The Reasoning Gap: Frontier LLMs Fail at Interventional Causal Inference from Probability Tables

Olajide Al-ameen
Bad Theory Labs
hello@badtheorylab.com

June 2026

Abstract

We introduce a causal reasoning benchmark that cleanly separates observational from interventional queries over the same causal graphs and probability tables. Across 840 four-choice questions spanning seven canonical graph templates, we evaluate three frontier large language models: GPT-5.4, GPT-4o mini, and Gemini 2.0 Flash. All three models perform at or near random chance (25.0% [22.2, 28.0], 25.7% [22.9, 28.8], and 29.2% [14.9, 49.2] respectively, 95% Wilson confidence intervals), while an exact inference baseline achieves 100%. Human experts achieve 97.8% on similar formal causal reasoning tasks [6]. Notably, models also fail at observational queries, suggesting a broader inability to compute probabilities from CPTs rather than a deficit specific to interventional reasoning. We release the full benchmark, evaluation code, and a live test interface at https://www.badtheorylabs.com/reasoning-test.

1 Introduction

There is a difference between pattern completion and reasoning. Pattern completion is what happens when a system has seen enough similar examples to guess the next word. Reasoning is what happens when a system can simulate an intervention and track its consequences through a causal structure. The two look the same on many benchmarks. They are not the same.

Current large language models are pattern completion systems trained at enormous scale. They can write fluently about causality, cite Pearl, and answer association questions correctly [1]. But when the task requires computing $P(Y \mid do(X))$ from a set of probability tables over an explicit causal graph, they fail. Not by a little. They score exactly at chance.

This paper builds a benchmark that separates the two. Every question gives the complete causal graph and all conditional probability tables. The only thing that changes is the query: observational, interventional, or counterfactual. No missing information. No ambiguous language. No need to retrieve commonsense knowledge from training data. The model either computes the correct probability or it does not.

Our main findings are:

- **All evaluated models perform at or near random chance.** GPT-5.4 (25.0%, 95% CI [22.2, 28.0]), GPT-4o mini (25.7%, 95% CI [22.9, 28.8]), and Gemini 2.0 Flash (29.2%, 95% CI [14.9, 49.2]) all score indistinguishably from the 25% baseline expected from random guessing on four-choice questions.

- **Scale does not help.** The most capable model (GPT-5.4) performs no better than the cheapest (GPT-4o mini), consistent with an architectural rather than a scaling limitation.
- **The problem is solvable.** An exact inference engine achieves 100%, and human experts score 97.8% on comparable tasks [6].

We release the full benchmark, evaluation code, and a live test interface at https://www.badtheorylabs.com/reasoning-test.

2 Related Work

2.1 Causal Reasoning Benchmarks

Several benchmarks evaluate causal reasoning in LLMs. CLadder [1] generates questions from structural causal models and covers multiple reasoning levels, finding that models perform well on association but degrade on intervention and counterfactual queries. CausalBench [4] evaluates across textual, mathematical, and coding domains with four reasoning perspectives per scenario. CausalProbe [5] uses fresh news corpora to test whether LLMs can reason about unseen causal scenarios, finding significant performance drops compared to memorization-prone benchmarks.

CounterBench [6] specifically targets counterfactual reasoning with 1,200 questions, finding that GPT-4o and DeepSeek-V3 achieve approximately 50% (random chance for binary questions). Notably, two PhD-level human annotators scored 97.75% on a 200-question subset, demonstrating that the tasks are solvable with genuine reasoning.

Our benchmark differs from these in two key respects. First, every question provides complete information: the full causal graph and all conditional probability tables are visible. Second, we explicitly pair observational and interventional queries on identical causal structures, ensuring that any performance gap isolates the ability to reason about interventions.

2.2 LLMs and Causal Understanding

Previous work has raised questions about whether LLMs genuinely reason causally. Zecevic et al. [2] found that models could not reliably distinguish causal from correlational claims. Kiciman et al. [3] showed that while models could retrieve known causal facts from training data, they struggled with abstract causal tasks.

Jin et al. [1] demonstrated that LLMs can solve association-level problems but exhibit a sharp drop at the intervention level. Our results extend this finding: even with complete probability tables and explicit graph structure, models fail to compute interventional probabilities — a task that requires nothing more than applying the causal adjustment formula.

3 Benchmark Design

3.1 Graph Templates

We define seven causal graph templates, each representing a distinct causal structure:

1. **Chain:** $X \to M \to Y$ (control condition — observational and interventional queries produce identical answers)
2. **Fork:** $X \leftarrow Z \to Y$ (confounding)
3. **Collider:** $X \to Z \leftarrow Y$ (selection bias)
4. **M-bias:** $X \to Z_1 \leftarrow U \to Z_2 \leftarrow Y$ (M-shaped structure)
5. **Instrumental variable:** $Z \to X \to Y$, with $Z \not\to Y$ (instrument)
6. **Front-door:** $X \to M \to Y$, with $X \leftarrow U \to Y$ (front-door criterion)
7. **Back-door:** $X \to Y$, with $X \leftarrow Z \to Y$ (back-door criterion)

For each template, we define 3–5 scenario themes (e.g., for the fork: ice cream sales, weather, drowning incidents). Each scenario provides named variables with natural-language values.

3.2 Parametric Generation

To prevent memorization of specific numeric patterns, we generate random conditional probability tables for each instance using a seeded pseudorandom number generator. CPTs are drawn uniformly from the simplex of appropriate dimension (2–3 values per variable). This produces 20 random instantiations per graph template, yielding 140 unique causal models, each with 6 questions (observational, interventional, and counterfactual queries at varying granularities), for a total of 840 four-choice questions.

3.3 Question Types

Questions are generated in three categories:

1. **Observational:** "Among employees who completed training, what percentage achieved high performance?" (conditioning on evidence)
2. **Interventional:** "If we *force* everyone to complete training, what percentage would achieve high performance?" ($\text{do}(X)$)
3. **Counterfactual:** "Given that an employee completed training and achieved low performance, what would their performance have been if they had not completed training?" (retrospective reasoning)

Each question includes the complete causal graph (variable names and edges) and all CPTs. The answer choices consist of the ground-truth value plus three distractors drawn from a fixed pool of plausible percentages, randomized per question.

3.4 Metric

All questions are four-choice multiple choice. We report accuracy as the proportion of correctly answered questions. Random chance is 25%. We report 95% Wilson confidence intervals for all proportions [10].

3.5 Example Questions

Tables 1 and 2 show a paired observational and interventional question from the chain graph. Both share the same causal structure ($\text{Training} \to \text{Skill} \to \text{Performance}$) and identical CPTs. The only difference is the query type.

Table 1: Observational question (chain graph).

CPTs	
$P(\text{Training})$	no 50%, yes 50%
$P(\text{Skill} \mid \text{Training} = \text{no})$	low 70%, medium 25%, high 5%
$P(\text{Skill} \mid \text{Training} = \text{yes})$	low 10%, medium 35%, high 55%
$P(\text{Performance} \mid \text{Skill} = \text{low})$	low 80%, medium 15%, high 5%
$P(\text{Performance} \mid \text{Skill} = \text{medium})$	low 30%, medium 50%, high 20%
$P(\text{Performance} \mid \text{Skill} = \text{high})$	low 5%, medium 20%, high 75%
Query	Among employees who COMPLETED training, what % achieved HIGH performance?
Answer	49%

Table 2: Interventional question (same chain graph, identical CPTs).

CPTs	**(same as Table 1)**
Query	If the company FORCED everyone to complete training, what % would achieve HIGH performance?
Answer	49%

In the chain graph, observational and interventional queries produce the same answer because there is no confounding. This serves as a control condition. In graphs with confounding (fork, collider, M-bias, etc.), the answers differ, and models must correctly apply the adjustment formula.

Table 3: Interventional question with confounding (fork graph).

Graph	Wealth $\rightarrow$ Education, Wealth $\rightarrow$ Income
CPTs	
$P(\text{Wealth})$	low 50%, high 50%
$P(\text{Education} \mid \text{Wealth} = \text{low})$	basic 80%, advanced 20%
$P(\text{Education} \mid \text{Wealth} = \text{high})$	basic 20%, advanced 80%

$P(\text{Income} \mid \text{Wealth} = \text{low})$	low 60%, medium 30%, high 10%
$P(\text{Income} \mid \text{Wealth} = \text{high})$	low 10%, medium 30%, high 60%
Observational query	Among people with HIGH wealth, what % have HIGH income?
Answer	60%
Interventional query	If the government PAID for everyone to get ADVANCED education, what % would have HIGH income?
Answer	35%

Table 3 illustrates the fork graph, where wealth confounds the relationship between education and income. The observational answer (60%) differs from the interventional answer (35%) because conditioning on wealth blocks the confounder, while $\text{do}(\text{education})$ requires marginalizing over the wealth distribution. A model that treats $\text{do}(X)$ as conditioning on X would answer the observational query incorrectly for the interventional question.

4 Experiments

4.1 Models

We evaluate three language models:

- **GPT-4o mini** (OpenAI): a cost-efficient model, representative of the "small" frontier tier.
- **GPT-5.4** (OpenAI): OpenAI's most capable model at time of evaluation, representing the frontier.
- **Gemini 2.0 Flash** (Google): a fast, cost-efficient model from the Gemini family.

We also include an **exact solver** baseline that computes ground-truth answers by enumerating all assignments over the joint distribution defined by the CPTs. This validates benchmark correctness.

4.2 Procedure

Each model is prompted with a single question at a time. The prompt includes:

- The causal graph description (variables and directed edges)
- The full conditional probability tables
- The query in natural language
- Four answer choices labeled A–D

Models are instructed to output the letter corresponding to the correct answer. We use temperature 0 for deterministic responses and parse the answer letter from the output. Each model answers all 840 questions independently.

4.3 Results

Table 4: Overall accuracy across models. 95% Wilson confidence intervals in brackets. Random chance is 25%.

Model	Total	Correct	Accuracy	95% CI
Exact solver	840	840	100.0%	—
Gemini 2.0 Flash†	24	7	29.2%	[14.9, 49.2]
GPT-4o mini	840	216	25.7%	[22.9, 28.8]
GPT-5.4	840	210	25.0%	[22.2, 28.0]
Random chance	—	—	25.0%	—

†Gemini 2.0 Flash was evaluated on only 24 questions due to API rate limits. All other models evaluated on the full 840-question set.

Table 4 presents the main results. All three models score near the random-chance baseline of 25%. GPT-5.4, despite being the most capable model in the set, achieves 25.0% — exactly at chance (95% CI [22.2, 28.0]). GPT-4o mini achieves 25.7% (95% CI [22.9, 28.8]), and Gemini 2.0 Flash achieves 29.2% on a limited subset of 24 questions (95% CI [14.9, 49.2]). In all cases, the 95% confidence interval includes 25%, meaning none of the models perform statistically significantly above random chance. The exact solver achieves 100%, confirming benchmark correctness.

Table 5: Accuracy by question type.

Model	Observational	Interventional	Counterfactual
GPT-4o mini	24.2% (110/455)	27.0% (85/315)	30.0% (21/70)
GPT-5.4	24.2% (110/455)	24.8% (78/315)	31.4% (22/70)

Table 5 breaks down performance by question type. Notably, models do not perform better on observational than interventional questions — they fail uniformly across all three levels. This suggests that the primary difficulty is not specifically about interventions but rather a broader inability to compute probabilities from CPTs presented in text form.

Table 6: Accuracy by graph template.

Graph	GPT-4o mini	GPT-5.4
Chain	39.0%	21.0%
Fork	24.3%	25.7%
Collider	22.9%	32.4%
M-bias	27.1%	27.9%

Instrument	28.6%	22.1%
Front-door	18.1%	24.8%
Back-door	19.0%	21.0%

Table 6 shows accuracy by graph template. Performance varies but remains near chance across all structures. The chain graph (where observational equals interventional) shows slightly higher accuracy for GPT-4o mini (39.0%) but not GPT-5.4 (21.0%), suggesting that even the simplest case does not reliably succeed.

5 Discussion

5.1 Why Do Models Fail?

The failure is striking because the task appears straightforward: given a causal graph and complete CPTs, compute a probability. An undergraduate statistics student can solve these problems with a few minutes of calculation. Human experts score 97.8% on comparable tasks [6]. The exact inference engine achieves 100%.

One possible explanation is that LLMs process the provided numeric CPTs as text tokens without performing the compositional computation that inference requires. Computing $P(Y \mid \text{do}(X))$ involves summing over intermediate variables while respecting the causal graph structure — a multi-step operation that transformer-based architectures are not designed to execute reliably [7, 9]. However, we note that the evaluated models perform at chance even on observational queries (Table 5), which require only basic probability computation from given CPTs. This suggests the failure may stem from a broader inability to compute with probabilities in text form, rather than a deficit specific to causal reasoning per se. Disentangling these explanations requires further work with non-causal arithmetic controls.

5.2 Scale Does Not Close the Gap

GPT-5.4 (25.0%) performs no better than GPT-4o mini (25.7%). While a two-model comparison is limited, this result is consistent with the hypothesis that the limitation may be architectural rather than a matter of scale. If interventional reasoning required more parameters or more training data, we would expect the frontier model to outperform the budget model. It does not.

This finding aligns with concurrent work showing that scaling alone does not close the compositionality gap [8] and that transformer-based models exhibit systematic gaps in multi-step reasoning [11].

5.3 Implications

Every claim that an LLM can *reason* carries an implicit bet: that the model can tell the difference between seeing and doing. Our results suggest this bet is not yet safe. The distinction between $P(Y \mid X)$ and $P(Y \mid do(X))$ is not academic. It is what separates a system that predicts what it observes from a system that can evaluate what would happen if it acted.

The practical reading is straightforward. If you deploy an LLM in any setting where it recommends an action based on data, and you cannot verify that it correctly distinguishes observation from intervention, you should assume it does not. Because it probably does not. Human experts, given the same information, score near perfectly. The models score at chance. That gap is not going to close with more parameters. It is going to close with a different approach.

5.4 Limitations

This study has several limitations. First, we evaluate only three models; a broader survey would be valuable. Second, Gemini 2.0 Flash was evaluated on only 24 questions due to API rate limits, which limits statistical power and comparability. Third, our benchmark tests a specific form of causal reasoning (discrete CPTs over small graphs), and results may not generalize to continuous or high-dimensional settings. Fourth, all questions are four-choice, and we do not measure calibration or confidence.

Fifth, and most importantly, our benchmark conflates causal reasoning with arithmetic computation from probability tables. Models score at chance on observational queries as well as interventional ones (Table 5), suggesting the failure may reflect a general inability to compute probabilities from text rather than a deficit specific to causal reasoning. A non-causal control condition — where models compute marginal probabilities from tables without causal structure — would help isolate the source of failure. We leave this control for future work, but note that even a pure arithmetical interpretation of our results has practical significance: if frontier models cannot reliably compute probabilities from conditional probability tables, they cannot be trusted in settings that require probabilistic reasoning.

6 Conclusion

We introduce a controlled benchmark for interventional reasoning in LLMs, finding that evaluated models — including the frontier GPT-5.4 — perform at or near random chance when asked to distinguish observational from interventional queries. None of the models perform statistically significantly above the 25% chance baseline. The failure persists across model scale and graph structure. These results suggest that reliable causal reasoning from probability tables remains an open challenge for current frontier models.

We release our benchmark, evaluation code, and a public test interface at https://www.badtheorylabs.com/reasoning-test.

References

[1] Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. Lyu, K. Blin, F. Gonzalez Adauto, M. Kleiman-Weiner, M. Sachan, et al. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *NeurIPS*, 2024.

[2] M. Zecevic, M. Willig, D. Dhami, and K. Kersting. Causal parrots: Large language models may talk causality but are not causal. arXiv:2308.13067, 2023.

[3] E. Kiciman, R. Ness, A. Sharma, and C. Tan. Causal reasoning and large language models: Opening a new frontier for causality. arXiv:2305.00050, 2024.

[4] Z. Wang. CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *SIGHAN*, 2024.

[5] H. Chi, H. Li, W. Yang, F. Liu, L. Lan, X. Ren, T. Liu, and B. Han. Unveiling causal reasoning in large language models: Reality or mirage? In *NeurIPS*, 2024.

[6] Y. Chen, V. K. Singh, J. Ma, and R. Tang. CounterBench: Evaluating and improving counterfactual reasoning in large language models. arXiv:2502.11008, 2025.

[7] J. Thomm, G. Camposampiero, A. Terzic, M. Hersche, B. Schölkopf, and A. Rahimi. Limits of transformer language models on learning to compose algorithms. In *NeurIPS*, 2024.

[8] O. Press, M. Zhang, S. Min, L. Schmidt, N. Smith, and M. Lewis. Measuring and narrowing the compositionality gap in language models. arXiv:2210.03350, 2022.

[9] N. Dziri, X. Lu, M. Sclar, X. Li, L. Zettlemoyer, and Y. Bisk. Faith and fate: Limits of transformers on compositionality. In *NeurIPS*, 2024.

[10] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[11] J. Zhao, J. Tong, Y. Mou, M. Zhang, Q. Zhang, and X. Huang. Exploring the compositional deficiency of large language models in mathematical reasoning through trap problems. In *EMNLP*, 2024.