixed, ensuring beverage consumption directly affected health outcomes positively or negatively, while lifestyle and health awareness independently affected both beverage consumption and health. Importantly, the beverage brand names ("HealthPlus" or "UltraSugar") themselves had no actual causal effect (see Figure 3 in the main paper).

Dataset Construction. We generated two datasets to represent four distinct scenarios, each combining a brand name ("HealthPlus", "UltraSugar") with a true health effect (beneficial or harmful):

1. Brand "HealthPlus", truly beneficial effect.
2. Brand "HealthPlus", truly harmful effect.
3. Brand "UltraSugar", truly harmful effect.
4. Brand "UltraSugar", truly beneficial effect.

Each dataset included 200 samples with variables: Consumption, Outcome (health impact), Health Awareness, and Lifestyle.

LLM Performance and Observations. LLMs were asked to assess if each beverage ("HealthPlus" or "UltraSugar") was beneficial or harmful based purely on the given data. Table 3 in the main paper summarizes the LLM conclusions across scenarios.

D.2 SPURIOUS CAUSAL INFERENCE FROM RANDOM PATTERNS.

We analyze the dataset `research_funding_rates` from the `dslabs` R package to illustrate how a rigorous causal analysis should be conducted.

We first construct a dataset containing the number of applications, awards, and success rates for each gender. Disciplines are then re-ordered by their overall success rate from the original dataset. To examine whether the data exhibit Simpson’s paradox, we plot success rates by discipline (ordered by overall success), using color to indicate gender and point size to reflect the number of applications.

As shown in Figure 5, there is no clear confounder driving the observed patterns. Nonetheless, some fields appear to favor men, while others favor women. Notably, the two disciplines with the largest differences favoring men are also those with the largest number of applications.

This raises the question: could some selection committees be biased while others are not? To investigate, we compute the log-odds ratio divided by its standard error for each discipline and examine their distribution using a Q-Q plot against the standard normal distribution.

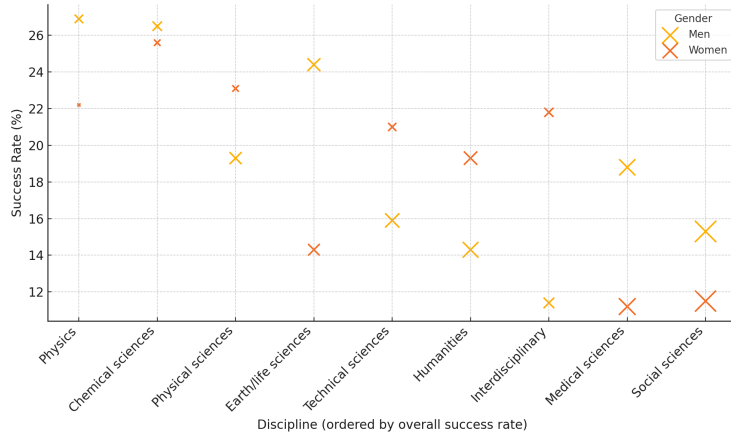


Figure 5: Success rates across disciplines for men and women.

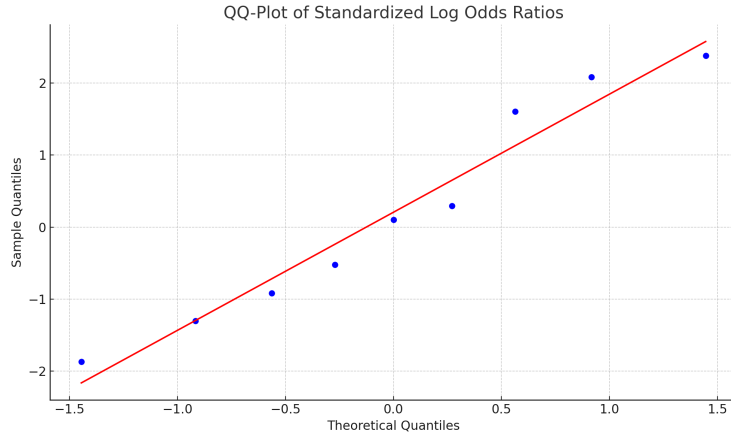


Figure 6: QQ-plot of standardized log odds ratios across disciplines.

Figure 6 shows that the points lie close to the reference line, indicating that the observed effect sizes across disciplines are consistent with random variation under the null hypothesis of no systematic bias. The single “significant” result in Medical Sciences, which emerges under a chi-square test, is plausibly attributable to chance when conducting multiple comparisons, rather than reflecting genuine committee-level bias.

E HUMAN-LLM GRADING ALIGNMENT

To assess the reliability of our GPT-4o-based automated evaluation approach, we conducted a validation involving expert human graders. This section elaborates on our detailed validation procedure.

Sampling Procedure. We randomly selected 150 LLM-generated responses from our complete evaluation dataset, employing **stratified** random sampling to ensure balanced representation across:

1. First, we divided responses into the six causal inference categories, selecting an equal number (25) from each category.
2. Within each category, we evenly sampled across the five difficulty levels (very easy, easy, medium, hard, very hard), selecting exactly 5 responses per difficulty level.
3. For each category-difficulty combination, we randomly selected responses from the evaluated LLMs, ensuring proportional representation of all models’ outputs.

This sampling approach ensured the validation set accurately represented the complexity, diversity, and balanced coverage of our entire evaluation dataset.

Human Grading Procedure. Each sampled response was independently evaluated by three PhD students majoring in statistics. Prior to grading, the evaluators received detailed instructions on the grading rubrics to ensure consistency in scoring. Each grader assessed responses independently based on these rubrics, after which we aggregated their individual scores by taking the arithmetic mean. The average *Gap* (absolute scoring differences) between human-expert and GPT-4o automated scores across causal inference categories, are summarized in Table 6.

Table 6: Average *Gap* (absolute scoring differences) between human graders and GPT-4o across causal inference categories (25 samples per category).

Causal Pitfall Category	Average Gap
Confounding biases and spurious associations	0.08
Interventions and experimental reasoning	0.05
Counterfactual reasoning and hypotheticals	0.15
Mediation and indirect causal effects	0.17
Causal discovery and structure learning	0.15
Causal generalization and external validity	0.05
Overall (all categories)	0.11

F ADDITIONAL EXPERIMENTAL SETUP

Our experiments included the following ten large language models: Llama3.1-8b, Llama3.1-70b, Mistral-7b, Mixtral-8x22b, Claude-3.5-sonnet, Gemma2-9b, Gemini-2.0-flash, Deepseek-chat, GPT-4.1, and GPT-o4-mini.

All experiments and analyses were conducted using a single NVIDIA A100 GPU. API interactions were handled using the official SDKs provided by each model vendor (OpenAI, Anthropic, Google, Deepseek).

Hyperparameters were consistently set across models to ensure fair comparisons. The maximum token length was standardized at 1000 tokens for all models. GPT o4-mini was configured using the parameter `reasoning_effort="medium"`.

G ADDITIONAL EXPERIMENTAL RESULTS

Tables 7 and 8 provide comprehensive evaluations of LLM reliability at the level of individual causal inference challenges, under both direct and code-assisted prompting protocols.

H CODE-ASSISTED PROMPTING WITH DEBUGGING

To further investigate whether the performance gap under code-assisted prompting arises from unreliable code generation, we introduce a variant: **Code-Assisted Prompting with Debugging**. This protocol extends Protocol 2 by allowing the model one opportunity to correct its code when execution fails. Specifically, when the generated code raises an error, we present the error message to the model and request a fixed version and re-execute the code.

Table 7: Direct Prompting: normalized scores (%) of LLMs across all individual causal inference challenges.

LLM	CE	CD	CI	CF	DS	MC	NS	OE	PS	SB	SM	SP	SU	TS	TM
Gemma2-9b	38.50	2.40	2.00	15.43	11.33	25.00	2.21	22.94	0.00	4.00	19.50	24.00	16.57	2.86	5.00
Llama3.1-8b	40.00	4.00	4.00	21.71	15.33	38.50	1.54	29.56	0.00	7.20	29.00	29.71	11.43	6.86	20.50
Llama3.1-70b	37.00	7.20	7.33	20.00	16.67	43.00	3.87	34.83	5.60	7.20	36.50	28.00	15.43	15.43	21.50
Mistral-7b	37.00	4.80	2.00	10.86	21.33	18.50	2.04	22.67	0.80	7.20	25.00	27.43	12.00	0.57	14.00
Mixtral-8x22b	39.50	6.40	5.33	16.00	18.00	35.00	2.64	24.67	4.80	8.80	33.00	24.00	12.00	10.86	14.50
Claude-3.5-sonnet	51.88	15.20	13.33	17.71	18.06	59.00	9.73	43.33	25.60	7.20	40.50	30.29	22.36	22.29	22.00
Gemini-2.0-flash	43.00	9.60	10.67	20.00	25.69	60.50	9.87	30.76	11.20	6.40	48.50	33.71	20.41	12.57	31.00
Deepseek-chat	58.00	20.80	19.33	24.00	25.33	80.50	5.80	46.83	40.00	15.20	48.00	36.57	20.88	30.29	33.00
GPT-4.1	36.72	18.40	16.67	10.29	22.67	65.79	6.59	31.56	38.40	8.80	61.31	25.71	14.29	18.29	37.00
GPT-o4-mini	50.62	40.80	32.00	25.14	28.67	83.00	16.51	39.79	54.40	36.00	52.50	46.86	33.14	62.86	37.50
Average	43.22	12.96	11.27	18.11	20.31	50.88	6.08	32.69	18.08	10.80	39.38	30.63	17.85	18.29	23.60

CE: Causal effect estimation; CD: Cause-effect direction; CI: Contextual interaction; CF: Counterfactual prediction; DS: Domain shift. MC: Mediator-outcome confounding; NS: Necessity and sufficiency; OE: Observational vs experimental; PS: Population shift; SB: Selection bias. SM: Sequential mediators; SP: Simpson’s paradox; SU: Structure uncertainty; TS: Temporal stability; TM: Treatment-mediator interaction.

Table 8: Code-Assisted Prompting: normalized scores (%) of LLMs across all individual causal inference challenges.

LLM	CE	CD	CI	CF	DS	MC	NS	OE	PS	SB	SM	SP	SU	TS	TM
Gemma2-9b	21.50	6.40	32.67	9.71	6.00	18.12	1.73	15.50	1.60	7.20	32.00	8.57	6.86	33.71	18.00
Llama3.1-8b	19.44	6.96	15.97	12.57	21.53	17.50	1.69	12.72	7.20	5.00	21.59	22.32	5.44	24.57	15.83
Llama3.1-70b	23.81	16.80	33.33	16.07	23.33	37.50	3.80	22.00	40.00	12.80	19.50	34.29	11.69	4.00	18.50
Mistral-7b	15.91	3.20	20.29	2.29	14.67	12.00	1.90	10.51	1.80	0.80	16.48	9.09	9.94	2.29	5.50
Mixtral-8x22b	29.17	18.18	30.67	9.94	20.29	33.70	2.90	31.67	16.80	11.67	26.70	33.33	9.77	29.14	18.75
Claude-3.5-sonnet	53.95	13.64	24.64	17.14	23.61	48.44	9.57	40.72	20.80	18.40	48.91	47.43	19.33	36.57	28.12
Gemini-2.0-flash	50.00	14.40	54.67	21.14	25.69	53.00	10.52	34.17	24.80	18.40	40.50	56.00	19.39	46.29	33.00
Deepseek-chat	48.96	28.00	46.00	20.00	26.00	61.00	5.37	48.44	56.80	18.40	47.50	58.86	20.78	53.71	32.29
GPT-4.1	37.50	26.40	51.96	21.14	28.00	60.16	6.96	44.50	60.00	32.00	56.55	62.29	20.71	55.43	36.50
GPT-o4-mini	50.62	22.61	50.72	22.98	18.06	68.48	15.75	53.24	69.17	48.00	44.50	76.00	31.58	64.88	38.50
Average	35.09	15.66	36.09	15.30	20.72	40.99	6.02	31.35	29.90	17.27	35.42	40.82	15.55	35.06	24.50

CE: Causal effect estimation; CD: Cause-effect direction; CI: Contextual interaction; CF: Counterfactual prediction; DS: Domain shift. MC: Mediator-outcome confounding; NS: Necessity and sufficiency; OE: Observational vs experimental; PS: Population shift; SB: Selection bias. SM: Sequential mediators; SP: Simpson’s paradox; SU: Structure uncertainty; TS: Temporal stability; TM: Treatment-mediator interaction.

Workflow for Code-Assisted Prompting with Debugging

1. **Code Generation:** Same as Protocol 2: provide the causal question, dataset location, column names, and a data sample. Request Python code for analysis.
2. **Code Execution:** Extract and run the generated code.
3. **Debugging (if execution fails):** Present the error message to the model and request corrected code. Execute the corrected code.
4. **Result Interpretation:** Show the model its code and the numerical results, and ask it to interpret the results in context.
5. **Output Collection:** Record the model’s interpretation for evaluation.

Table 9 compares causal reliability across all three protocols. Debugging primarily benefits models that frequently fail on the first code attempt. For example, Mistral-7B improves from 7.65% to 17.55% and Llama3.1-8B from 12.86% to 19.04%, recovering to or above their direct-prompting baselines. In contrast, stronger models such as GPT-o4-mini and DeepSeek-chat show mild but

consistent improvements, as their initial code-error rates are already low. This pattern suggests that for weaker models, code-assisted prompting introduces a code-generation failure mode that masks their underlying reasoning ability, whereas stronger models leverage computation to improve causal inference.

Table 9: Causal reliability (%) across three evaluation protocols. Causal reliability is the normalized score averaged across all six pitfall categories.

LLM	Direct	Code-Assisted	Code-Assisted + Debug
Gemma2-9B	13.69	13.24	15.93
Llama3.1-8B	17.86	12.86	19.04
Llama3.1-70B	19.78	19.89	20.42
Mistral-7B	14.43	7.65	17.55
Mixtral-8×22B	16.99	20.59	20.84
Claude-3.5-Sonnet	26.22	29.38	30.17
Gemini-2.0-Flash	24.36	31.80	33.30
DeepSeek-Chat	32.43	36.12	36.72
GPT-4.1	25.24	37.32	37.98
GPT-o4-Mini	40.72	43.03	45.15

I BENCHMARK COVERAGE COMPARISON

Table 10 compares the challenge coverage of CausalPitfalls against four existing benchmarks. Existing evaluations have focused either on semantic causal reasoning (Jin et al., 2023) or on estimation accuracy in standard settings (Liu et al., 2024; Wang, 2024). In contrast, CausalPitfalls targets common statistical pitfalls and failure modes across all six categories.

Table 10: Coverage of CausalPitfalls challenges across existing benchmarks. “Symbolic only” refers to semantic reasoning without tabular data.

Challenge	CLadder	Corr2Cause	CausalBench	QRData	Ours
Simpson’s paradox	Symbolic only	No	No	Partial	Yes
Selection bias (Berkson’s)	Symbolic only	No	No	No	Yes
Observational vs experimental	Symbolic only	No	Yes	Partial	Yes
Causal effect estimation	Symbolic only	No	Yes	Partial	Yes
Counterfactual prediction	Symbolic only	No	Yes	Limited	Yes
Causal necessity & sufficiency	Symbolic only	No	Partial	No	Yes
Mediator–outcome confounding	Symbolic only	No	Partial	No	Yes
Sequential mediators	Symbolic only	No	No	No	Yes
Treatment–mediator interaction	No	No	No	No	Yes
Cause–effect direction	Symbolic only	Symbolic only	Yes	No	Yes
Uncertainty in causal structures	Partial	No	Partial	No	Yes
Population shift & transferability	No	No	No	No	Yes
Temporal stability	No	No	No	No	Yes
Contextual interaction & moderation	No	No	Partial	Partial	Yes
Domain shift & transportability	No	No	No	No	Yes