WHEN OBSERVATIONAL CONTEXT OVERRIDES INTERVENTIONAL EVIDENCE IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a transformer use interventional evidence when its context also contains conflicting observations? We study this question in synthetic structural causal models where observational association and the interventional effect have opposite signs, training decoder-only transformers from scratch on symbolic records while varying the training mixture and the context composition separately. We first formalize what each record type identifies: interventional probes determine the queried effect, observational records do not, and an ideal reader can still be harmed by observations on worlds where the two disagree. Against this reference, the composition of the context strongly shapes the answer. On 50 positive-effect evaluation worlds, a probe-only model produces 41 correct-sign slopes and 4 reversals, compared with 18 and 19 for a separately trained mixed-context model. At a single fixed checkpoint, adding observational records while keeping four probes reduces correct-sign slopes from 31 to 17 on 37 commonly parsed worlds and raises reversals from 1 to 11, an outcome that an ideal reader under the positive-effect training prior avoids, and replacing the observational values with zeros yields 25 correct-sign slopes of 33, compared with 16 for the world's own values. Activation patching shows that representations at the observational-record positions are sufficient to carry much of this effect in the tested subsets. The models also exhibit a positive-sign prior: on binary causal graphs, both models reverse on most negative-effect worlds, and sign-balanced retraining improves in-distribution sign reading for the probe-only model while leaving the out-of-distribution positive offset in place. Evaluations of causal learning should therefore vary inference-time evidence composition, not only the training mixture.

1 INTRODUCTION

Observational and interventional records answer different questions. An association between X and Y describes how they co-vary, whereas an intervention asks how Y changes when X is externally set (Pearl, 2009). With hidden confounding, even the signs of these relationships can disagree. For a model receiving both types of evidence, successful prediction therefore requires more than reading numerical examples: it must use the evidence appropriate to the query.

Transformers can learn estimators from synthetic training tasks, including function interpolation and causal effect estimation (Garg et al., 2022; Akyürek et al., 2023; Robertson et al., 2025; Balazadeh et al., 2025). This motivates a complementary question: *when interventional examples are available, does observational context interfere with their use?* We investigate this question by independently varying the share of interventional training examples and the evidence records supplied in context.

Two distinctions are essential. First, causal effects are not generally identifiable from observations under unrestricted hidden confounding. A failure with observation-only context cannot, by itself, demonstrate a failure to use available causal information. Conversely, under a restricted prior in which effects are always positive, the sign can be predicted without reading any evidence. Second, comparing models trained on different context formats conflates learning and inference. We therefore distinguish comparisons across checkpoints from interventions on a fixed checkpoint, and interpret positive-sign performance alongside a sign-transfer audit.

Our testbed uses synthetic worlds with a designated confounded pair: observational association is negative while the endpoint interventional effect is positive. We train 25.7M-parameter decoder-only

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

transformers on symbolic numerical records. Interventional probes set the query variable and reveal sampled responses; observational records are sampled without intervention. Predictions across an intervention grid yield an estimated response slope, which we classify as correct, flat, or reversed. This is a study of *effect direction for a specified ordered pair*, not discovery of the orientation of an unknown graph edge as in amortized causal discovery (Ke et al., 2023; Lorch et al., 2022).

Across separately trained models, probe-only contexts yield more correct slopes than mixed contexts. At a fixed mixed-context checkpoint, increasing observational dose reduces correct slopes and increases reversals on a common evaluation subset. Replacing the observational values with zeros improves performance while retaining the record format. These interventions establish that the answer depends on observational context at fixed weights, the reversals they induce depart from ideal inference under the training prior, and a pooling model of evidence integration reproduces their dose dependence (Section 3). Additional analyses examine training mixtures, seeds, model size, activation patching, and transfer to binary causal graphs, where good performance on positive-effect worlds turns out not to imply reliable discrimination of positive and negative effects.

Our contributions are as follows.

1. We formalize evidence integration in mixed contexts, showing what probes and observational records each identify, when additional evidence can harm even an ideal reader, that under a positive-effect prior the ideal reader never reverses, and how a simple pooling reader predicts dilution and reversal (§3).
2. We introduce a controlled testbed that separates the training mixture, evidence formats across checkpoints, and interventions on the context of a single checkpoint (§4–§4.1).
3. We show that at fixed weights observational records suppress the use of interventional probes in a dose- and content-dependent way, and that activation transplants at the observational-record positions are sufficient to recover much of the effect (§5.2–§5.5).
4. We show that models trained on positive effects exhibit a sign prior that transfers poorly to negative effects and survives sign-balanced retraining out of distribution (§5.6).

2 Related Work

Evaluating causal reasoning. CLadder evaluates language models on formally specified causal questions (Jin et al., 2023); Corr2Cause examines inference from correlational statements (Jin et al., 2024); and CausalPitfalls studies statistical pitfalls in causal inference (Du et al., 2026). Other work investigates causal knowledge and reasoning in large language models (Kiciman et al., 2024) and whether their estimates of causal quantities compose consistently (Maasch et al., 2025). These evaluations differ in available information and task assumptions, so their accuracy scores need not measure the same capability. Our controlled symbolic training setting complements them by manipulating the evidence presented to a model. Our CLadder-derived audit uses binary SCMs, not the original natural-language benchmark or its accuracy protocol.

Learning causal estimators and strategies. Do-PFN and CausalPFN study amortized causal effect estimation with prior-fitted networks (Robertson et al., 2025; Balazadeh et al., 2025). More recently, CausalFM formalizes SCM-based priors for several causal adjustment settings (Ma et al., 2026), and amortized inference of causal mechanisms given a graph supports interventional predictions (Mahajan et al., 2025). Related studies examine in-context linear regression with endogeneity (Liang et al., 2025) and causal reasoning through axiomatic training (Vashishtha et al., 2025). Meta-trained agents can learn in context to draw causal conclusions from observational and interventional data (Dasgupta et al., 2019), and agents and language models can acquire active causal strategies from passive data (Lampinen et al., 2023). These approaches use different assumptions, observations, and objectives. Our results do not contradict their identifiability or consistency guarantees; we study competition between two evidence types in a particular symbolic prediction task. Synthetic training distributions impose implicit priors in amortized causal discovery while remaining subject to identifiability constraints (Montagna et al., 2025). This distinction motivates our separation of answers consistent with a training prior from predictions that demonstrably depend on interventional evidence.

In-context learning and conflicting evidence. Transformers can learn simple function classes from context (Garg et al., 2022), their behavior can be compared with explicit learning algorithms (Akyürek et al., 2023), and gradient training can make attention encode the latent causal structure of in-context

sequences (Nichani et al., 2024). Interventional probes make our task an analogous response-estimation problem, with observational records supplying a potentially misleading additional signal, akin to spurious features (Harutyunyan et al., 2026) or false demonstrations (Halawi et al., 2024) in in-context learning. Work on knowledge conflicts examines disagreements between contextual information and parametric knowledge (Longpre et al., 2021; Xie et al., 2024), including the attention heads through which both sources influence the output (Li et al., 2025). Here the conflict is between observational and interventional samples in the context, under a learned task prior. We use Simpson-type sign reversals (Pearl, 2009) to make these signals distinguishable, without assuming that the model implements an explicit evidence-selection algorithm of the kind transformers can in principle express (Bai et al., 2023).

3 Problem Formulation

Evidence integration as estimation. A world W is drawn from a prior Π over structural causal models, and a designated ordered pair (X, Y) defines the target response $\mu_W(x) = \mathbb{E}_W[Y \mid do(X = x)]$. A context $C = (O, P)$ contains a set O of observational records, each a draw of the observed variables from $P_W(V)$, and a set P of interventional probes, each a pair $(v, \text{responses})$ drawn from $P_W(V \mid do(X = v))$. Given C and a query value x, a predictor outputs $f(C, x)$, and the question is how well $f(C, \cdot)$ tracks μ_W. The benchmark reference is the *ideal reader* $f^*(C, x) = \mathbb{E}_\Pi[\mu_W(x) \mid C]$, which minimizes expected squared error under the prior. A trained transformer is an amortized approximation to some such reader, and the experiments ask which records it relies on when the two record types disagree.

What each record type identifies. The two record types carry different information about μ_W. Probes are samples from the interventional distribution itself: their responses are unbiased for $\mu_W(v)$ at the probed values, so, for a continuous response and probe values that cover the query grid, the grid slope of μ_W is identified from probes alone as their number grows. Observational records identify $P_W(V)$ but, with a hidden confounder of (X, Y), do not identify μ_W: Appendix A gives two SCMs with the same observational distribution and opposite interventional effects. In a mixed context, observational records can therefore help an ideal reader only through the prior: they are informative about μ_W to the extent that Π couples observational and interventional parameters across worlds.

The Simpson condition in a linear model. The evaluation family is defined by a disagreement in sign, and a linear-Gaussian model shows when it arises. Let $Z \sim \mathcal{N}(0, 1)$ be the hidden confounder and let all noise terms be mutually independent, $X = uZ + \epsilon_X$ with $\mathrm{Var}(\epsilon_X) = \sigma_X^2$, and $Y = \beta X + vZ + \epsilon_Y$. The interventional slope is β, whereas the observational regression slope is

$$\beta_{\mathrm{obs}} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \beta + \frac{uv}{u^2 + \sigma_X^2}. \tag{1}$$

A Simpson trap in the sense of Eq. (4) occurs exactly when $\beta > 0$ and $uv < -\beta(u^2 + \sigma_X^2) = -\beta\,\mathrm{Var}(X)$: the two confounding paths have opposite signs and their product exceeds the direct effect times the variance of X. When both paths strengthen together, $|uv|$ grows, which is why the evaluation family is drawn at the largest confounding strength. The generator uses nonlinear mechanisms and larger graphs, so Eq. (1) illustrates the condition rather than specifying it; Figure 1 shows a sampled evaluation world.

Two comparisons, two questions. Let $h(\theta, C)$ denote the hit rate of parameters θ under context type C on the evaluation family. Comparing a model trained and evaluated with probes only (θ_P, C_P) with one trained and evaluated with mixed contexts (θ_M, C_M) conflates two effects:

$$h(\theta_P, C_P) - h(\theta_M, C_M) = \underbrace{\big[h(\theta_P, C_P) - h(\theta_M, C_P)\big]}_{\text{what training learned}} + \underbrace{\big[h(\theta_M, C_P) - h(\theta_M, C_M)\big]}_{\text{what the context selects}}. \tag{2}$$

The first term changes the weights and holds the context fixed; the second holds the weights fixed and changes the context. Only the second term isolates an inference-time effect, and the fixed-checkpoint experiments estimate its analogue by varying the observational records in the context while the weights stay fixed. Because it is estimated within a single set of weights, it is a paired comparison on identical worlds and requires no assumption that separately trained checkpoints are exchangeable. Table 1 lists each comparison in the paper with the factor it varies and the question it answers.

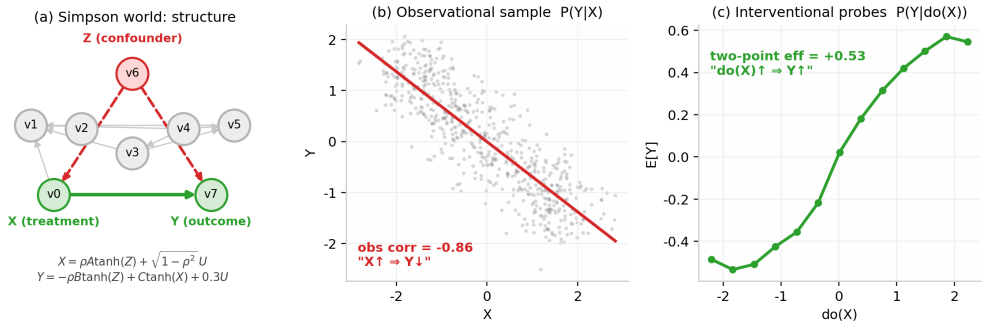

Figure 1: A sampled evaluation world. (a) The designated pair (X, Y) has a positive direct effect, while a hidden confounder drives X and Y in opposite directions. (b) Observational samples show a strong negative association. (c) Interventional probes on X trace a positive response. Observational records in a context carry the sign of panel (b); probes carry the sign of panel (c).

Table 1: Experimental design. Each comparison varies one factor and answers a distinct question; only the fixed-checkpoint rows estimate the context term of Eq. (2).

Comparison	Varies	Holds fixed	Question
Across checkpoints (§5.1)	training and evaluation context type	mixture $\alpha = 1$, worlds	Do models trained on different evidence formats differ?
Observational dose (§5.3)	number of observational records	checkpoint, four probes	Does observational dose change the answer at fixed weights?
Observational content (§5.3)	values in the observational records	checkpoint, probes, record format	Does observational content matter beyond format?
Activation patching (§5.5)	activations at observational-record positions	checkpoint, context	Are those representations sufficient to carry the effect?
Mixture sweep (§5.4)	interventional share α	observation-only context	Does interventional training change the sign under an observational context?
Sign audit (§5.6)	sign distribution of effects	checkpoints	Does performance transfer to negative effects?

When more evidence can hurt. Because f^* is a conditional expectation, adding records cannot increase its expected error under the prior: for any additional set of records R drawn from the same world as C, $f^*(C \cup R, x)$ is the L^2 projection of $\mu_W(x)$ onto a space of functions that contains the space onto which $f^*(C, x)$ projects, so $\mathbb{E}_\Pi\big[(f^*(C \cup R, x) - \mu_W(x))^2\big] \le \mathbb{E}_\Pi\big[(f^*(C, x) - \mu_W(x))^2\big]$. This guarantee holds on average over the prior, not on every subset of worlds. If under Π observational and interventional signs usually agree, observational records raise the ideal reader's posterior weight on the observational sign, which lowers error on agreeing worlds and can raise it on the complementary set of Simpson traps.

A drop in hit rate on Simpson traps can therefore arise from ideal inference under a prior in which signs usually agree. Reversals are different when the prior excludes negative effects.

Proposition 1. *If Π places all its mass on worlds whose interventional grid slope $b^\star_W$ is positive, then for every context C shared by the query points, the grid slope of $f^*(C, \cdot)$ equals $\mathbb{E}_\Pi[b^\star_W \mid C] > 0$; the ideal reader never reverses.*

The least-squares grid slope is linear in the responses, so the slope of $f^*(C, \cdot) = \mathbb{E}_\Pi[\mu_W(\cdot) \mid C]$ is the conditional expectation of $b^\star_W$, which is positive. Our evaluation draws a fresh context at each query point and averages eight draws, and because the least-squares slope is linear in the responses, the ideal reader's measured slope then has a positive expectation, the average of the slope above over contexts, plus a zero-mean draw-to-draw term whose variance the averaging reduces eightfold. Under the positive-effect training family, the reversals produced at fixed weights by adding observational records, eleven of 37 worlds at the largest dose (Section 5.3), therefore indicate a departure from

ideal inference under the training prior rather than a consequence of it, without computing the ideal reader itself.

A pooling model of dilution and reversal. A simple reader makes the fixed-checkpoint predictions concrete; transformers can implement readers that select among estimators according to the context (Bai et al., 2023). Suppose the model forms its slope as a weighted average of the slopes carried by each evidence type, giving each of n_p probes unit weight and each of n_o observational records weight $\omega > 0$:

$$\hat{b}(n_o) = \frac{n_p\, b_{\text{int}} + \omega\, n_o\, b_{\text{obs}}}{n_p + \omega\, n_o}, \tag{3}$$

where $b_{\text{int}} > 0$ is the probe slope and b_{obs}, negative on Simpson traps, the slope carried by the observational values. Equation (3) predicts that (i) the estimate decreases monotonically in n_o when $b_{\text{obs}} < b_{\text{int}}$; (ii) it reverses sign once $n_o > n_p\, b_{\text{int}}/(\omega|b_{\text{obs}}|)$, so reversals appear first in worlds whose observational slope is largest relative to the probe slope; and (iii) observational records whose values carry no slope, $b_{\text{obs}} = 0$, attenuate the estimate without ever reversing it. The dose and content experiments of Section 5.3 agree with (i) and (iii): hits fall and reversals rise monotonically with observational dose, and zero-valued records nearly eliminate reversals, leaving one where same-world records leave nine; (ii) is a quantitative prediction for designs that vary the observational slope. Cross-world and scrambled values give outcomes between the same-world and zero-valued conditions, indicating that the effective weight also depends on the values of the records and not only on their number.

4 Experimental Setup

Worlds and estimands. A world $W = (\mathcal{G}, \mathcal{F}, P_U)$ is a structural causal model with mechanisms $V_i = f_i(\text{pa}(V_i), U_i)$. The generator uses 5–12 nodes, Erdős–Rényi or scale-free graph structure with equal probability, and mixed linear and nonlinear mechanisms. A designated triple has confounding paths $Z \to X$ and $Z \to Y$, with confounding strength $\rho \in \{0, .3, .6, .9\}$. The hidden confounder is not available as an adjustment variable to the predictor. Let $\mu_w(x) = \mathbb{E}_W[Y \mid do(X = x)]$. The main evaluation uses 50 held-out worlds at $\rho = .9$ selected to satisfy

$$\text{Corr}_{\text{obs},w}(X, Y) < 0, \qquad \text{eff}_w = \frac{\mu_w(1) - \mu_w(-1)}{2} > 0. \tag{4}$$

The mean endpoint effect is approximately 0.57. This family isolates the conflict between association and intervention. Worlds are generated online with disjoint train, validation, test, and larger-graph seed domains. Online sampling avoids reuse of fixed examples.

Evidence and queries. Observational records contain sampled observed-variable values. An interventional record additionally marks which variable was externally set and its value. A *probe* intervenes on the queried variable X; a random-node intervention need not do so. Queries ask for Y under $do(X = x)$ on $\mathcal{X} = \{-1, -.5, 0, .5, 1\}$. Probe values are drawn from a two-sided range with equal probability of either side. In the primary comparison, the observation-only model receives four observational records, C1 receives *four observations plus two probes*, and C2 receives four probes. These contexts differ in both evidence type and total length. The fixed-checkpoint dose experiment instead holds the number of probes at four and varies the number of observations. Records are serialized as an ordered sequence, rather than treated as a permutation-invariant set.

Training factors. The primary model is a 25.7M-parameter GPT-style decoder with eight layers, vocabulary size 41, and maximum sequence length 640. Training uses 20,000 steps, batch size 64, learning rate 6×10^{-4}, and 500 warmup steps. The training mixture selects an interventional example with probability α, independently per sample. This example-level mixture and the composition of its evidence context are separate factors: $\alpha = 1$ does not imply probe-only context. The primary evidence comparison fixes $\alpha = 1$; a separate sweep varies α while retaining observation-only context. Table 2 summarizes the training and evaluation settings.

Identifiability boundary. As Section 3 shows, observational records alone do not identify the interventional effect, so the observation-only comparison diagnoses behavior under the learned prior rather than failure of an identifiable estimator. Moreover, the context-free rule $\hat{y}(x) = 0.1x$ has slope $0.1 > \tau_0$ and therefore achieves 50/50 hits on the all-positive main family by construction, without

Table 2: Training and primary evaluation configuration. The context composition is independent of the example-level intervention mixture α.

Component	Setting
Primary decoder	25.7M parameters; 8 layers; vocabulary size 41; maximum sequence length 640
Training schedule	20,000 steps; batch size 64; learning rate 6×10^{-4}; 500 warmup steps
World families	5–12 nodes; Erdős–Rényi or scale-free structure; mixed linear/nonlinear mechanisms
Confounding strengths	$\rho \in \{0, .3, .6, .9\}$; primary test family $\rho = .9$
Mixture sweep	$\alpha \in \{0, .1, .3, .5, .7, 1\}$; observation-only context; one run per level
Primary comparison	$\alpha = 1$; baseline: 4 obs; C1: 4 obs + 2 probes; C2: 4 probes
Effect evaluation	50 held-out worlds; grid $\{-1, -.5, 0, .5, 1\}$; 8 context draws per query; slope threshold .05
Fixed-C1 interventions	4 probes; dose: 1/2/4 obs; content: same/cross/scramble/zero values
Larger-graph evaluation	13–16 nodes; 20 attempted worlds
Larger decoder	0.93B parameters; 20,000 steps; learning rate 3×10^{-4}
Activation patching	4 context draws per query; responsive reversed-world subsets

estimating the correct effect magnitude. This analytic sanity check shows why hit rate on this family alone cannot certify evidence use. The principal evidence for context interference therefore comes from changes within a checkpoint, and sign generalization is tested separately (Section 5.6).

4.1 Measurement and statistical scope

For each world, predictions are averaged across $n_{\text{rep}} = 8$ independent context draws per query and fitted by least squares:

$$\bar{y}_w(x) = \frac{1}{n_{\text{rep}}} \sum_{r=1}^{n_{\text{rep}}} \hat{y}_w(x; C_{w,x,r}), \quad (\hat{a}_w, \hat{b}_w) = \arg\min_{a,b} \sum_{x \in \mathcal{X}} (\bar{y}_w(x) - a - bx)^2. \tag{5}$$

Let $b^\star_w$ denote the analogous grid slope of the Monte Carlo oracle response $\mu_w(x)$, and define the normalized slope $r_w = \hat{b}_w / b^\star_w$. Values of r_w near one indicate recovery of the oracle grid-slope magnitude; values near zero indicate an attenuated response. The endpoint effect eff_w and grid slope $b^\star_w$ are different estimands for nonlinear responses. We therefore distinguish endpoint-sign classification from normalized grid slopes; the two direction summaries are interchangeable only when their oracle signs agree.

Using the threshold $\tau_0 = .05$, a *hit* has $|\hat{b}_w| > \tau_0$ and the same sign as eff_w; a *flat* estimate has $|\hat{b}_w| \leq \tau_0$; and a *reversal* has $|\hat{b}_w| > \tau_0$ and the opposite sign.

Units and missing outputs. Worlds, not grid points or repeated context draws, are the units of comparison. We report denominators after parsing and restrict paired comparisons to worlds parsed under every compared condition. For example, the dose and content analyses retain 37/50 and 33/50 worlds, respectively. Appendix D gives bounds that count all excluded worlds.

Uncertainty. The primary three pairwise hit comparisons use two-sided exact McNemar tests, with Holm correction across those three contrasts. Wilson intervals describe uncertainty in world-level hit proportions, and training-seed variation is reported separately (Section 5.4). A repeated-probe simulation with a least-squares reader provides a reference error rate (Appendix C).

5 Results

5.1 Comparison across checkpoints

Table 3 compares models trained and evaluated with the same context type. At fixed $\alpha = 1$, the probe-only model C2 has 41/50 hits and 4/50 reversals, compared with 18/50 and 19/50 for mixed-context C1. The observation-only baseline has 7/50 hits and 29/50 reversals. Mean normalized

Table 3: Primary comparison: 50 common worlds, $\rho = .9$, $\alpha = 1$, and eight context draws per query. Each row is a different checkpoint. Hit, flat, and reversal counts use the same 0.05 slope threshold. The interval is a 95% Wilson interval for the hit proportion, not variability across training seeds.

Model (obs/probes)	Hit	Flat	Rev.	Mean r	Hit (%)	95% CI
Baseline (4/0)	7	14	29	$-.288$	14	$[7.0, 26.2]$
C1 (4/2)	18	13	19	$-.092$	36	$[24.1, 49.9]$
C2 (0/4)	41	5	4	$+.418$	82	$[69.2, 90.2]$
C2-curr2 (0/4)	40	6	4	$+.482$	80	$[67.0, 88.8]$

Table 4: Fixed-checkpoint interventions. Dose rows use the same 37 commonly parsed worlds; content rows use a separate common subset of 33 worlds. Counts and means must be compared within a block. All rows contain four interventional probes.

Intervention	Condition	Hit	Flat	Rev.	Mean r
Dose ($n = 37$)	1 observation	31	5	1	$+.472$
	2 observations	22	12	3	$+.204$
	4 observations	17	9	11	$+.037$
Content ($n = 33$)	Same-world values	16	8	9	$+.059$
	Cross-world values	21	6	6	$+.192$
	Scrambled values	20	8	5	$+.203$
	Zero values	25	7	1	$+.378$

slope also increases across these settings, from $-.288$ to $-.092$ to $+.418$. This comparison contrasts complete training-and-context configurations, which also differ in probe count and context length; the fixed-checkpoint experiments below isolate the effect of context.

We compare hits using three paired contrasts. C2 versus baseline has 35 wins and one loss; C2 versus C1 has 25 wins and two losses; C1 versus baseline has 14 wins and three losses. Their two-sided exact McNemar p-values are 1.08×10^{-9}, 5.65×10^{-6}, and .0127, respectively; Holm-adjusted values are 3.23×10^{-9}, 1.13×10^{-5}, and .0127. Because the family is all-positive, C2's success is complemented by the sign audit of Section 5.6.

5.2 Fixed-checkpoint interventions

Small preliminary interventions on frozen C1 motivate a larger context analysis. In the observation-removal ablation, removing observational records gives 7/9 hits and mean $r = +.56$. In the probe-count ablation, increasing the probe count from two to four while retaining observations gives 4/8 hits and two reversals. These small subsets motivate the common-subset dose and content experiments below.

5.3 Observational dose and content

All conditions in Table 4 use the same C1 checkpoint, four probes, and eight context draws per query. The dose experiment varies the number of observations; the content experiment retains four observational records while changing their values. Both depart from C1's four-observation/two-probe training format.

Dose. Increasing the number of observations from one to four reduces hits from 31/37 to 17/37 and increases reversals from 1/37 to 11/37; by Proposition 1, these reversals indicate a departure from ideal inference under the training prior. Mean r falls by $.435$, from $+.472$ to $+.037$. At two observations it is $+.204$, with 12 flat estimates and three reversals. The exploratory one-versus-four comparisons give McNemar $p = .0013$ and signed-rank $p = 1.0 \times 10^{-5}$. The direction of the hit-count difference survives even a worst-case allocation of the 13 worlds omitted from the common subset (Appendix D). Because adding observations also lengthens the context, the content experiment holds the record count fixed and varies only the values.

Content. SAME uses observations from the queried world; CROSS substitutes a different world's observations; SCRAMBLE is intended to retain marginals while disrupting correlations; and ZERO replaces observational values with zeros. Compared with SAME, ZERO increases hits from 16/33

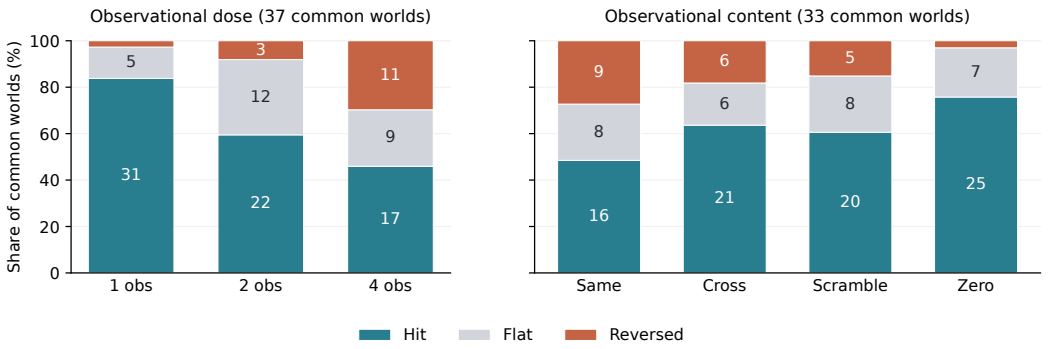

Figure 2: Observed class proportions at a fixed C1 checkpoint, redrawn from Table 4. Increasing observational dose shifts mass from hits toward flat and reversed estimates (left). Replacing observed values with zeros improves the complete-case outcome (right). The two panels use different common-world subsets.

to 25/33 and mean r from $+.059$ to $+.378$. The paired p-values are .023 for hits and .0002 for normalized slopes. Because record markers remain, the contrast is driven by the observational values rather than the record format.

CROSS and SCRAMBLE give intermediate outcomes (their comparisons with SAME give $p \approx .07$–$.08$), placing the queried world's own values at the harmful end of the range. Both observational dose and observational values therefore affect predictions at fixed weights.

5.4 TRAINING MIXTURE

Training mixture. A separate six-level sweep uses observation-only context and $\alpha \in \{0, .1, .3, .5, .7, 1\}$, with one run per level. Slope magnitude rises from .169 to .384 while the aggregate sign remains negative. Increasing the interventional training share therefore does not correct the sign error under an observation-only context. The sweep also yields sign accuracies of .5 at $\alpha = 0$ and .2–.4 otherwise; .5 is not an appropriate universal chance baseline for a test family whose true effects all have the same sign. Other small-context ablations and training trajectories are documented in Appendix E.

Seeds and transfer. C1's hit/flat/reversed profiles are 18/13/19, 18/13/18, and 18/13/19 across three seeds, with one missing parse for the second seed. C2 has 41/5/4 and 44/3/3 across two seeds. C1's per-world slopes correlate across seeds ($r = .825$ and .872 in two comparisons), suggesting repeatable world-specific structure within this model family. A C2 model trained at $\rho = .6$ and evaluated at .9 gives 39/8/3, and the harder-curriculum C2 gives 40/6/4. On the larger 13–16-node domain, 16/20 worlds parse, of which 12 are hits: this is 75% of parsed worlds or 60% of all attempted worlds. The latter denominator is necessary when assessing end-to-end transfer.

5.5 ACTIVATION PATCHING

Because patching conclusions depend on the corruption and the metric (Zhang & Nanda, 2024), we state ours explicitly: the analysis replaces activations in a same-world observational context with activations from a zeroed-observation context, holding the checkpoint fixed. It is restricted to originally reversed worlds with a sufficiently large baseline release under zeroing. Patching observational-record residual states in early and middle layers restores much of that release in the selected subsets; a search over sparse head/MLP components does not identify a reliably effective subset (Appendix G). The effect of context is thus carried by the representations of the observational records in early and middle layers. The absence of an effect at the final layer is expected architecturally: after the final attention operation, no cross-token operation can transmit an edit at non-query positions to the prediction.

5.6 SIGN GENERALIZATION

The main family admits a positive-sign shortcut (Geirhos et al., 2020), analogous to the surface heuristics that language models adopt when inferring causal relations from text (Joshi et al., 2024). To probe this limitation, we map CLadder-derived binary SCMs into the symbolic channel, using $x \in \{0, 1\}$ so the response difference is an average treatment effect (ATE). The confounding-family suite contains 23 positive-ATE and 27 negative-ATE worlds. C1 reverses on 25/27 negative worlds and C2 on 24/27; prediction–ATE correlations are $.53$ and $.48$ on positive-ATE worlds but approximately $-.08$ and $-.10$ on negative-ATE worlds, so both models track effect magnitude only when the effect is positive. This is strong evidence against general sign competence under that transfer protocol. The broader suite retains nine of ten graph families for which exact enumeration agrees with the reference ground truth (Appendix H).

Regressing predicted effects on true ATEs gives positive intercepts of $.386$ for C1 and $.233$ for C2. These offsets are consistent with a training prior favoring positive effects. Sign-randomized retraining provides a further test: the in-distribution normalized slope is $+.74$ for C2-sr and $-.06$ for C1-sr, while positive offsets persist on the binary audit. This suggests that balancing effect signs alone does not remove the out-of-distribution bias. The retraining and binary-SCM results are detailed in Appendix H.

Scale. At 0.93B parameters, C1 has 14 reversals among 44 parsed worlds and C2 has three among 50, alongside substantially smaller slope magnitudes. These runs use the same step budget as the 25.7M runs and have different parse rates. Because reversal requires crossing a fixed absolute threshold, reversal frequency is not invariant to changes in prediction magnitude; the full counts are in Appendix F.

6 DISCUSSION AND LIMITATIONS

The central result is a context-dependent change in a fixed model: observational dose and values change interventional predictions even when informative query-targeted probes remain in the context. The cross-checkpoint comparison, the fixed-checkpoint interventions, and the activation transplants answer different causal questions, and together they place the effect at inference time and implicate the representations at the positions that carry the observational records. The formulation of Section 3 identifies the general condition behind it: whenever training data mostly contain agreeing observational and interventional evidence, a reader that pools the two will be most wrong exactly on the confounded cases that interventional data are meant to resolve.

The practical implication is that causal learning should be evaluated along two axes rather than one. The training mixture shapes the magnitude of the interventional response a model produces; the evidence composition at inference time determines whether it uses that ability, echoing findings that in-context behavior reflects competition among strategies shaped by the training distribution (Park et al., 2025; Singh et al., 2025). Evaluations that vary only the mixture, or report accuracy on a family in which effects share a sign, can miss both the interference shown here and the sign prior exposed by the binary-graph audit. Proposition 1 separates the two for reversals. The proposition also provides a general test for other amortized causal learners: under a sign-restricted training prior, any sign reversal induced by additional context identifies suboptimal evidence use without reference to the learner's internals. For in-context causal estimation more broadly, the results imply that appending observational data to an interventional query is not neutral: at fixed weights it can pull the answer toward the observational sign, so a model's sensitivity to observational dose should be measured before its answer on mixed evidence is trusted.

Limitations. Our models are small symbolic transformers trained from scratch, which gives exact control over the evidence presented to the model. The main evaluation family has positive effects, so sign inference is assessed through the binary-graph audit, and record count and context length vary together in the dose experiment, which the content experiment complements at a fixed record count. Activation patching establishes that the observational-record representations suffice to carry the effect for the tested perturbation, and the statistical support comes from 50 worlds and a small number of training seeds per configuration.

Ethics Statement

The experiments use synthetic causal systems and a symbolic audit derived from a public benchmark; they involve no human subjects or personal data. The main risk our results point to is overtrust: a model that predicts effects well on agreeing evidence can be systematically wrong on confounded cases, so causal conclusions drawn by learned models should not inform scientific, medical, or policy decisions without domain-specific causal assumptions and evaluation.

Reproducibility Statement

The main text states the estimation problem, the assumptions of the illustrative models, the experimental factors, and the estimands, aggregation, and statistical comparisons. The appendix gives the identifiability derivation, the model configuration, training schedule, evaluation settings, and version identifiers, the reference-error simulation, the missing-world bounds, the context formats, and the protocols for the larger-model runs, activation patching, and the binary-SCM audit.

AI Use Statement

We used generative AI tools to assist with structuring, writing, and editing the manuscript, checking the consistency of claims and summary statistics, searching for and verifying related literature, and preparing figures from aggregate results. AI tools also assisted in deriving and writing up the illustrative models and the proposition in the problem formulation; the authors checked every derivation and its connection to the data. We take responsibility for the final content of this paper, including text, claims, proofs, and figures produced with the aid of generative AI.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. In *Advances in Neural Information Processing Systems*, 2023.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse C. Cresswell, and Rahul G. Krishnan. CausalPFN: Amortized Causal Effect Estimation via In-Context Learning. In *Advances in Neural Information Processing Systems*, 2025.

Ishita Dasgupta, Jane X. Wang, Silvia Chiappa, Jovana Mitrovic, Pedro A. Ortega, David Raposo, Edward Hughes, Peter W. Battaglia, Matthew M. Botvinick, and Zeb Kurth-Nelson. Causal Reasoning from Meta-reinforcement Learning. *arXiv preprint arXiv:1901.08162*, 2019.

Jin Du, Li Chen, Xun Xian, An Luo, Fangqiao Tian, Ganghua Wang, Charles Doss, Xiaotong Shen, and Jie Ding. Ice Cream Doesn't Cause Drowning: Benchmarking LLMs Against Statistical Pitfalls in Causal Inference. In *International Conference on Learning Representations*, 2026.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. In *Advances in Neural Information Processing Systems*, 2022.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. Overthinking the Truth: Understanding how Language Models Process False Demonstrations. In *International Conference on Learning Representations*, 2024.

Hrayr Harutyunyan, Rafayel Darbinyan, Samvel Karapetyan, and Hrant Khachatrian. In-context Learning in Presence of Spurious Correlations. *Transactions on Machine Learning Research*, 2026.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing Causal Reasoning in Language Models. In *Advances in Neural Information Processing Systems*, 2023.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. Can Large Language Models Infer Causation from Correlation? In *International Conference on Learning Representations*, 2024.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. LLMs Are Prone to Fallacies in Causal Inference. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Nan Rosemary Ke, Silvia Chiappa, Jane X. Wang, Jörg Bornschein, Anirudh Goyal, Mélanie Rey, Theophane Weber, Matthew M. Botvinick, Michael Curtis Mozer, and Danilo Jimenez Rezende. Learning to Induce Causal Structure. In *International Conference on Learning Representations*, 2023.

Emre Kiciman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *Transactions on Machine Learning Research*, 2024.

Andrew K. Lampinen, Stephanie C. Y. Chan, Ishita Dasgupta, Andrew J. Nam, and Jane X. Wang. Passive learning of active causal strategies in agents and language models. In *Advances in Neural Information Processing Systems*, 2023.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming Knowledge Conflicts in Language Models. In *International Conference on Machine Learning*, 2025.

Haodong Liang, Krishna Balasubramanian, and Lifeng Lai. Transformers Handle Endogeneity in In-Context Linear Regression. In *International Conference on Learning Representations*, 2025.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-Based Knowledge Conflicts in Question Answering. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized Inference for Causal Structure Learning. In *Advances in Neural Information Processing Systems*, 2022.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation Models for Causal Inference via Prior-Data Fitted Networks. In *International Conference on Learning Representations*, 2026.

Jacqueline R. M. A. Maasch, Alihan Hüyük, Xinnuo Xu, Aditya V. Nori, and Javier González. Compositional Causal Reasoning Evaluation in Language Models. In *International Conference on Machine Learning*, 2025.

Divyat Mahajan, Jannes Gladrow, Agrin Hilmkil, Cheng Zhang, and Meyer Scetbon. Amortized Inference of Causal Models via Conditional Fixed-Point Iterations. *Transactions on Machine Learning Research*, 2025.

Francesco Montagna, Max Cairney-Leeming, Dhanya Sridhar, and Francesco Locatello. Demystifying amortized causal discovery with transformers. *Transactions on Machine Learning Research*, 2025.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How Transformers Learn Causal Structure with Gradient Descent. In *International Conference on Machine Learning*, 2024.

Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. Competition Dynamics Shape Algorithmic Phases of In-Context Learning. In *International Conference on Learning Representations*, 2025.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-PFN: In-Context Learning for Causal Effect Estimation. In *Advances in Neural Information Processing Systems*, 2025.

Aaditya K. Singh, Ted Moskovitz, Sara Dragutinovic, Felix Hill, Stephanie C. Y. Chan, and Andrew M. Saxe. Strategy Coopetition Explains the Emergence and Transience of In-Context Learning. In *International Conference on Machine Learning*, 2025.

Aniket Vashishtha, Abhinav Kumar, Atharva Pandey, Abbavaram Gowtham Reddy, Kabir Ahuja, Vineeth N. Balasubramanian, and Amit Sharma. Teaching Transformers Causal Reasoning through Axiomatic Training. In *International Conference on Machine Learning*, 2025.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. In *International Conference on Learning Representations*, 2024.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, 2024.

A AN ILLUSTRATIVE IDENTIFIABILITY EXAMPLE

This example illustrates the boundary of the observation-only experiment. Let Z, E be independent standard normal variables and $X = Z + E$. Consider two SCMs with independent Gaussian outcome noise:

$$W_+ : \quad Y = X - 3Z + \epsilon_+, \qquad \epsilon_+ \sim \mathcal{N}(0, 1), \tag{6}$$

$$W_- : \quad Y = -X + Z + \epsilon_-, \qquad \epsilon_- \sim \mathcal{N}(0, 5). \tag{7}$$

In both models (X, Y) is zero-mean Gaussian with covariance

$$\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 6 \end{pmatrix}.$$

Indeed, $Y_+ = E - 2Z + \epsilon_+$ and $Y_- = -E + \epsilon_-$ both have variance six and covariance -1 with X. Yet intervening on X yields $\mathbb{E}[Y \mid do(X = x)] = x$ in W_+ and $-x$ in W_-. Thus, observing only (X, Y) cannot identify the effect sign over this two-model class. Under equal prior probabilities, every observational-data sign classifier has average error $1/2$ because the data distributions coincide. A prior that excludes one model, such as a family constrained to positive effects, or a context containing informative interventional probes makes the sign predictable.

B TRAINING AND EVALUATION DETAILS

World generation and replay. Training, validation, test, and larger-graph seed domains are disjoint. The primary test family uses `make_world(SEED_TEST + i, rho=0.9, curriculum=1)` for $i = 0, \ldots, 49$. The intervention-value cache persists across generator calls, so exact replay requires explicitly resetting its random state in addition to the world seed.

Evaluation version identifiers. The MD5 identifiers of the evaluation scripts are:

`eval.py`	`e9c347704ef8ddf9524c63ddb06bf094`
`eval_ckpt.py`	`f877ac2b3e7a02c3d910f29b7df69bde`
`mech_patch.py`	`542e2dd7d0e88081c6b8758bb63cb8c0`

The primary behavioral comparisons use eight context draws; the patching analysis uses four and is reported separately.

C REPEATED EVIDENCE AND THE REFERENCE ERROR SIMULATION

The reference simulation uses a least-squares reader of four query-targeted probes, a five-point grid, and 20 repetitions on ten worlds. Its sign-error rates for $n_{\text{rep}} = 1, 2, 4, 8$ are 26%, 17%, 12%, and 9%, respectively. These rates characterize this reader, world subset, and sampling protocol, and they count ungated sign errors rather than thresholded reversals.

For independent context draws with finite conditional variance $\sigma_w^2(x)$, averaging predictions gives

$$\mathrm{Var}[\bar{y}_w(x) \mid w] = \frac{\sigma_w^2(x)}{n_{\text{rep}}}.$$

This reduces sampling variance but not systematic bias. The grid slope is linear in the mean responses,

$$\hat{b}_w = \sum_{x \in \mathcal{X}} a_x \bar{y}_w(x), \qquad a_x = \frac{x - \bar{x}}{\sum_{u \in \mathcal{X}} (u - \bar{x})^2},$$

so its variance also depends on covariance between grid responses when contexts or randomness are shared. For a biased model, the sign-error rate therefore depends on the bias as well as on this variance.

D PAIRED COMPARISONS AND MISSING-WORLD BOUNDS

For discordant hit counts b, c, the two-sided exact McNemar test is

$$p = \min\left\{1,\ 2\sum_{j=0}^{\min(b,c)}\binom{b+c}{j}2^{-(b+c)}\right\}.$$

The primary comparisons use $(b, c) = (35, 1), (25, 2), (14, 3)$. Holm correction is applied across this three-comparison family. These paired tests condition on the trained checkpoints; training-seed variation is reported in Section 5.4.

For a condition with h hits among m commonly parsed worlds out of N attempted worlds, its full-set hit proportion lies in

$$\left[\frac{h}{N}, \frac{h+N-m}{N}\right]$$

when each outcome outside the common subset may be either a hit or a non-hit. These bounds hold for every assignment of the missing outcomes.

Table 5: Full-set hit bounds from the common subsets, with $N = 50$.

Experiment	Condition	Hits/common worlds	Full-set hit bound
Dose	1 obs	31/37	$[62, 88]\%$
	2 obs	22/37	$[44, 70]\%$
	4 obs	17/37	$[34, 60]\%$
Content	Same	16/33	$[32, 66]\%$
	Cross	21/33	$[42, 76]\%$
	Scramble	20/33	$[40, 74]\%$
	Zero	25/33	$[50, 84]\%$

For the one-versus-four observation comparison, the full-set hit difference is bounded by $[2, 54]$ percentage points: the observed difference is 14 hits, while only 13 worlds lie outside the common subset. Its direction therefore holds under every assignment of those missing outcomes. For zero versus same content, the analogous bound is $[-16, 52]$ points, and the direction rests on the paired common-subset comparison of Section 5.3.

E CONTEXT FORMATS AND ADDITIONAL RESULTS

C1 is trained with four observations and two probes. With each probe serialized on two lines, this gives an eight-line evidence format; increasing the probe count to four gives twelve lines. The 25.7M C1 has the same aggregate outcomes under two and four probes (18/13/19 and mean $r = -.092$).

Training trajectory and small ablations. A no-context ablation gives zero reversals in ten worlds with slope magnitude .018, indicating near-flat predictions. A random-node-intervention ablation, whose interventions need not target X, gives seven reversals in ten worlds. On ten worlds with a single context draw, the normalized slope of C1 falls from $+.172$ to $-.266$ after training step 12k.

Replication summaries. The C1 reversed-world sets overlap in 14 worlds between seeds 0 and 1, and 15 between seeds 0 and 2. Eleven strong reversals persist in a matched comparison that excludes one parse failure. This repeatability holds on the evaluation world set for the 25.7M training configuration.

Further checks. C2 with injected observations has zero parsing success. The eight-observation C1 dose condition has parse rate .16. An association between suppression and observational-correlation magnitude observed in a ten-world pilot does not reach significance in the 50-world follow-up ($p = .14$).

F THE LARGER-MODEL COMPARISON

The 0.93B runs share the 20,000-step budget of the 25.7M runs and use learning rate 3×10^{-4}. Because the 0.05 slope threshold is absolute, reversal frequency also depends on the overall magnitude

Table 6: Scale comparison, using the primary context formats. All conditions attempt 50 worlds; the Missing column counts worlds without a parseable output.

Model	Parameters	Hit	Flat	Reversed	Missing	Mean r
C1	25.7M	18	13	19	0	$-.092$
C2	25.7M	41	5	4	0	$+.418$
C1	0.93B	12	18	14	6	$+.008$
C2	0.93B	14	33	3	0	$+.072$

of predictions, which is smaller at 0.93B. The comparison therefore describes the joint change of model size and training recipe.

Among the 19 C1 reversals at 25.7M, six remain reversed at 0.93B, seven become flat, four become hits, and two fail to parse. Eight previously non-reversed worlds become reversed. The Spearman correlation is $.084$ on 44 paired worlds. The world-level pattern of reversals therefore changes with the model and recipe. At the longer four-probe C1 format, the 0.93B model's parse rate is $.32$, compared with $.88$ under its training-length format.

G Activation patching: scope and controls

For each selected reversed world, define slopes s_{same} under same-world observational values and s_{zero} under zeroed values. The release statistic is

$$R = \frac{s_{\text{patch}} - s_{\text{same}}}{s_{\text{zero}} - s_{\text{same}}}, \qquad s_{\text{zero}} - s_{\text{same}} > .05.$$

The guard restricts the analysis to worlds that respond to zeroing. The analysis uses four repeated contexts per query. Seed 0 divides its 19 reversed worlds into ten development and nine holdout worlds; the guard retains eight in each subset. Seed 1 uses 18 reversed worlds without a development/holdout split and retains 17.

Interventions. On seed-0 development worlds, patching observational-record residual states gives $R = 1.01$–1.04 at layers 0–3 and $.81$–$.92$ at layers 4–6. The corresponding seed-1 values span $.94$–1.04 at layers 0–6. Both scans give exactly zero at layer 7. Non-observation patches are smaller in the seed-0 scan (maximum $.36$) and negative in seed 1 ($-.13$ to $-.26$). The development-selected top-eight component union gives holdout $R = .20$, with 95% interval $[-.43, .67]$ and comparison against random components $p = .19$. Thus the component search does not isolate a reliable sparse circuit.

Interpretation. The row-level interventions show that the residual states at observational-record positions are sufficient to change predictions relative to the chosen baselines in responsive reversed worlds. The final-layer null follows from the computation graph: after the last cross-token operation, editing non-query states cannot affect the current prediction.

H Binary-SCM audit and sign randomization

The CLadder-derived audit uses binary SCMs serialized in the same symbolic channel as the synthetic training task. Interventions are evaluated at $x \in \{0, 1\}$, so the difference in expected responses is the ATE. The all-family suite contains 45 worlds, with five worlds in each of nine graph families retained after agreement between exact enumeration and the reference ground truth. The confounding-family suite contains 50 worlds. All 285 model–world evaluations across three model configurations on these two suites parse successfully.

Asymmetric effect prediction. The confounding suite contains 23 positive-ATE and 27 negative-ATE worlds. Regressing predicted effects on true ATEs gives intercepts $.386$ for C1 and $.233$ for C2. Within positive-ATE worlds, prediction–ATE correlations are $.53$ and $.48$; on negative-ATE worlds they are approximately $-.08$ and $-.10$. C1 and C2 reverse on 25/27 and 24/27 negative-ATE worlds, respectively. A second C1 seed gives 13/27 such reversals, showing that the severity of this failure varies across runs. Positive offsets together with asymmetric magnitude tracking indicate a positive-sign prior.

Sign-randomized retraining. C1-sr and C2-sr use the 25.7M training recipe with a world-specific sign applied to the causal and confounding pathways. On the training-distribution probe battery, normalized slopes are $+.74$ for C2-sr, $-.06$ for C1-sr, and $+.18$ for the corresponding non-randomized probe-only comparison. On the binary audit, intercepts remain near $+.21$ across the randomized model–suite combinations. Thus, the training modification improves the in-distribution probe-only response while leaving a positive out-of-distribution offset. Sign balance in the training family and calibration under a change in variable domain and graph distribution are therefore separate problems.