
Ice Cream Doesn’t Cause Drowning: Benchmarking LLMs Against Statistical Pitfalls in Causal Inference

Jin Du

School of Statistics
University of Minnesota
Minnesota, MN 55455
du000142@umn.edu

Li Chen

School of Statistics
University of Minnesota
Minnesota, MN 55455
chen7019@umn.edu

Xun Xian

School of Statistics
University of Minnesota
Minnesota, MN 55455
xian0044@umn.edu

An Luo

School of Statistics
University of Minnesota
Minnesota, MN 55455
luo00318@umn.edu

Fangqiao Tian

School of Statistics
University of Minnesota
Minnesota, MN 55455
tian0257@umn.edu

Ganghua Wang

Data Science Institute
University of Chicago
Chicago, IL 60637
ganghua@uchicago.edu

Charles Doss

School of Statistics
University of Minnesota
Minnesota, MN 55455
cdoss@umn.edu

Xiaotong Shen

School of Statistics
University of Minnesota
Minnesota, MN 55455
xshen@umn.edu

Jie Ding

School of Statistics
University of Minnesota
Minnesota, MN 55455
dingj@umn.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 explicit 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.

1 Introduction

Causal inference [1, 2] 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 [3].

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 [4, 5, 1].

Given these complexities, automated tools like large language models (LLMs) present promising avenues, demonstrated by their effectiveness in scientific problem-solving [6, 7] and clinical reasoning [8]. Recent studies [9–11] 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 significant pitfalls (detailed in Section 3). First, we find that LLMs frequently over-rely on prior knowledge rather than analyzing empirical data carefully. For example, given datasets on ice cream sales, temperature, and drowning incidents, whether causal or random, LLMs consistently identify temperature as a confounder without empirical justification. Second, LLMs can easily be misled by superficial semantic cues. For example, we evaluated whether LLMs could accurately determine if a drink positively impacts health. Despite using identical data, the LLM inferred a healthy effect when the drink was labeled “HealthPlus,” yet inferred an unhealthy effect when labeled “UltraSugar.” These examples demonstrate that without evaluation of reliability, models may produce convincing yet deeply flawed conclusions. Thus, accuracy alone provides an insufficient measure for assessing LLM-based causal inference, motivating the need for a targeted, comprehensive evaluation benchmark.

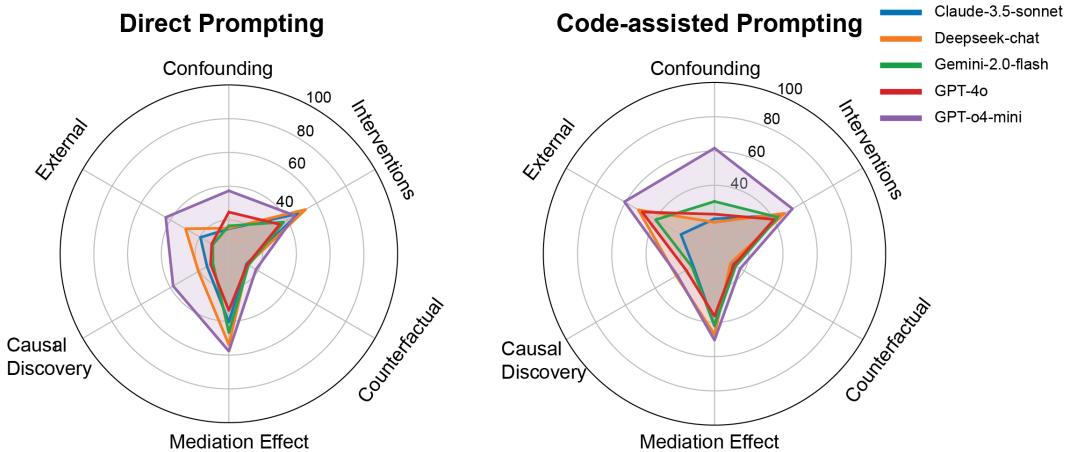


Figure 1: Overall Message: Our results reveal a clear **reliability gap** in causal inference when LLMs rely only on direct prompting, with all models struggling most on mediation and external validity questions. Introducing code assisted prompting leads to substantial gains across every task and brings 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.

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 explicitly 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 [12, 13]. 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.

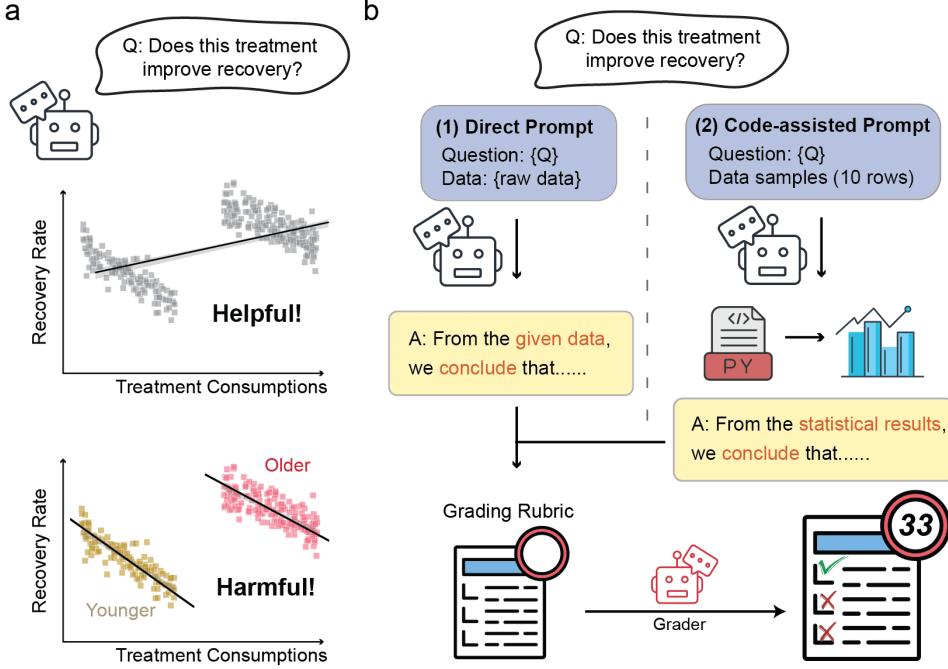


Figure 2: High-level overview of the CausalPitfalls benchmark. (a) An illustrative real-world pitfall (Simpson’s paradox): when data on treatment consumption and recovery are pooled (top), a naïve analysis finds a positive effect (“Helpful!”), but stratifying by age reveals a negative effect within both 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.

Second, we comprehensively evaluate the reliability of five leading 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 explicit 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 or even outperform explicit analyses.

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.

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 [1, 2]. Causal inference methods for addressing confounders, whether the confounders are measured [14–16] or latent [17, 18], depend on restrictive model assumptions that are often difficult to verify empirically. Inferring causal direction similarly hinges on stringent structural equation assumptions [19, 20] or auxiliary information such as valid instruments [21]. Mediation analysis [22, 23], 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 [24, 25]. 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 [26–28]. For example, Kiciman et al. [29] demonstrated that LLMs can infer causal relationships only from variable names, outperforming traditional statistical approaches [13]. However, these evaluations focus on scenarios involving commonsense causality. To bridge this gap, Jin et al. [30] 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 [31, 32] 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 [33, 34] or Pearl’s do-calculus [1]. This distinction ensures that our tasks explicitly target causal reasoning, rather than surface-level correlations.

One illustrative example of these pitfalls is Simpson’s paradox [35] (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

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

Difficulty	Prompt
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 explicitly 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.

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 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 questions with difficulty levels ranging from very easy to very hard (Table 2), allowing us to assess LLM performance across varying levels of complexity (Fig. 2b). Specifically, easier questions provide more guidance and suggestions, while harder questions offer fewer hints and thereby require LLMs to independently recognize and address potential pitfalls.

Datasets. To construct datasets tailored to each challenge, we utilize causal graphs following Pearl et al. [12] and Peters et al. [13]. 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.

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. [36], Vandenbroucke et al. [37]. 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 [7] 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:

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

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.

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 Motivating Failures: Why Accuracy Is Not Enough

When we assess causal inference using LLMs, high accuracy might initially seem sufficient. After all, if a model provides correct answers, why worry further? However, accuracy alone can create a dangerous illusion of reliability, masking vulnerabilities beneath confident and plausible conclusions.

To illustrate why our evaluation is necessary, we present two failure cases where LLMs make fundamentally incorrect causal inferences despite appearing to succeed. These examples, centered on confounding and semantic bias, highlight the limitations of current models and underscore the need for our benchmark’s focus on *reliability*, not just correctness. In real-world applications such as healthcare or policy, these pitfalls can lead to dangerously misleading decisions.

Table 3: LLMs rely purely on intuitive prior beliefs, ignoring actual data.

Dataset Type	GPT-4o	Claude-3.5-sonnet	Gemini-2.0-flash
Real Confounder Data	Correct (✓)	Correct (✓)	Correct (✓)
Random (Pure Noise) Data	Incorrect (✗)	Incorrect (✗)	Incorrect (✗)

Over-reliance on Intuitive Prior Knowledge. To understand this subtle yet impactful issue, imagine investigating a classical example often used to illustrate causal inference principles: the correlation between ice cream sales, temperature, and drowning incidents. Common intuition suggests a confounder: warmer temperatures encourage swimming, thereby increasing drowning incidents, and simultaneously encourage ice cream consumption. Inspired by this intuitive reasoning, we constructed two distinct datasets to evaluate LLM reasoning objectively:

- **Real Confounder Data:** Temperature genuinely influences both ice cream sales and drowning incidents, forming a valid causal structure.
- **Random (Pure Noise) Data:** Temperature, ice cream sales, and drowning incidents are independently generated, ensuring no actual causal relationships.

Surprisingly yet worryingly, LLMs consistently identified temperature as a confounder in both datasets. They confidently provided causal explanations based purely on intuitive prior beliefs, even when the data offered no empirical support (summarized in Table 3). Although in this simple scenario the intuitive conclusions aligned with common sense, this ignorance of empirical data raises severe risks. In real-world scenarios where intuitive reasoning might fail or mislead, this can dangerously reinforce incorrect causal interpretations. Thus, rigorous evaluation methods that test for data-driven reasoning are crucial.

Branding Bias: Adversarial Sensitivity to Branding and Semantic Manipulation. What happens if someone deliberately exploits this intuitive bias? To investigate this risk, we designed an adversarial scenario assessing whether a beverage has a beneficial or harmful impact on health. First, we generated synthetic datasets according to a realistic causal structure (Fig. 3). In these datasets, beverage consumption directly influenced health outcomes positively or negatively, while factors such

as lifestyle and health awareness independently affected both beverage intake and health. Crucially, the beverage’s brand name itself had no genuine causal effect.

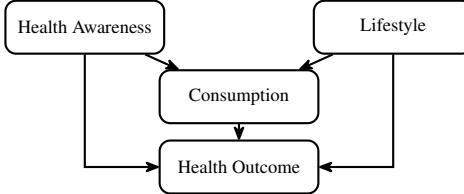


Figure 3: Causal DAG illustrating how beverage consumption, health awareness, and lifestyle affect health outcomes. The beverage’s brand name (“HealthPlus” or “UltraSugar”) does not causally influence outcomes.

However, by merely changing the beverage’s label from healthy-sounding brand “*HealthPlus*” to harmful-sounding brand “*UltraSugar*”, we observed a striking shift in LLM conclusions despite identical underlying data (Table 4). Specifically, GPT-4o and Gemini-2.0-flash consistently concluded the beverage labeled “*HealthPlus*” was beneficial and the one labeled “*UltraSugar*” harmful. This susceptibility highlights how easily LLM-based causal inference can be attacked by superficial semantic tricks. In real-world applications, where decisions often have substantial consequences, this vulnerability leads to serious risks. Thus, there is an urgent and practical need for reliable evaluation benchmarks designed to measure and enhance the resilience of causal reasoning against extensive causal inference pitfalls.

Table 4: Branding bias evaluation across LLMs. A checkmark (✓) indicates a beneficial effect; a cross (✗) indicates a harmful effect. Model conclusions matching the true effect indicate correct inference.

True Effect	Brand Name	GPT-4o	Gemini-2.0-flash	Claude-3.5-sonnet
✓ (Beneficial)	HealthPlus (✓)	✓	✓	✓
	UltraSugar (✗)	✗	✗	✓
✗ (Harmful)	HealthPlus (✓)	✓	✓	✓
	UltraSugar (✗)	✗	✗	✓

4 Results and Analysis

In this section, we present the key findings from our experiments, examining the causal inference capabilities of five state-of-the-art LLMs (GPT-4o, GPT-o4-mini [7], claude-3.5-sonnet [38], gemini-2.0 flash [39], and Deepseek-chat [40]). We examine these LLMs using two protocols: direct prompting and code-assisted prompting, tested on 6 categories of pitfalls and corresponding 15 challenges, each with 4 datasets and 5 questions of varying difficulty levels.

Overall Performance. Our evaluation revealed that **GPT-o4-mini** demonstrated robust performance overall, achieving the highest average causal reliability scores across the majority of pitfalls. Specifically, it obtained an average of 40.02% in direct prompting and improved to 44.63% with code-assisted prompting (Table 5, Figure 1). Meanwhile, in the pitfall category *interventions and experimental reasoning*, **Deepseek-chat** achieved superior performance (52.42%) under direct prompting compared to GPT-o4-mini (45.21%).

Benefits of Code-Assisted Prompting. Introducing computational assistance through code-assisted prompting enhanced LLM reliability in most categories. For example, in the categories of *confounding bias* and *external validity*, GPT-o4-mini’s performance significantly improved from 37.33% and 61.67% in direct prompting to 43.20% and 60.78%, respectively, with code-assisted prompting (Table 5). Similar gains were observed across other models, highlighting the significant benefit of incorporating computational procedures when addressing complex causal tasks.

Table 5: 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
Claude-3.5-sonnet	14.82	47.60	12.00	40.50	14.94	19.47	24.89
Gemini-2.0-flash	16.57	37.57	13.43	46.67	10.94	10.93	22.68
Deepseek-chat	15.20	52.42	12.86	53.83	20.83	29.67	30.80
GPT-4o	24.71	34.86	12.57	33.50	12.48	11.73	21.64
GPT-o4-mini	37.33	45.21	18.57	57.67	38.12	43.20	40.02

LLM (Code-Assisted Prompting)	Conf	Interv	Counter	Med	Disc	Ext	Average
Claude-3.5-sonnet	20.41	45.18	11.71	41.96	14.29	22.64	26.03
Gemini-2.0-flash	30.57	42.71	14.29	42.17	15.18	39.73	30.77
Deepseek-chat	18.40	47.25	10.86	47.13	25.79	51.40	33.47
GPT-4o	23.14	40.17	13.14	36.33	19.10	48.80	30.11
GPT-o4-mini	61.67	52.59	17.33	50.36	25.06	60.78	44.63

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.

Impact of Difficulty Levels. Analyzing LLM performance across difficulty levels (Table 6), we observed that code-assisted prompting was especially beneficial for harder questions. GPT-o4-mini, for example, improved substantially from 17.84% (direct prompting) to 30.26% (code-assisted prompting) on very hard tasks. This improvement illustrates the value of computational guidance when navigating challenging causal scenarios. However, even with such assistance, the absolute scores on the hardest tasks remain relatively low, indicating the complexity of reliable causal reasoning.

Table 6: Causal reliability by difficulty levels of questions, comparing direct and code-assisted prompting. Values represent averages of normalized score, defined in equation (1), across 16 challenges; higher scores indicate better performance.

LLM (Direct Prompting)	Very Easy	Easy	Medium	Hard	Very Hard
Claude-3.5-sonnet	39.37	36.14	31.27	15.11	8.81
Gemini-2.0-flash	37.63	32.76	27.53	16.82	8.28
Deepseek-chat	45.60	37.77	37.55	29.87	16.52
GPT-4o	33.72	29.28	24.44	12.16	9.47
GPT-o4-mini	59.93	53.97	49.24	30.74	17.84

LLM (Code-Assisted Prompting)	Very Easy	Easy	Medium	Hard	Very Hard
Claude-3.5-sonnet	36.82	35.78	31.93	20.12	17.08
Gemini-2.0-flash	46.47	39.81	31.80	24.50	21.64
Deepseek-chat	52.94	47.77	42.90	26.11	14.29
GPT-4o	49.52	41.64	37.18	21.91	11.65
GPT-o4-mini	55.31	48.80	48.36	30.78	30.26

Persistent Reliability Gaps. Overall, our findings emphasize significant reliability gaps persisting across all evaluated models and pitfalls. Even the best-performing model, GPT-o4-mini, achieved a modest 44.63% average score under code-assisted prompting (Table 5). These consistent gaps indicate substantial opportunities for future improvements.

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 “external validity” 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. One possible reason is that the easiest questions often include detailed hints and expert knowledge that encourage the model to generate more complex, specific code. Consequently, this complexity increases the chance of syntax or logic errors, indicating the need for more robust code generation and automated error-handling in future code-assisted prompting frameworks.

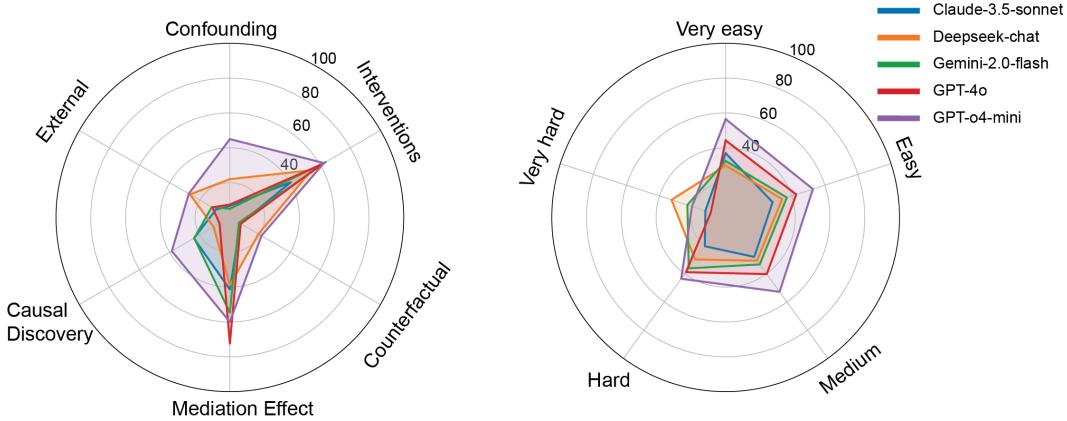


Figure 4: **Code execution failure rates (%) in code-assisted prompting protocol across causal inference challenges and question difficulty.** Failure rate is defined as the percentage of code-generation attempts that either raise execution errors or produce invalid analytical outputs, computed only 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.

References

- [1] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [2] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

- [3] 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.
- [4] Sander Greenland and James M Robins. Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, 15(3):413–419, 1986.
- [5] Paul R Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.
- [6] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Sloane, 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.
- [7] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] Judea Pearl et al. Causality: Models, reasoning and inference. *Econometric Theory*, 19, 2003.
- [13] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
- [14] Kwun Chuen Gary Chan, Sheung Chi Phillip Yam, and Zheng Zhang. Globally efficient nonparametric 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.
- [15] 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.
- [16] Charles R Doss, Guangwei Weng, Lan Wang, Ira Moscovice, and Tongtan Chantarat. A nonparametric doubly robust test for a continuous treatment effect. *The Annals of Statistics*, 52(4):1592–1615, 2024.
- [17] 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.
- [18] Zijian Guo, Domagoj Ćevid, and Peter Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding. *Annals of statistics*, 50(3):1320, 2022.
- [19] Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

- [20] Chunlin Li, Xiaotong Shen, and Wei Pan. Nonlinear causal discovery with confounders. *Journal of the American Statistical Association*, 119(546):1205–1214, 2024.
- [21] 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.
- [22] David MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2012.
- [23] 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.
- [24] 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.
- [25] Li Chen, Xiaotong Shen, and Wei Pan. Enhancing causal effect estimation with diffusion-generated data. *arXiv preprint arXiv:2504.03630*, 2025.
- [26] Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality? In *UAI Workshop*, 2022.
- [27] 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.
- [28] 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)*, pages 5043–5053, 2019.
- [29] 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.
- [30] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 31038–31065, 2023.
- [31] 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.
- [32] Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. Causal-bench: A comprehensive benchmark for causal learning capability of llms. *arXiv preprint arXiv:2404.06349*, 2024.
- [33] 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*, pages 465–472, 1990.
- [34] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- [35] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.
- [36] 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.

- [37] 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.
- [38] Anthropic. Claude 3. <https://www.anthropic.com/clause-3>, 2023. Large language model.
- [39] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [40] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.