

Choosing Experiments for Causal Prediction: An Exact Bayes-Risk Benchmark of Open Language Models

Arya Somu

October 2026

Abstract

Observationally identical causal models can disagree about the effect of an intervention. A useful experimental-design agent must choose an intervention that resolves uncertainty about a particular future causal query, rather than merely name a plausible experiment. We construct finite structural causal model tasks for which every candidate causal order reproduces the same observational distribution, while interventions have distinct values for a held-out query. Exact enumeration gives the prior-averaged one-step Bayes risk of each action. We prove that the expected reduction in this risk equals the mutual information between the experiment outcome and the future query outcome. This yields a deterministic, auditable action score without realized-outcome noise.

Across successive locked studies, unaided Qwen2.5-32B, Qwen2.5-72B, and Qwen3-235B conditions did not meet their registered better-than-random action-selection rules. In the v4 randomized-label test, the 235B thinking condition chose an optimal action in 102 of 300 informative cases; its risk exceeded uniform random by 0.00151 nats with a 95% interval spanning zero. A 32B selection control given exact public-input action risks chose the optimum in all 300 cases. A fresh v5 diagnostic finds that a 32B calculator-using agent reduces risk by 0.02536 nats relative to its raw-stage counterpart on 150 new cases, although it never queried one of the three actions. This tool-assisted contrast is descriptive and does not establish unaided causal calculation. The results locate a substantial gap between access to a finite causal specification and effective numerical experiment choice, within a deliberately narrow synthetic setting.

1 Introduction

Interventions separate causal models that can be indistinguishable under observation [1, 2]. Recent work evaluates language models on interventional reasoning [3] and on active causal discovery [4]. Here the objective is narrower: the agent is given a finite list of candidate causal worlds and chooses one experiment to improve prediction of a *specified* untried intervention. This closed-world construction removes graph-search and observational-estimation error. It makes the expected value of every experiment exactly computable.

The central empirical question is whether a model uses the supplied causal specification to select low-risk actions. We progressively remove confounds: realized forecast noise, semantic action defaults, and inability to select a minimum when values are supplied. The control that provides exact action risks deliberately supplies the computational answer. It tests selection and interface compliance, not unaided causal reasoning.

2 Finite causal prediction problem

Let $V = (X, Y, Z)$ be binary and let $p(v) > 0$ be a joint observational law. Candidate world w specifies a total order π_w and the complete DAG in which each variable has all preceding variables as parents. Conditional probability tables are derived from the same p :

$$p_w(v) = \prod_{j=1}^3 p(v_{\pi_w, j} \mid v_{\pi_w, 1}, \dots, v_{\pi_w, j-1}) = p(v). \quad (1)$$

Thus the worlds agree on *every* observational cell. They share the full joint distribution, not only a pattern of conditional independence.

For action $a = \text{do}(V_k = c)$, its interventional distribution follows truncated factorization:

$$K_a(v \mid w) = \mathbf{1}\{v_k = c\} \prod_{\pi_{w,j} \neq k} p(v_{\pi_{w,j}} \mid v_{\pi_{w,<j}}). \quad (2)$$

The intervention deletes the target’s conditional factor while leaving all others intact. Since the factors depend on order, $K_a(\cdot \mid w)$ can differ even though (1) does not. The model observes p , the candidate orders, a uniform prior $b(w)$, the allowed actions, and the query. It does not observe which world generated the episode.

The held-out query is $Q = \Pr(Z = 1 \mid \text{do}(X = 1))$. Write q_w for its exact value in world w , and let T be a hypothetical Bernoulli draw with $\Pr(T = 1 \mid W = w) = q_w$. A chosen experiment a returns a complete assignment O_a from $K_a(\cdot \mid W)$. We treat T and O_a as conditionally independent draws given W ; no same-unit counterfactual coupling is assumed. This convention is essential to the information identity below.

2.1 Exact posterior and query risk

For outcome o with positive predictive mass,

$$m_a(o) = \sum_w b(w) K_a(o \mid w), \quad b_{a,o}(w) = \frac{b(w) K_a(o \mid w)}{m_a(o)}. \quad (3)$$

The posterior predictive probability is $\bar{q}_b = \sum_w b(w) q_w$. Under Bernoulli log loss, the expected excess loss of this prediction relative to knowing W is

$$\begin{aligned} \mathcal{R}(b) &= \sum_w b(w) D_{\text{KL}}(\text{Bern}(q_w) \parallel \text{Bern}(\bar{q}_b)) \\ &= H(\bar{q}_b) - \sum_w b(w) H(q_w) = I(W; T \mid b), \end{aligned} \quad (4)$$

where $H(u) = -u \log u - (1 - u) \log(1 - u)$ in nats. The expected one-step action risk is

$$R_a(b) = \sum_{o: m_a(o) > 0} m_a(o) \mathcal{R}(b_{a,o}), \quad R_{\text{stop}}(b) = \mathcal{R}(b). \quad (5)$$

All terms are finite sums over at most three candidate worlds and eight assignments. The exact policy minimizes $\{R_{\text{stop}}, R_a : a \in \mathcal{A}\}$. The random baseline is the mean of R_a over the three allowed actions.

Proposition 1 (Query-targeted information value). *Under the conditional independence $T \perp O_a \mid W$, the expected reduction in query risk is*

$$\mathcal{R}(b) - R_a(b) = I(T; O_a). \quad (6)$$

Consequently $0 \leq R_a(b) \leq \mathcal{R}(b) \leq \log 2$.

Proof. Equation (4) gives $\mathcal{R}(b) = I(W; T)$ and (5) gives $R_a(b) = I(W; T \mid O_a)$. The chain rule implies

$$I(W; T) - I(W; T \mid O_a) = I(T; O_a) - I(T; O_a \mid W).$$

The last term vanishes by conditional independence. Nonnegativity of mutual information gives $R_a \leq \mathcal{R}(b)$; conditional mutual information gives $R_a \geq 0$; and $I(W; T) \leq H(T) \leq \log 2$. $\square$

This is a *query-targeted* design criterion. Maximizing $I(W; O_a)$, information about the entire causal world, can select a different action. For example, let $W = (U, V)$ contain two independent fair bits and let $T = U$. An action revealing V perfectly has $I(W; O) = \log 2$ but $I(T; O) = 0$. An action observing U through a binary error channel with error probability $\epsilon \in (0, 1/2)$ has less world information, $\log 2 - H(\epsilon)$, yet positive query value of the same amount. The benchmark scores the latter objective.

We independently evaluated both sides of (6) for every action in the v3, v4, and v5 locked test manifests. The maximum absolute numerical discrepancy was 3.11×10^{-16} nats over 3,150 action checks. This verifies the implementation’s direct Bayes-risk enumeration against a separately computed T, O_a joint table; it is not a statistical result about the models.

3 Benchmark construction and evaluation

The common observational law is a positive three-bit log-linear distribution. For each assignment $v \in \{0, 1\}^3$,

$$p_\theta(v) = \frac{\exp\{\sum_{\emptyset \neq S \subseteq \{1,2,3\}} \theta_S \prod_{i \in S} (2v_i - 1)\}}{\sum_{u \in \{0,1\}^3} \exp\{\sum_{\emptyset \neq S} \theta_S \prod_{i \in S} (2u_i - 1)\}}. \quad (7)$$

Seven coefficients determine the eight positive cell probabilities. Each episode samples two or three distinct total orders, uses the same p_θ for all, and offers $\text{do}(X = 0)$, $\text{do}(Y = 0)$, and $\text{do}(Y = 1)$. The held-out intervention $\text{do}(X = 1)$ is forbidden as an experiment. Rejection gates require a meaningful range of query values and a minimum random-to-oracle opportunity. Independent joint-parameter draws define development and test cases, with no parameter overlap. The informative strata balance the optimal semantic action before any model call.

For case i and policy f , define

$$d_i(f) = R_{f(i),i} - \frac{1}{3} \sum_{a \in \mathcal{A}} R_{a,i}, \quad \delta(f) = \frac{1}{n} \sum_{i=1}^n d_i(f). \quad (8)$$

Negative δ favors the model. The opportunity gap is $G = n^{-1} \sum_i (R_{\text{random},i} - R_{\text{oracle},i})$. Intervals use 2,000 bootstrap resamples of independent parameter draws. This is uncertainty over the benchmark’s case distribution, conditional on each saved model response; it does not include repeated decoding seeds or model retraining. An interval that spans zero does not prove equality. A positive interval reports worse action choice on this sampled task distribution, not a general impossibility theorem.

4 Completed model results

4.1 Earlier locked studies

The initial 32B factorial study used end-to-end submitted forecasts and had only a 0.00057-nat realized random-to-oracle gap, making action discrimination weak despite a large model-minus-random forecast loss. The v2 study improved opportunity but still used realized outcomes. V3 moved to the exact expected risk (5), eliminating that outcome noise. On 300 informative v3 cases, unaided Qwen2.5-32B, Qwen2.5-72B, and Qwen3-235B each selected $\text{do}(Y = 0)$ in all 450 test cases, including 150 near-tie controls. Each was optimal in 100 of 300 informative cases and had model-minus-random risk +0.000172 nats with an interval crossing zero. Two Qwen2.5 selection controls supplied exact public-input action values and chose the optimum in 299 of 300 cases.

V4 randomized the mapping from semantic actions to presented A/B/C identifiers. Its 300 informative cases had mean random risk 0.129469, oracle risk 0.102272, and opportunity gap 0.027196 nats. Table 1 reports all five complete conditions. Only the 235B thinking model versus random was a registered confirmatory comparison; its interval spans zero. The other intervals and condition contrasts are descriptive.

V4 condition	δ (nats)	95% interval	Optimal / 300
32B, literal IDs, raw SCM	−0.002666	[−0.006155, 0.000751]	100
32B, opaque IDs, raw SCM	+0.001551	[−0.001237, 0.004396]	102
32B, opaque IDs, public likelihoods	+0.000717	[−0.002222, 0.003753]	109
32B, opaque IDs, exact action risks	−0.027196	[−0.030334, −0.024161]	300
235B thinking, opaque IDs, raw SCM	+0.001509	[−0.001350, 0.004370]	102

Table 1: Locked v4 informative cases. Each condition covers the same 300 cases; 150 near-tie controls are excluded from δ . Lower risk is better. The exact-risk condition supplies the values needed for selection.

The two raw opaque-ID models agreed on 298 of 300 informative choices. The 32B model selected B or C in 212 and 88 informative cases; the 235B model selected A, B, or C in 1, 210, and 89. Both chose an optimal action in 102 cases. Their similar choices despite different checkpoints are a behavioral observation, not an isolated effect of size or thinking. A post hoc analysis found that the 235B model chose B in 71 of 101 cases where B was optimal and 139 of 199 where it was not. Because this label was singled out after seeing responses, that analysis is exploratory. The saved thinking traces do not identify the first incorrect numerical step.

4.2 Fresh calculator diagnostic

V5 freezes 150 new informative cases, disjoint from v4, with opportunity gap $G = 0.027566$ nats and optimal presented labels A/B/C in 57/51/42 cases. Its registered confirmatory comparison is a longer-versus-shorter 235B thinking budget; a 32B raw-stage versus calculator-agent contrast is descriptive. The raw-stage condition reports per-world query probabilities, a probe likelihood, a posterior, all three action risks, and an action. The calculator agent may request exact risk for at most three named actions and then chooses one.

Before v5 test inference, the 18-case 32B development pilot revealed that 10 raw-stage JSON responses hit a 256-token output cap and were incomplete. A recorded pretest execution amendment raised only that direct JSON cap to 512. The amended development run had no incomplete responses. The frozen case set, prompts, scoring, and paired decision rule were unchanged. Table 2 gives the complete, independently replayed 32B test results. These do not complete the five-condition v5 study or decide its 235B confirmatory criterion.

V5 condition	Mean risk	δ	95% interval	Optimal / 150
32B, raw numerical stages	0.141043	+0.015592	[0.012604, 0.018433]	7
32B, exact-risk calculator agent	0.115679	-0.009773	[-0.015194, -0.004600]	76

Table 2: V5 32B diagnostic on the same 150 new cases. Uniform-random mean risk is 0.125452 nats and exact-oracle mean risk is 0.097885. The within-case calculator-minus-raw difference is -0.025365 nats, with descriptive paired interval $[-0.032520, -0.018185]$.

The raw-stage model chose STOP in 134 of 150 cases. Its mean absolute error was 0.411 for per-world query probabilities, 0.287 for the prespecified probe likelihood, 0.213 for the probe posterior, and 0.538 nats for reported action risks among 148 numerically valid risk objects. These errors are separately scored but are not independent causal mechanisms: an early error can propagate. The calculator agent chose an action in every case, requested A and B exactly once in every case, never requested C, and selected C in 102 cases. It therefore gained substantially from access to exact risk values while leaving an unevaluated-action bias in the workflow. Supplying a calculator changes the system being tested.

An exact policy restricted to the two values the agent actually requested would have mean risk 0.112236 nats, below the agent’s 0.115679. This post hoc calculation shows that the system did not consistently take the lower of even its observed A/B values. It is a diagnostic of incomplete tool use, not an additional registered model comparison.

The pinned 235B checkpoint loaded on two B200 GPUs, but an eight-case short-budget *development* batch took 199 seconds of generation after loading. The full 150-case short and long conditions could not be budgeted reliably under the new \$30 RunPod cap. We made no v5 235B test call and deleted the pod. Therefore v5 has no result for its registered confirmatory criterion; its completed contribution is the 32B diagnostic above.

5 Reproducibility and limitations

Each study has a frozen manifest binding generation settings, episode records, baseline audit, prompts, analysis, and source hashes. Raw model outputs, token counts, selected actions, tool transcripts, and per-case exact scores are retained. Independent verifiers rebuild public prompts and risk tables, parse saved responses, and check episode coverage. Frozen source bytes are archived for versions whose working tree later changed. GPU inference used pinned public model revisions and vLLM 0.10.2. The release repository contains the paper source, code, manifests, verification scripts, and public result ledgers. The v5 235B test conditions remain unrun.

The public source and evidence repository is <https://github.com/arpjw/causal-interventions>. The paper is available under Creative Commons Attribution 4.0 International (CC BY 4.0); code is available under the MIT License. The repository README gives a CPU-only replay path for the saved results and distinguishes that audit from new model inference.

These are synthetic three-variable tasks with a known candidate set, a uniform prior, one held-out query form, and one action. The generator conditions on informative opportunity and balanced semantic optima, so results do not estimate performance on a natural distribution of causal problems. The v4 controls and v5 calculator give derived numerical values from the public specification; success with them is not evidence that a model computes those values unaided. The v3 and v4 model comparisons mix architecture, training, quantization, decoding, and output budget. Opaque labels limit a semantic default but can introduce presentation preferences of their

own. Test-case bootstrap intervals do not capture decoding variability. We did not test generalization to larger graphs, uncertain observational estimates, costly interventions, partially observed outcomes, or real-world causal decision utility.

The mathematical identity (6) clarifies the target: a good action predicts the *query outcome*, even when another action reveals more about the whole graph. Empirically, the verified negative unaided results and large assisted contrast motivate smaller, stage-level evaluations that can locate numerical failures. They do not justify a general claim that open language models cannot reason causally.

References

- [1] J. Pearl. A causal calculus for statistical research. *Proceedings of Machine Learning Research*, R0:430–449, 1995. <https://proceedings.mlr.press/r0/pearl195a.html>.
- [2] A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012. <https://www.jmlr.org/papers/v13/hauser12a.html>.
- [3] T. Kasetty, D. Mahajan, G. K. Dziugaite, A. Drouin, and D. Sridhar. Evaluating interventional reasoning capabilities of large language models. *arXiv:2404.05545*, 2024. <https://arxiv.org/abs/2404.05545>.
- [4] S. Deb, D. Shah, and A. Krishnan. Active causal discovery benchmark: Evaluating LLM agents under budgeted interventions. *arXiv:2609.31675*, 2026. <https://arxiv.org/abs/2609.31675>.

A Direct finite-table audit formulas

The proof of Proposition 1 uses the standard chain rule, but the equality can also be checked without a belief-update implementation. For $t \in \{0, 1\}$, define $s_w(t) = q_w^t(1 - q_w)^{1-t}$ and construct the joint table

$$J_a(t, o) = \sum_w b(w) s_w(t) K_a(o | w). \quad (9)$$

Its marginals are $J_a(t) = \sum_o J_a(t, o)$ and $J_a(o) = \sum_t J_a(t, o)$. Then

$$I(T; O_a) = \sum_{t, o: J_a(t, o) > 0} J_a(t, o) \log \frac{J_a(t, o)}{J_a(t) J_a(o)}. \quad (10)$$

The release checker computes (10) from the stored public joint and causal orders and compares it with the difference between (4) and (5). It checked 1,350 actions in v3, 1,350 in v4, and 450 in v5, with maximum absolute error 3.11×10^{-16} nats. This catches mistakes in outcome ordering, posterior normalization, or target-factor truncation without using a model response.

Proposition 2 (Risk-estimation margin). *Let $\mathcal{A}_0$ include STOP and the allowed interventions. If estimates $\hat{R}_a$ obey $\max_{a \in \mathcal{A}_0} |R_a - \hat{R}_a| \leq \varepsilon$, and $\hat{a}$ minimizes $\hat{R}_a$, then*

$$0 \leq R_{\hat{a}} - \min_{a \in \mathcal{A}_0} R_a \leq 2\varepsilon. \quad (11)$$

In particular, if the true best action beats every other action by more than 2ε , it is also the unique estimated best action.

Proof. Let a^* minimize true risk. Adding and subtracting the two estimates gives

$$R_{\hat{a}} - R_{a^*} = (R_{\hat{a}} - \hat{R}_{\hat{a}}) + (\hat{R}_{\hat{a}} - \hat{R}_{a^*}) + (\hat{R}_{a^*} - R_{a^*}).$$

The middle term is nonpositive by the choice of $\hat{a}$; the others are each at most ε . For the margin claim, a competing action with true excess risk above 2ε remains worse under all permitted estimation errors. $\square$

The bound explains why scoring all reported action risks is useful: a final choice can be correct by luck even when its numerical estimates are poor. The v5 reported mean absolute stage errors do not imply a uniform ε bound, so (11) is a structural guarantee, not a measured performance certificate for these model outputs.