

CausalPrism: Component-Level Evaluation of Intervention-Based Causal Reasoning in Large Language Models

Jinyue Feng

University of Toronto, Toronto, Canada
Vector Institute, Toronto, Canada
jinyue@cs.toronto.edu

Frank Rudzicz

Dalhousie University, Halifax, Canada
Vector Institute, Toronto, Canada
frank@dal.ca

Abstract

Evaluating causal reasoning in large language models (LLMs) is complicated by benchmarks that conflate distinct reasoning operations into aggregate scores. We propose CausalPrism, an evaluation framework grounded in structural causal model semantics and cognitive models of causal inference, which decomposes intervention-based causal reasoning into three functional primitives—Causal Prior, Forward State Tracing, and Belief Updating—plus a consistency primitive, Agency Sensitivity. Primitives are instantiated in directed acyclic graphs and evaluated across three narrative domains of varying semantic richness. Across 14 open-source and frontier LLMs, Belief Updating is the dominant bottleneck: the best model reaches only 0.613 accuracy, while GPT-5.2 and GPT-5-mini score 0.531 and 0.517 despite near-ceiling performance on Causal Prior and Forward State Tracing. Prior exhibits systematic domain sensitivity and over-rejection. Agency framing effects are family-dependent for weaker models. We argue against treating causal reasoning as a unitary capability and motivate targeted evaluation and training along multiple primitive axes.

1 Introduction

Large language models (LLMs) are increasingly used to interpret, explain, and act on causal narratives (Zadrozny, 2026; Chen et al., 2025; Gkoutouras et al., 2025), yet evaluating “causal reasoning” in LLMs remains challenging. Existing benchmarks often report aggregate accuracy across heterogeneous query types (Jin et al., 2023; Wang, 2024; Chi et al., 2024; Feng and Rudzicz, 2025), making it difficult to determine *why* a model succeeds or fails: is failure driven by misunderstanding causal structure, mis-propagating interventions, incorrect belief revision, or framing sensitivity?

Both interventionist causal theory and cognitive models of causal inference suggest that

causal reasoning comprises multiple distinct operations rather than a single capability (Pearl, 2009; Woodward, 2003; Griffiths and Tenenbaum, 2009; Steyvers et al., 2003). Prior theory establishes that such operations differ but does not provide a decomposition suited to diagnosing LLM behaviour. We propose **CausalPrism**, a diagnostic decomposition designed to localize failures in intervention-based causal reasoning rather than aggregate heterogeneous causal queries into a single score. We provide three functional primitives — CAUSAL PRIOR, FORWARD TRACING, and BELIEF UPDATING — that each target a distinct reasoning operation, along with a consistency primitive, AGENCY SENSITIVITY¹, which evaluates whether causal reasoning remains invariant under contrasting epistemic framings. All primitives are instantiated in deterministic DAG-based systems across three narrative domains of varying semantic richness (Section 3); see Example 1 for a worked illustration.

We identify four robust patterns: (i) PRIOR and TRACING often saturate for frontier models, but BELIEF UPDATING remains challenging; (ii) updating failures concentrate on *contradict* predictions, and stratified analysis of label-wise signatures reveals distinct behavioural profiles across performance tiers; (iii) PRIOR shows the strongest domain sensitivity and a characteristic over-rejection failure mode; and (iv) AGENCY framing is largely negligible in stronger models but becomes family-dependent and diagnostic for weaker ones.

Contributions. (1) We introduce a factored framework decomposing intervention-based causal reasoning into four primitives grounded in SCM semantics and cognitive models. (2) We synthesize a benchmark dataset instantiating the framework across three controlled narrative domains with algorithmically-determined ground truth. (3) We

¹The term “agency” is used strictly in the sense of epistemic framing, and does not point to autonomy or “AI agents”.

evaluate a diverse suite of LLMs and localize failure modes that inform evaluation design and model training objectives.

2 Related Work

Our framework is grounded in interventionist causal theories formalized by structural causal models (Pearl, 2009), and in cognitive research showing that causal reasoning comprises multiple component operations rather than a single unitary skill (Griffiths and Tenenbaum, 2009; Tenenbaum et al., 2011; McCormack et al., 2016). An extended review is in Appendix A.

Causal reasoning evaluation in LMs. Recent benchmarks evaluate causal reasoning using structured tasks spanning intervention and counterfactual queries (Wang, 2024; Chi et al., 2024; Jin et al., 2023; Komanduri et al., 2025), including interactive evaluation as in CausalLink (Feng and Rudzicz, 2025), the work closest related to ours. However, aggregate scores may obscure which reasoning components drive success or failure (Ribeiro et al., 2020; Kim et al., 2025). We address this by decomposing performance into primitive-level diagnostics for intervention-based causal reasoning in LLMs.

Diagnostic and agentic evaluation. Diagnostic and reliability-oriented evaluation is an active area, probing specific capabilities and failure modes across tasks and scales (Ribeiro et al., 2020; Yang et al., 2024; Kim et al., 2025; You et al., 2024; Srivastava et al., 2023; Liang et al., 2023; Wang et al., 2024). Agentic evaluation has similarly grown, with benchmarks analysing LLM behaviour under sequential interaction (Liu et al., 2024; George et al., 2024; Mohammadi et al., 2025). Our Agency Sensitivity primitive extends this line by isolating epistemic role framing from planning ability under information-equivalent evidence.

3 Methods

We propose *CausalPrism*², a configurable framework for evaluating qualitative causal reasoning in LLMs, grounded in both interventionist causal theory and cognitive models of causal inference (Figure 1). On these accounts, causal reasoning is not unitary — it comprises multiple operations including hypothesis evaluation, inferring effects from interventions, and belief revision (Griffiths

and Tenenbaum, 2009; Pearl, 2009; Meder et al., 2010; McCormack et al., 2016). We ground our decomposition in prior theoretical studies, and introduce four **primitives**: Causal Prior, Forward State Tracing, Belief Updating, and Agency Sensitivity.

For each primitive, we specify a *primitive task*: a normatively-defined input-output mapping. We ground all *primitive tasks* in explicitly defined deterministic causal systems represented as directed acyclic graphs (DAGs). A *test case* is one concrete instance of a task. Each *test case* is generated from a causal graph $G = (V, E)$, where V denotes variables and E denotes directed causal relations. Each variable X_i assumes a binary state (active/inactive). The system dynamics in which interventions act as transient exogenous inputs that do not override internal causal structure follow CausalLink (Feng and Rudzicz, 2025).

In this work, we include all canonical DAGs of two or three variables (Figure 1) for controlled analysis of model behaviours, but the framework itself can be scaled to more variables and more complex structures. Prior work (Feng and Rudzicz, 2025) shows that competence in minimal, 2-3 variable settings marks the onset of causal reasoning ability, and that larger-system failures often reflect breakdowns already present at this scale. By pairing the framework with a set of canonical causal structures and three narrative domains of differing semantic richness (Section 3.5), we synthesize a benchmark dataset and evaluate a wide range of language models to illustrate *CausalPrism*’s diagnostic value. Example 1 walks through a set of test cases from all four primitives on a shared causal structure.

3.1 CAUSAL PRIOR

Motivation and Targeted Ability. When causal structure is unknown, observations constrain which structural hypotheses remain admissible. This primitive tests whether a model can determine the evidential status of an observation with respect to a proposed directed causal claim, without access to the underlying graph (Steyvers et al., 2003; Griffiths and Tenenbaum, 2009; Pearl, 2009).

Task Construction. For each causal structure, we enumerate all valid variable state configurations. Under each configuration, we enumerate all candidate directed causal claims between all ordered variable pairs. Each test case features the variable states as a snapshot (*the observation*) and a proposed causal claim, but not the causal graph itself.

²Code and datasets available at <https://github.com/JinyueF/CausalPrism>.

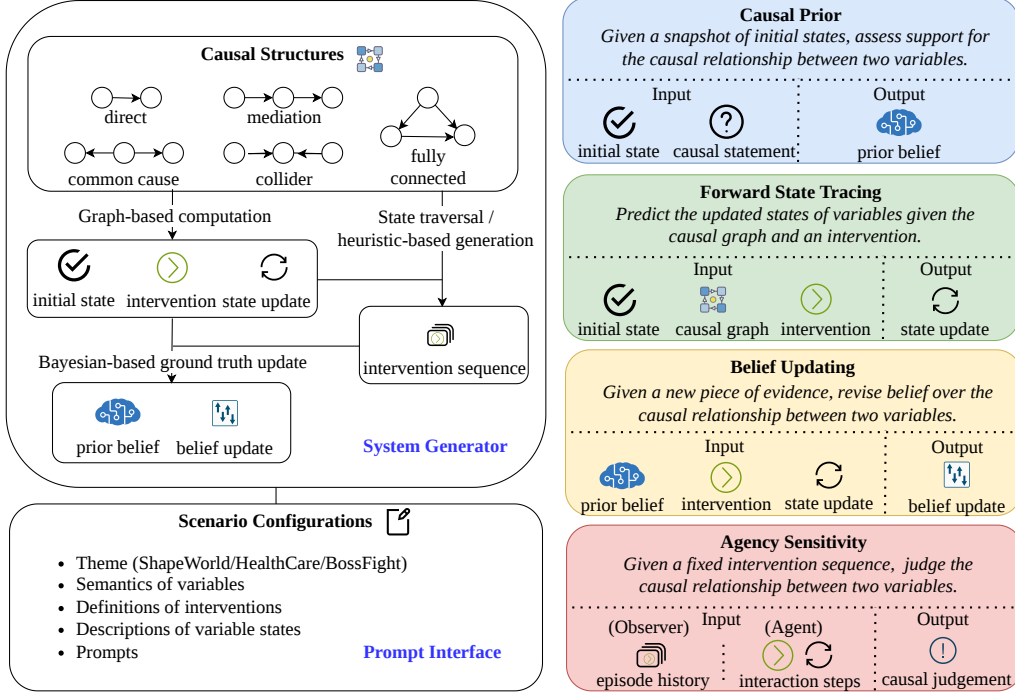


Figure 1: **Diagnostic framework overview.** We generate deterministic DAG-based causal systems (2–3 variables) spanning five canonical motifs and instantiate them across three narrative domains. The framework is customizable and scalable both in terms of underlying structures and scenario narratives.

A model must classify the informational value of the observation relative to the claim into one of four categories: **impossible** (the observation contradicts the claim); **plausible** (the observation is consistent with and supports the claim); **irrelevant** (the observation does not constrain the claim); **certain** (the observation deductively entails the claim).

The label *certain* is normatively unreachable from a single snapshot and is included to detect overconfident assertions. Both *plausible* and *irrelevant* are structurally consistent with the candidate relation but differ in their statistical informativeness: *plausible* observations are more likely under the candidate relation than under its absence, while *irrelevant* observations are equally likely either way. When the presumed cause is active and the presumed effect inactive, the observation contradicts the hypothesis under deterministic semantics and is labelled *impossible*. We additionally report a relaxed accuracy score where *irrelevant* is considered correct for *plausible* cases to accommodate strictly deductive interpretation of the problem.

3.2 FORWARD STATE TRACING

Motivation and Targeted Ability. A core component of interventionist reasoning is computing the

consequences of a manipulation within a known causal structure (Pearl, 2009). This primitive evaluates whether a model can (i) correctly interpret a provided causal graph and (ii) accurately propagate state changes through it under a given intervention.

Task Construction. For each causal structure, we enumerate all valid pre-intervention states, iterate through all non-trivial interventions (i.e., interventions that change the target variable’s state), and compute the ground truth post-intervention states using graph propagation. Each test case provides the model with the full causal graph, the initial state, an intervention applied to a target variable, and the activation rules. Interventions are transient inputs: activating interventions introduce exogenous activation; deactivating interventions remove endogenous activation but cannot override activation sustained by an active parent. The model must determine the final states of all variables.

3.3 BELIEF UPDATING

Motivation and Targeted Ability. Beyond predicting intervention outcomes under known causal relations, causal reasoning requires revising structural hypotheses in light of experimental evidence (Eberhardt and Scheines, 2007). This primitive

Running example (ShapeWorld). Shapes move under deterministic, acyclic causal rules with no hidden confounders: a shape moves if it is directly intervened on or if a parent shape moves. Interventions are transient inputs that do not override the internal rules — *Move* applies a force to initiate movement; *Stop* removes a shape’s momentum, but cannot hold a shape static if an **active parent** is moving it. The example below uses the graph $\text{square} \rightarrow \text{circle} \rightarrow \text{triangle}$.

Causal Prior. Given an observation ($\{\text{square, circle, triangle}\}$ moving) and a candidate claim ($\text{square} \rightarrow \text{circle}$), classify the claim as impossible / plausible / irrelevant / certain. The causal graph is withheld for this task. *Ground truth:* plausible.

Forward State Tracing. Given the full graph, an all-static initial state, and the intervention *Move square*, predict every shape’s final state. *Ground truth:* all shapes moving.

Belief Updating. Given prior beliefs over $\text{square} \rightarrow \text{circle}$ and $\text{circle} \rightarrow \text{square}$, and the outcome of *Move square* (square and circle move; triangle stays static), label each direction strengthen / contradict / no impact. *Ground truth:* $\text{square} \rightarrow \text{circle}$ strengthen; $\text{circle} \rightarrow \text{square}$ contradict.

Agency Sensitivity. The same intervention sequence (*Stop square*, then *Move square*) and causal question (does square’s movement cause circle’s?) are presented either all at once (OBSERVER) or executed step-by-step with interactive state updates (AGENT); the causal judgment should be invariant. *Ground truth:* yes.

Example 1: A worked ShapeWorld example illustrating all four primitives on a single shared causal structure. Each primitive is shown with its task inputs and normative ground truth, demonstrating how the same causal system instantiates four distinct diagnostic tasks.

evaluates whether a model updates prior qualitative beliefs about a directed relation in the appropriate direction given interventional evidence.

Task Construction. Each test case provides the model with: (i) a pre-intervention observation, (ii) a prior qualitative belief over two competing directed relations between a pair of variables, (iii) a single intervention applied to one variable, and (iv) the resulting post-intervention observation. The prior is supplied as a given input and is not inferred from the preceding observations, decoupling the evaluation of belief revision from prior formation. The model must determine, for each directional hypothesis, whether the new evidence *strengthens*, *contradicts*, or has *no impact* on the prior belief.

Normative Ground Truth. To generate test cases, we simulate intervention trajectories with a breadth-first traversal that enumerates all reachable graph states, followed by a depth-first walk over this state-transition graph. The yielded structured sequence of intervention–observation pairs

are used to compute belief updates. Ground-truth belief updates are computed using a simple Beta–Bernoulli model over directed edges. For each ordered pair of variables (i, j) , we maintain a Beta prior over the existence of a directed relation: $\theta_{ij} \sim \text{Beta}(\alpha_{ij}, \beta_{ij})$.

Qualitative Mapping. Our evaluation uses only the *direction* of belief change, avoiding conflation with numerical computation and remaining robust to alternative update rules. Priors are presented qualitatively as *plausible* ($\mathbb{E}[\theta_{ij}] > 0.5$), *implausible* ($\mathbb{E}[\theta_{ij}] < 0.5$), or *neutral*. Update labels combine prior polarity with the direction of change: *strengthen* if the posterior shifts further in the prior’s direction, *contradict* if it shifts against it, and *no impact* if there is no change (with neutral priors labeled solely by increase vs. decrease). Crucially, *strengthen* and *contradict* are defined relative to the supplied prior belief, not simply relative to the causal edge itself. I.e., evidence supporting an edge strengthens a plausible prior about that edge but contradicts an implausible prior. For each pair of non-neutral directed edges, we enumerate all four plausible/implausible assignments over the two directions; neutral priors are passed through unflipped. This deliberately includes cases in which the supplied prior is incorrect (so that the normatively correct behaviour is to revise against the given prior), and it improves the balance of the ground truth label distribution, which trajectory-true priors alone would skew.

3.4 AGENCY SENSITIVITY

Motivation and Targeted Ability. Causal reasoning may be influenced by epistemic framing independently of structural information. Research in developmental psychology shows that active intervention and passive observation can yield different causal attributions even when informational content is equivalent (Sobel and Legare, 2014; McCormack et al., 2016); yet from a formal standpoint, rational inference should depend only on available evidence, not on one’s epistemic role. This primitive tests whether causal judgments remain invariant under informationally equivalent but role-differentiated formulations. The term *agency* refers strictly to task framing — whether the model executes prescribed interventions or observes a history of episodes — and carries no implication of autonomy, intention, or goal-directed behaviour.

Task Construction. Each test case uses a

paired design with identical causal structure, intervention sequence, and causal question. In the **observer** condition, the model receives the complete intervention sequence and resulting system states, then answers a query about a directed causal relation. In the **agent** condition, the model executes the same sequence step-by-step in a conversational format, receiving the updated system state after each intervention before proceeding; it answers the same question after all steps are complete. The intervention sequence is computed using a graph-aware heuristic that iteratively deactivates active ancestors of the cause and effect variables until both are inactive, then activates the cause to probe its downstream effect (Appendix B.2). The design guarantees informational equivalence between conditions and isolates epistemic role framing from planning ability. We measure the comparative performance across the two conditions.

3.5 Narrative Domains

Design Principle: Isolation from Parametric Knowledge. A causal reasoning benchmark must ensure a model’s answer derives from the provided structures and evidence, not memorized facts from pretraining. Prior work isolates reasoning from recall by substituting variables with anti-factual counterparts (Jin et al., 2023; Roewer-Després et al., 2025), stripping variables’ semantic meanings (Jin et al., 2023), or constructing fictional scenarios (Srivastava et al., 2023). Instead, we enforce one unified constraint that allows us to control the semantic richness of the variables and preserve natural coherence in our descriptions. Recognizing that parametric knowledge becomes confounding when it renders one causal direction trivially impossible, we choose our variables based on one requirement: given two variables X and Y , both $X \rightarrow Y$ and $Y \rightarrow X$ must be conceptually coherent.

Narrative Domains and Initiation. Following the design principle above, we introduce three narrative domains with varying semantic richness. Each narrative domain is a fixed set of variables mapped one-to-one onto the nodes of the causal graphs; the same graph structures are instantiated across all three. The domain names (*ShapeWorld*, *HealthCare*, *BossFight*) are thematic labels used for convenient reference when reporting results; they indicate only the vocabulary from which variable names are drawn and are not intended as faithful representations of those real-world domains. *ShapeWorld* is adopted from CausalLink (Feng and

Rudzicz, 2025) and uses abstract, semantically neutral variables, trivially enforcing the directional-non-triviality constraint. *HealthCare* uses health-themed variables, constructed so that neither causal direction is decidable from domain knowledge. *BossFight* uses variables from a fictional action-game setting that are semantically structured but do not correspond to any real-world causal system. Table 1 lists the variables, states, and interventions for each domain.

Domain	Variables	States / Interventions
ShapeWorld	shape movement	moving/static; move/hold
HealthCare	EMR software adoption; physician screen time; diagnostic errors	high/low; increase/decrease (regime switches)
BossFight	character buffs; debuffs; environmental effects	on/off; activate/deactivate

Table 1: Domain configurations. Each domain is a fixed set of variables mapped one-to-one onto causal-graph nodes (Section 3.5).

3.6 Construct Validity.

CausalPrism’s construct validity rests on label correctness and distinctness of the primitives. Every label is a deterministic function of the input scenario under DAG semantics. It carries no annotation subjectivity and admits no inter-rater disagreement; we release the full generation logic and code so any label can be independently recomputed. Empirically, the primitives are not merely a single capability proxy: Claude 4.6 Sonnet obtains the highest Belief Updating score of any model, 0.613, while trailing GPT-5.2/GPT-5-mini on Causal Prior and Forward State Tracing (Table 2). Conversely, GPT models approach a ceiling on Prior/Tracing but fall to 0.517-0.531 on Updating. Agency framing gap is nearly uncorrelated with Prior, Tracing, and Updating, respectively ($r = -0.116, -0.091$, and $+0.018$), further supporting its status as a useful diagnostic probe; see Appendix E.

3.7 Evaluation Setup

Dataset Statistics. For each causal structure, we enumerate all causally valid initial setups and expand them into primitive-specific test cases. CAUSAL PRIOR and AGENCY SENSITIVITY each yield 114 cases per narrative domain (one per ordered cause–effect pair per setup); FORWARD

STATE TRACING yields 60 cases (one per non-trivial intervention per setup); BELIEF UPDATING yields 701 cases, as each setup is expanded across multiple intervention steps and all variable pairs at each step, with prior combinations enumerated as described in Section 3.3. This yields a total of 989 test cases per narrative domain, and 2,967 test cases across all three domains. Ground-truth labels are distributed as follows: CAUSAL PRIOR: impossible / plausible / irrelevant / certain = 18% / 32% / 50% / 0%; FORWARD STATE TRACING: activation intervention / deactivation intervention = 50% / 50%; BELIEF UPDATING: strengthen / contradict / no impact = 26.2% / 28.0% / 45.7%; AGENCY SENSITIVITY: yes / no = 41% / 59%.

Prompt Construction. We design and refine prompts through iterative experimentation to ensure observed model shortcomings are not artifacts of prompt design. In each iteration, we identify failure cases attributable to semantic ambiguity in model outputs and revise instructions to eliminate these ambiguities.³ To better isolate reasoning from instruction-following, we deliberately avoid eliciting predefined step-by-step chains of thought (full templates and a revision log in Appendix F).

Models and Configuration. We evaluate 14 open-source and frontier LLMs spanning multiple parameter scales and model families; full details of model selection, deployment, and inference configuration are provided in Appendix B.1.

4 Results

We report pooled accuracies across three domains in Table 2 and domain-specific scores in Appendix C.1. As illustrated by Figure 2, BELIEF UPDATING is the dominant bottleneck and the main separator between strong and weak models⁴. Pooled updating accuracy is $33.9\% \pm 15.3\%$, far below priors and tracing (Table 2). Within open-source model families, CAUSAL PRIOR and FORWARD STATE TRACING performance generally increases with parameter scale. BELIEF UPDATING does not follow this pattern consistently: Qwen 2.5-70B (0.263) scores lower than both Qwen 2.5-32B (0.294) and Qwen 2.5-14B (0.290), suggesting that scale alone does not reliably improve belief revi-

³One author manually verified a stratified subset (20% of the test cases) to ensure construction soundness.

⁴JSON parsing failures that affected Claude-Haiku, DeepSeek 8B, Llama 8B, and Llama 3B are rescued by regex-based extractions. Less than 0.4% of the test cases remain unparsable and are negligible.

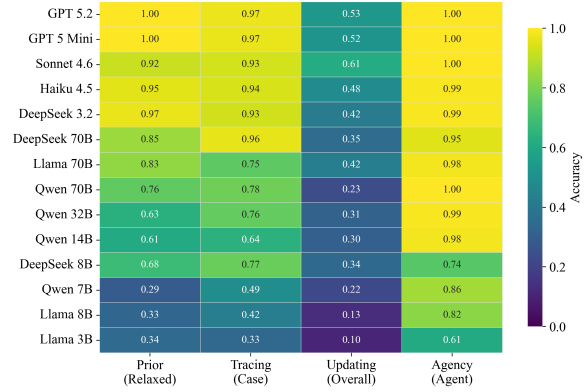


Figure 2: **Performance landscape across primitives.** Heat map shows pooled accuracies across the three domains for four headline metrics: relaxed accuracy for Causal Prior, case-level accuracy for Forward State Tracing (Case), accuracy for Belief Updating, and agent-version accuracy for Agency Sensitivity.

sion ability. Observer Agency accuracies saturate in most models, but the framing effects are visible away from the performance ceiling.

To examine how behaviour varies with overall performance, we stratify models by a composite score computed from PRIOR (relaxed), TRACING (case-level), and UPDATING. The top group comprises Claude 4.6 Sonnet, GPT-5.2, GPT-5-Mini, and DeepSeek-v3.2-Chat; the bottom group comprises Llama 3 (3B, 8B) and Qwen 2.5 (7B, 14B).

4.1 Behaviour and Failure Mode Analysis

Causal Prior. Over-rejection is the dominant failure mode across all domains, accounting for 91.8%, 85.4%, and 67.8% of errors in ShapeWorld, BossFight, and HealthCare respectively. The *irrelevant* label is consistently the hardest (mean error rates: 0.500, 0.580, 0.395), while *impossible* is the easiest (0.116, 0.048, 0.003). The label *certain* yields zero cases, confirming that premature causal conclusion is not a prevailing reasoning flaw.

Weaker models collapse toward rejection: they fail on both *plausible* and *irrelevant* cases by predicting *impossible*. Qwen 2.5-7B fails on 100% of *plausible* cases in ShapeWorld and 75.0% (27/36) in HealthCare; Llama 3-8B fails on 83.3% in ShapeWorld. These models treat observations as either conclusive or uninformative, with no intermediate evidential status.

Stronger models exhibit a qualitatively different failure concentrated in HealthCare: *irrelevant* error rates spike from near-zero in BossFight to 59.7% for Claude 4.6 Sonnet, 22.8% for DeepSeek-

Model	Prior		Tracing		Updating	Agency		
	strict	relaxed	case-level	variable-level	overall	observer	agent	gap
GPT-5.2	0.982	<u>0.997</u>	0.972	0.990	<u>0.531</u>	<u>0.997</u>	1.000	-0.003
GPT-5-mini	<u>0.971</u>	1.000	<u>0.967</u>	<u>0.985</u>	0.517	<u>0.997</u>	<u>0.997</u>	0.000
Claude-4.6-Sonnet	0.848	0.915	0.933	0.966	0.613	1.000	1.000	0.000
Claude-4.5-Haiku	0.901	0.953	0.939	0.962	0.479	<u>0.997</u>	0.991	0.006
DeepSeek-R1-Llama-8B	0.547	0.681	0.772	0.889	0.284	0.975	0.738	0.236
DeepSeek-R1-Llama-70B	0.775	0.851	0.961	<u>0.985</u>	0.376	<u>0.997</u>	0.953	0.044
DeepSeek-v3.2-Chat	0.947	0.968	0.928	<u>0.964</u>	0.416	1.000	0.988	0.012
Qwen2.5-7B	0.284	0.287	0.489	0.753	0.177	0.909	0.863	0.047
Qwen2.5-14B	0.579	0.608	0.639	0.845	0.290	0.980	0.982	-0.003
Qwen2.5-32B	0.632	0.635	0.761	0.877	0.294	<u>0.997</u>	0.988	0.009
Qwen2.5-70B	0.696	0.757	0.778	0.885	0.263	0.974	<u>0.997</u>	-0.023
Llama3.2-3B	0.275	0.342	0.333	0.640	0.095	0.685	0.608	<u>0.077</u>
Llama3.1-8B	0.316	0.330	0.421	0.688	0.101	0.779	0.825	-0.046
Llama3.3-70B	0.784	0.833	0.750	0.900	0.308	0.991	0.982	0.009

Table 2: **Pooled performance across domains.** For **Causal Prior**, we report strict and relaxed accuracy; for **Forward State Tracing**, case-level and variable-level accuracy; for **Belief Updating**, overall accuracy; for **Agency Sensitivity**, observer and agent accuracy and the framing gap. Accuracies are averaged over BossFight, HealthCare, and ShapeWorld. *gap* is observer minus agent. The highest in column is **bold**; the second highest is underlined.

v3.2-Chat, and 12.3% for GPT-5-mini. Inspection of model reasoning reveals the mechanism: under identical observations and causal structures, swapping which variable appears as cause versus effect causes models to switch from *impossible* to *plausible* without any change in the structural ground truth. Models acknowledge the same evidence but reach different conclusions, suggesting parametric preference possibly related to their world knowledge (see Appendix D.1 for paired examples).

Forward State Tracing. A consistent case-variable accuracy gap across all domains (mean 11.3–13.1%, maximum 30–33 points) indicates that models frequently trace part of the causal graph but miss at least one variable state. The gap is near-zero for top-group models (2.6%) but large for the bottom group (26.1%). Inactive states are systematically harder to predict than active ones (mean difference = 9.9%, 9.6%, and 3.9% in BossFight, HealthCare, and Shapeworld, respectively).

Belief Updating. As the most dominant bottleneck, Belief Updating reveals two distinct failure profiles across model tiers. A label-wise accuracy analysis across different tiers of overall performance (Figure 3) indicates that *contradict* is the most discriminative label. Mid- and bottom-tier models underperform across all label categories, but particularly lack the ability to accommodate contradictory evidence. In contrast to weaker models’ over-rejection tendency in Causal Prior, results on Belief Updating suggest representation or

reasoning failure beyond a tendency to associate rejection-based semantics to causal judgement.

Frontier API models exhibit much stronger handling of *contradict* cases but fail for structurally different reasons. The update labels are inherently relational — *strengthen* and *contradict* describe how evidence moves a belief. Frontier models often fail to condition on the prior beliefs and resort to superficially reasoning about the causal statement itself. Their outputs are heavily anchored by the prior labels, predicting *strengthen* in 30–43% cases with *plausible* belief but only 3–16% cases with *implausible* belief. Additionally, reasoning may become erroneous when it requires more than one hop. For example, models also inconsistently apply acyclicity: when evidence supports one direction of the causality, the implied contradiction of the opposite is missed in 22–33% of cases. We provide failure examples in Appendix D.2 and evidence against insufficient prompting in Appendix D.3.

Agency Sensitivity. Agency framing effects are most visible away from the performance ceiling. Two patterns stand out. First, framing sensitivity is family-dependent: DeepSeek shows the largest and most consistent observer advantage (mean gap 9.7%, positive in all domains), while Llama and Qwen exhibit smaller mean gaps with mixed directions. Second, observer competence is generally observed even in smaller models. The confusion breakdown further suggests that framing effects are primarily due to agent-specific failures. Finally,

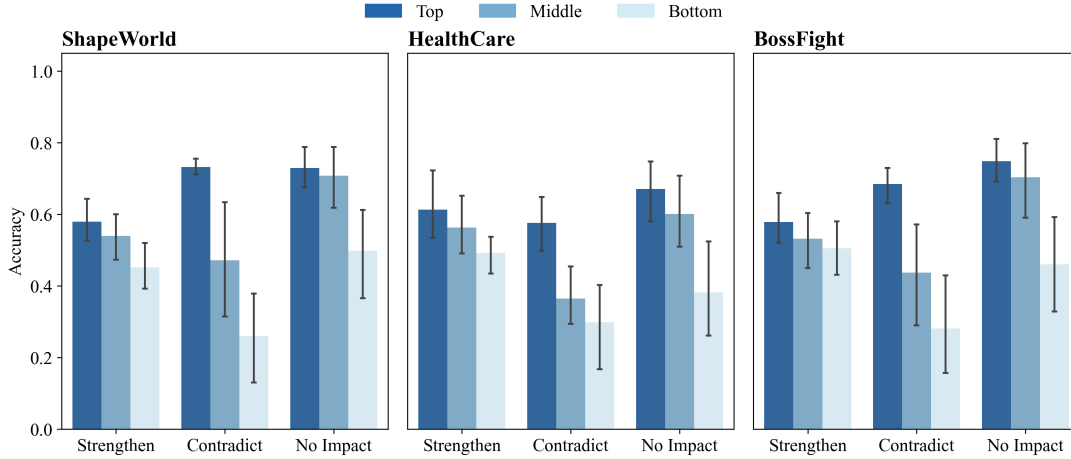


Figure 3: **Belief updating analysis by label types and performance tiers.** Accuracy on *strengthen*, *contradict*, and *no impact* is shown for each domain. Models are stratified into Top, Middle and Bottom groups following the performance based-stratification in Section 4. Error bars indicate variability within each performance group.

the framing gap is largely distinct from the other primitives; for instance, its correlations with PRIOR and UPDATING are near zero. This suggests that role-conditioned framing sensitivity is a distinct diagnostic axis rather than a byproduct of general reasoning accuracy.

4.2 The Effects of Semantic Richness

Holding causal structure fixed and varying only the semantic richness of the variables reveals two contrasting effects of domain knowledge on causal reasoning. Semantic context can make causal structures more readily accessible, but it can also increase the tendency to form judgments without grounding them in the observed evidence.

Causal Prior is the most domain-sensitive primitive. In addition to observations in Section 4.1, we quantify domain sensitivity via the relaxed-strict accuracy gap: the rate at which models default to *irrelevant* on cases whose ground truth is *plausible*. This gap is near-zero in ShapeWorld (mean 0.006), where semantic neutrality forces reliance on the provided snapshot. BossFight raises strict Prior accuracy above ShapeWorld for 8 of 14 models (mean delta +0.085): structured fictional semantics make priors more accessible without triggering associations strong enough to override snapshot-grounded reasoning. HealthCare reverses this: the relaxed-strict gap reaches a mean of 0.095 (maximum 0.211, DeepSeek-R1-8B), indicating that real-world semantic associations increase the tendency to misclassify informative observation. The failure mechanisms underlying this domain gradient are analyzed in Section 4.1. **Belief**

Updating shows a parallel pattern where HealthCare appears to be the most challenging scenario.

Forward State Tracing is modestly more accurate in richer domains. Case-level accuracy is higher in both HealthCare and BossFight than in ShapeWorld for a majority of models ((SW→HC: mean +0.029, 7/14 positive; SW→BF: mean +0.045, 8/14 positive)). For a structural task requiring only graph propagation, semantic context may provide scaffolding for interpretation.

Agency framing sensitivity is domain-invariant. The observer-agent gap shows near-zero mean deltas across all three domain comparisons, confirming that framing sensitivity is not modulated by semantic richness.

5 Conclusion

We presented an LLM evaluation framework that decomposes intervention-based reasoning into four primitives: CAUSAL PRIOR, FORWARD TRACING, BELIEF UPDATING, and AGENCY SENSITIVITY. By grounding tasks in explicit deterministic DAG systems and varying only the narrative semantics across three domains, we obtained precise normative labels and isolated distinct failure modes.

Across a diverse set of models, we find that belief updating is the dominant bottleneck: models that perform strongly on priors and forward tracing can still fail to revise beliefs appropriately. Updating errors exhibit two structurally different patterns, marking a general lack of causal understanding in weaker models and a prevailing failure to reason about revision in stronger models. We also

find that Causal Prior is the most domain-sensitive primitive, with systematic over-rejection and under-commitment behaviours that vary with both domain and model capability. Finally, agency framing effects are small for saturated frontier models but diagnostic for non-saturated models, where sensitivity clusters by family and primarily appears as selective degradation under agent framing. Our results argue against treating “causal reasoning” as a single end-to-end capability, and motivate targeted evaluation and training focused on belief revision under contradiction, calibration of prior judgments, and invariance to epistemic framing.

6 Limitations

6.1 Controlled Synthetic Setting

Our benchmark is instantiated in small deterministic structural causal systems and realized through tightly controlled narrative templates. This design provides explicit normative ground truth and enables clean analysis of specific reasoning primitives, but it also makes the evaluation substantially simpler and more scaffolded than most real-world causal reasoning problems, which typically involve stochastic relations, latent confounding, partial observability, richer semantics, and larger interacting structures. The intervention sequences in Agency Sensitivity test cases are generated with a graph-aware heuristic and provide easily accessible information to draw causal conclusions; consequently, the reasoning difficulty is minimal: even perfect accuracy in one or both conditions does not by itself imply robust causal-reasoning ability.

Templated synthesis may also introduce benchmark-specific artifacts: because the tasks are templated, some apparent successes may reflect sensitivity to surface regularities or other task-specific cues, while some failures may weaken in richer settings where world knowledge and discourse coherence provide additional support.

6.2 Scope of the Framework

The four primitives in our framework are theoretically motivated but not exhaustive. In particular, models are not required to identify relevant variables, infer the graph, or determine informative interventions in our primitive tasks. Important aspects of causal reasoning also remain outside the present scope, including latent confounding, probabilistic inference, and counterfactual reasoning.

A related limitation is that language models are

stateless at inference time rather than persistent learners with evolving internal beliefs. Our results therefore speak to behavioural performance on formal task structures, not to human-like causal cognition in a broader sense.

6.3 Evaluation and Model Coverage

Our evaluation relies on tasks with precise normative answers and therefore emphasizes behavioural accuracy over process-level evidence. As a result, it is better suited to characterizing performance outcomes than to identifying the mechanisms underlying them. The model set is constrained and is not representative of all model families, scales, or training regimes, so our conclusions apply most directly to the models and inference settings evaluated here.

7 Ethical Considerations

This study raises limited direct ethical concerns. The benchmark is constructed from synthetic scenarios and does not involve human participants, personal data, or real clinical or organizational records.

The main ethical consideration is responsible interpretation of the results, especially for models that appear top-performing under our evaluation. Our framework is not a test of reliable causal competence in naturalistic or high-stakes environments. Accordingly, benchmark performance should not be treated as evidence that a model is suitable for deployment in consequential domains such as healthcare, law, or policy. The benchmark should be strictly viewed as a diagnostic instrument for controlled analysis rather than a comprehensive measure of causal competence or safety.

Acknowledgements

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/#partners. Frank Rudzicz is supported by a Canada CIFAR Chair in AI.

References

- Anthropic. 2025. [Claude Haiku 4.5 system card](#). System card (PDF). Accessed: 2026-03-08.
- Anthropic. 2026. [Claude Sonnet 4.6 system card](#). System card (PDF). Accessed: 2026-03-08.

- Wei Chen, Jiahao Zhang, Haipeng Zhu, Boyan Xu, Zhifeng Hao, Keli Zhang, Junjian Ye, and Ruichu Cai. 2025. [Causal-aware large language models: Enhancing decision-making through learning, adapting and acting](#). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 4292–4300. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024. [Unveiling causal reasoning in large language models: Reality or mirage?](#) In *Advances in Neural Information Processing Systems*, volume 37, pages 96640–96670. Curran Associates, Inc.
- DeepSeek-AI. 2025. [DeepSeek-V3.2-Exp: Boosting long-context efficiency with deepseek sparse attention](#).
- Frederick Eberhardt and Richard Scheines. 2007. [Interventions and causal inference](#). *Philosophy of Science*, 74(5):981–995.
- Jinyue Feng and Frank Rudzicz. 2025. [CausalLink: An interactive evaluation framework for causal reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22313–22326.
- Sonny George, Chris Sypherd, and Dylan Cashman. 2024. [Probing the capacity of language model agents to operationalize disparate experiential context despite distraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15447–15459, Miami, Florida, USA. Association for Computational Linguistics.
- John Gkountouras, Matthias Lindemann, Phillip Lippe, Efstratios Gavves, and Ivan Titov. 2025. [Language agents meet causality – bridging LLMs and causal world models](#). In *The Thirteenth International Conference on Learning Representations*.
- Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. 2004. [A theory of causal learning in children: causal maps and bayes nets](#). *Psychological review*, 111(1):3.
- Alison Gopnik and Laura Schulz. 2004. [Mechanisms of theory formation in young children](#). *Trends in cognitive sciences*, 8(8):371–377.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Thomas L. Griffiths, Charles Kemp, and Joshua B. Tenenbaum. 2008. [Bayesian models of cognition](#). *The Cambridge handbook of computational psychology*, pages 59–100.
- Thomas L. Griffiths and Joshua B. Tenenbaum. 2009. [Theory-based causal induction](#). *Psychological review*, 116(4):661.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [DeepSeek-R1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. [CLADDER: assessing causal reasoning in language models](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Dongjun Kim, Gyuho Shim, Yongchan Chun, Minhyuk Kim, Chanjun Park, and Heuseok Lim. 2025. [Benchmark profiling: Mechanistic diagnosis of LLM benchmarks](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15635–15650, Suzhou, China. Association for Computational Linguistics.
- Aneesh Komanduri, Karuna Bhaila, and Xintao Wu. 2025. [CausalVLBench: Benchmarking visual causal reasoning in large vision-language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30660–30680, Suzhou, China. Association for Computational Linguistics.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng

- Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. [AgentBench: Evaluating LLMs as agents](#). In *The Twelfth International Conference on Learning Representations*.
- Teresa McCormack, Neil Bramley, Caren Frosch, Fiona Patrick, and David Lagnado. 2016. [Children’s use of interventions to learn causal structure](#). *Journal of experimental child psychology*, 141:1–22.
- Bjorn Meder, Tobias Gerstenberg, York Hagmayer, and Michael R. Waldmann. 2010. [Observing and intervening: Rational and heuristic models of causal decision making](#). In *The Open Psychology Journal*.
- Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. [Evaluation and benchmarking of LLM agents: A survey](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD ’25*, page 6129–6139, New York, NY, USA. Association for Computing Machinery.
- OpenAI. 2025a. [GPT-5 system card](#). System card (PDF). Accessed: 2026-03-08.
- OpenAI. 2025b. [Update to GPT-5 system card: GPT-5.2](#). Web page. Accessed: 2026-03-08.
- Judea Pearl. 2009. [Causality](#), 2 edition. Cambridge University Press.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. [Elements of Causal Inference: Foundations and Learning Algorithms](#). MIT press.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. [GPQA: A graduate-level google-proof q&a benchmark](#). In *First Conference on Language Modeling*.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- François Roewer-Després, Jinyue Feng, Zining Zhu, and Frank Rudzicz. 2025. [ACCORD: Closing the commonsense measurability gap](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3799–3829, Albuquerque, New Mexico. Association for Computational Linguistics.
- David M Sobel and Cristine H Legare. 2014. [Causal learning in children](#). *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(4):413–427.
- Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. [Causation, Prediction, and Search](#). MIT press.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. [Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*. Featured Certification.
- Mark Steyvers, Joshua B. Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum. 2003. [Inferring causal networks from observations and interventions](#). *Cognitive Science*, 27(3):453–489.
- Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. [How to grow a mind: Statistics, structure, and abstraction](#). *science*, 331(6022):1279–1285.
- Marshall Wang. 2026. [Vector Inference: Efficient LLM inference on Slurm clusters](#).
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. [MMLU-Pro: A more robust and challenging multi-task language understanding benchmark](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290. Curran Associates, Inc.
- Zeyu Wang. 2024. [CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models](#). In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151, Bangkok, Thailand. Association for Computational Linguistics.
- Jason Weston, Antoine Bordes, Sumit Chopra, and Tomáš Mikolov. 2016. [Towards AI-complete question answering: A set of prerequisite toy tasks](#). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- James Woodward. 2003. [Making Things Happen: A Theory of Causal Explanation](#). Oxford University Press.
- Zhe Yang, Yichang Zhang, Tianyu Liu, Jian Yang, Junyang Lin, Chang Zhou, and Zhifang Sui. 2024. [Can large language models always solve easy problems if they can solve harder ones?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1531–1555, Miami, Florida, USA. Association for Computational Linguistics.

Jiaxuan You, Mingjie Liu, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. [LLM-Evolve: Evaluation for LLM’s evolving capability on benchmarks](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16937–16942, Miami, Florida, USA. Association for Computational Linguistics.

Wlodek W. Zadrozny. 2026. [Challenges and opportunities in causality analysis using large language models](#). *Entropy*, 28(1).

A Extended Related Work

Interventionist semantics and structural causal models. Interventionist theories characterize causal relations in terms of how the system would respond under manipulations: roughly, X causes Y when intervening on X would change Y in an appropriate way (Woodward, 2003). Structural causal models (SCMs) formalize this idea by representing a system as a directed graph together with structural equations that specify how each variable is generated from its parents and exogenous noise (Pearl, 2009). Within this framework, causal questions are not reducible to associative patterns in observational data; instead they are defined by a family of distributions induced by interventions (and, for counterfactuals, by interventions combined with evidence about realized outcomes) (Pearl, 2009; Peters et al., 2017). SCMs thus separate (i) *statistical* information (what is observed) from (ii) *causal* information (what would happen under manipulation), and provide a normative basis for deciding which causal conclusions are licensed by which kinds of evidence.

A key implication is that causal reasoning depends on the *regime* of evidence. Observational evidence supports inferences about association and may constrain causal structure, but it generally cannot, by itself, identify causal direction without additional assumptions (Spirtes et al., 2000; Peters et al., 2017). Interventions change the data-generating process by surgically fixing a variable, which breaks certain dependencies and thereby disambiguates structures that are observationally equivalent (Pearl, 2009; Spirtes et al., 2000). In turn, the same observed outcome can carry different implications depending on whether it arose naturally or under manipulation, making it essential to track the observation–intervention distinction.

Multiple inference modes: prediction, intervention, and counterfactuals. SCMs also clarify that “causal reasoning” encompasses different query types with different computational requirements. Predicting from observation can often be cast as conditioning (e.g., $P(Y \mid X = x)$), whereas predicting consequences of an intervention requires computing an interventional distribution (e.g., $P(Y \mid do(X = x))$) that generally differs from observational conditioning when confounding is present (Pearl, 2009). Even in simplified settings where confounding is absent and mechanisms are deterministic, these modes remain conceptually

distinct and impose different normative constraints on what can be concluded from a given evidence pattern.

Structure uncertainty and causal discovery.

Causal discovery frameworks treat structure learning as selecting among candidate graphs consistent with observed conditional independences (under assumptions such as causal Markov and faithfulness), and show how interventions can dramatically reduce ambiguity by ruling out equivalence classes (Spirtes et al., 2000; Peters et al., 2017). Importantly, this view already decomposes causal competence: one must represent and compare candidate structures (hypothesis evaluation), reason about implications of interventions within a hypothesized structure (forward causal computation), and revise the hypothesis space as new evidence arrives (belief updating) (Spirtes et al., 2000; Peters et al., 2017).

Cognitive models: hypothesis evaluation over structured representations. Cognitive science offers a complementary but aligned picture. Bayesian approaches model learners as maintaining probabilistic beliefs over hypotheses about latent structure and updating those beliefs in light of data (Griffiths et al., 2008). In causal learning, the hypotheses are often causal graphs or related structured representations, and inference involves evaluating competing structures against evidence to infer both causal direction and mechanism (Gopnik et al., 2004; Griffiths and Tenenbaum, 2009). This framing naturally yields a separation between (i) *prior expectations* over plausible structures and mechanisms and (ii) *likelihood-based updating* from observations and interventions. Theory-based causal induction further emphasizes that causal inference is constrained by structured prior knowledge about plausible causal relations and mechanisms, rather than being reducible to correlational heuristics (Griffiths and Tenenbaum, 2009). More broadly, accounts of cognition as probabilistic inference over structured hypothesis spaces argue that learning and reasoning proceed by abstraction and compositional structure, supporting systematic generalization across contexts (Tenenbaum et al., 2011).

Interventions in human causal learning: doing vs. seeing. Empirical work in cognitive and developmental psychology also supports the centrality of interventions. People can infer causal structure

from observation alone to some degree, but performance improves substantially when they can intervene, and human learners select interventions that are informative for disambiguating competing hypotheses (Steyvers et al., 2003). Developmental findings show that children use interventions to discover causal structure, suggesting that sensitivity to the evidential status of interventions emerges early and plays a distinct role from passive observation (McCormack et al., 2016). Reviews of causal learning in children emphasize that causal knowledge is built through an interplay of observation, intervention, and explanation, and that multiple interacting processes contribute to causal competence (Sobel and Legare, 2014; Gopnik and Schulz, 2004).

Causal reasoning evaluation in language models.

A growing body of work evaluates causal and counterfactual reasoning in LMs using structured query families derived from formal causal formalisms. Benchmarks such as CLADDER translate causal graphs and queries into natural language to probe associational, interventional, and counterfactual inference (Jin et al., 2023). Other datasets broaden coverage across domains and task formats, including textual, mathematical, and code-based causal reasoning (Wang, 2024), and targeted analyses that interrogate whether apparent causal competence reflects robust reasoning or superficial pattern matching (Chi et al., 2024). Multimodal extensions similarly test whether large vision-language models can infer structure or predict intervention and counterfactual outcomes when causal information is presented visually (Komanduri et al., 2025). These efforts provide valuable standardized testbeds, but they typically report end-to-end performance over heterogeneous instances, leaving open which specific reasoning operations (e.g., intervention propagation vs. hypothesis discrimination) drive errors.

Interactive and structured elicitation methods offer a complementary perspective. CausalLink, for example, evaluates causal reasoning through interactive questioning designed to elicit and probe causal understanding more directly than single-shot QA (Feng and Rudzicz, 2025). Our work shares the motivation to move beyond static, aggregate evaluation, but differs in emphasis: rather than adding interaction primarily to improve elicitation, we use controlled decompositions into primitives to attribute performance to distinct causal operations under matched causal structures and evidence regimes.

Diagnostic evaluation and the limits of aggregate benchmark scores. Beyond causal reasoning specifically, a broader evaluation literature argues that aggregate scores can obscure the underlying skill mixtures and failure modes that matter for deployment. Behavioural testing frameworks like CheckList propose systematic, capability-oriented test design to uncover brittle behaviour not reflected in average accuracy (Ribeiro et al., 2020). Dynamic evaluation platforms further emphasize that benchmark performance can be gamed or overfit, motivating evaluation regimes that evolve with models and highlight systematic weaknesses (Kiela et al., 2021). Related benchmark-auditing work argues that benchmark labels often conflate multiple competencies and that improvements in overall score may not translate to improvements in the intended ability (Kim et al., 2025). Reliability-oriented benchmarks and analyses also document inconsistencies where models may succeed on nominally harder instances while failing on easier ones, suggesting that end-to-end success can mask fragile reasoning processes (Yang et al., 2024).

Large-scale evaluation suites provide complementary breadth by aggregating diverse tasks, domains, and testing conditions. BIG-bench and HELM are representative efforts to broaden coverage and standardize reporting across models and settings (Srivastava et al., 2023; Liang et al., 2023). More targeted high-difficulty benchmarks, such as MMLU-Pro and GPQA, are designed to increase robustness and reduce contamination or shortcut learning (Wang et al., 2024; Rein et al., 2024), while task-specific suites such as bAbI historically served as controlled probes of prerequisite reasoning skills (Weston et al., 2016). Our contribution is orthogonal to breadth: we aim to increase interpretability of causal evaluation by factorizing causal competence into primitives and testing each operation under controlled semantic variation.

B Supplementary Materials for Methods

B.1 Models and Configurations

Our diverse model suite includes Llama 3 (3.2-3B-Instruct, 3.1-8B-Instruct, 3.3-70B-Instruct, Grattafiori et al., 2024), Qwen 2.5 (7B-Instruct, 14B-Instruct, 32B-Instruct, 70B-Instruct, Qwen et al., 2025), DeepSeek R1 Distilled Llama (8B, 70B, Guo et al., 2025), DeepSeek-v3.2-Chat (DeepSeek-AI, 2025), Claude 4.6 Sonnet (Anthropic, 2026), Claude 4.5 Haiku (Anthropic,

2025), GPT-5.2 (OpenAI, 2025b), and GPT-5-Mini (OpenAI, 2025a). Open source model licenses are: for Llama models, Llama Community License; for Qwen models, Apache 2.0; for DeepSeek models, MIT License.

Open-source models are deployed using the vec-inf library (Wang, 2026), released under Apache 2.0 license, on a cluster with hybrid standard compute (NVIDIA L40S 48GB) and performance compute (NVIDIA H100 SXM 80GB). The GPU usage is centrally controlled to optimize for efficiency.

We selected models to balance popularity and representativeness subject to availability and computational constraints. Notably, we selected the latest available version of Llama 3 for each model size, meaning the specific version numbers vary across the size categories. Our usage of scientific artifacts strictly follow their intended uses.

Unless otherwise specified, we follow developer-recommended settings and use defaults when available. For non-reasoning models (Llama 3, Qwen 2.5, and DeepSeek-v3.2-Chat), we set the temperature to 0.3 due to the logical nature of our tasks. For API-based reasoning models, we set the reasoning-effort parameter to medium.

B.2 Algorithm for Intervention Sequence Generation in Agency Sensitivity

We generate intervention sequences for the Agency Sensitivity task using a graph-aware procedure (Algorithm 1) that guarantees the causal query is decidable from the resulting observations. The resulting sequences are not intended to model realistic agent behaviour in the simulated environment; rather, they should be understood as an optimistically informative, “best-case” intervention history. Accordingly, this procedure is specific to our Agency Sensitivity evaluation and should not be treated as a general-purpose policy or used as the basis for training data for action-identification tasks.

C Additional Results

C.1 Full Performance Tables

The per-model performances for ShapeWorld, HealthCare, and BossFight are summarized in Table 3, 4, and 5 respectively.

C.2 Inter-Primitive Correlation Structure

Table 6 reports the full inter-primitive correlation matrix pooled across models and domains ($n =$

Model	Prior		Tracing		Updating	Agency		
	strict	relaxed	case-level	variable-level	overall	observer	agent	gap
GPT-5.2	0.991	0.991	0.950	0.983	0.518	1.000	1.000	0.000
GPT-5-mini	0.991	1.000	<u>0.967</u>	<u>0.989</u>	0.505	1.000	1.000	0.000
Claude-4.6-Sonnet	0.860	0.860	0.833	0.908	0.622	1.000	1.000	0.000
Claude-4.5-Haiku	0.904	0.930	0.850	0.897	<u>0.535</u>	0.991	0.991	0.000
DeepSeek-v3.2-Chat	0.965	0.965	0.833	0.908	0.432	1.000	0.991	0.009
DeepSeek-R1-Llama-70B	0.754	0.772	0.983	0.994	0.469	1.000	0.947	0.053
DeepSeek-R1-Llama-8B	0.553	0.570	0.717	0.879	0.298	1.000	0.722	0.278
Qwen2.5-7B	0.263	0.263	0.450	0.707	0.178	0.930	0.842	0.088
Qwen2.5-14B	0.377	0.377	0.550	0.810	0.307	0.956	0.982	-0.026
Qwen2.5-32B	0.561	0.561	0.817	0.914	0.330	1.000	0.974	0.026
Qwen2.5-70B	0.605	0.605	0.883	0.948	0.308	0.982	1.000	-0.018
Llama3.2-3B	0.237	0.254	0.333	0.638	0.106	0.775	0.602	<u>0.173</u>
Llama3.1-8B	0.281	0.281	0.417	0.672	0.100	0.798	0.904	-0.105
Llama3.3-70B	0.737	0.737	0.717	0.879	0.351	0.991	1.000	-0.009

Table 3: Per-model performance on **ShapeWorld**.

Model	Prior		Tracing		Updating	Agency		
	strict	relaxed	case-level	variable-level	overall	observer	agent	gap
GPT-5.2	0.956	1.000	0.967	0.989	<u>0.532</u>	1.000	1.000	0.000
GPT-5-mini	<u>0.939</u>	1.000	0.950	0.971	0.514	0.991	0.991	0.000
Claude-4.6-Sonnet	0.702	0.904	0.983	0.994	0.570	1.000	1.000	0.000
Claude-4.5-Haiku	0.825	0.930	0.967	0.989	0.421	1.000	0.991	0.009
DeepSeek-v3.2-Chat	0.886	0.939	0.983	0.994	0.377	1.000	0.982	0.018
DeepSeek-R1-Llama-70B	0.702	0.851	0.900	0.960	0.253	1.000	0.947	0.053
DeepSeek-R1-Llama-8B	0.570	0.781	0.783	0.885	0.233	0.961	0.778	0.183
Qwen2.5-7B	0.342	0.351	0.550	0.787	0.146	0.895	0.868	0.026
Qwen2.5-14B	0.614	0.675	0.717	0.879	0.251	0.991	0.982	0.009
Qwen2.5-32B	0.570	0.570	0.683	0.851	0.225	0.991	0.991	0.000
Qwen2.5-70B	0.649	0.816	0.733	0.862	0.174	0.939	0.991	-0.053
Llama3.2-3B	0.360	0.509	0.267	0.592	0.107	0.643	0.536	<u>0.107</u>
Llama3.1-8B	0.404	0.430	0.400	0.661	0.106	0.780	0.781	-0.001
Llama3.3-70B	0.737	0.825	0.817	0.931	0.290	0.982	0.982	0.000

Table 4: Per-model performance on **HealthCare**.

42, Pearson correlations with Benjamini–Hochberg correction).

D Case Studies for Qualitative Analysis

D.1 Causal Prior Examples

Differing behaviours on the same structural evidence

D.2 Belief Updating Examples

The following example (from Claude 4.6 Sonnet) demonstrates two failure modes:

Prior Belief: Triangle → Circle: "implausible", Circle → Triangle: "implausible"

Pre-intervention: Circle and Triangle are static.

Intervention: Move the circle.

Post-intervention: Circle and Triangle are moving.

Ground Truth of Belief Update: strengthen, contradict

(The observation supports the claim that the movement of Circle causes the movement of Triangle, strengthening the "implausible" belief from Triangle to Circle and contradicting the "implausible" belief from Circle to Triangle.)

Model Prediction: no impact, strengthen

Model Reasoning: "In State 1, both shapes are static. We intervene by moving the circle. In State 2, both circle and triangle are moving. Since the circle moved due to the direct intervention (not due to triangle), the circle's movement cannot be attributed to triangle. Therefore, **this experiment**

Model	Prior		Tracing		Updating	Agency		
	strict	relaxed	case-level	variable-level	overall	observer	agent	gap
GPT-5.2	1.000	1.000	1.000	1.000	0.543	0.991	1.000	-0.009
GPT-5-mini	0.982	1.000	0.983	0.994	0.534	1.000	1.000	0.000
Claude-4.6-Sonnet	0.982	0.982	0.983	0.994	0.648	1.000	1.000	0.000
Claude-4.5-Haiku	0.974	1.000	1.000	1.000	0.483	1.000	0.991	0.009
DeepSeek-v3.2-Chat	0.991	1.000	0.967	0.989	0.440	1.000	0.991	0.009
DeepSeek-R1-Llama-70B	0.868	0.930	1.000	1.000	0.405	0.991	0.965	0.026
DeepSeek-R1-Llama-8B	0.518	0.693	0.817	0.902	0.322	0.963	0.714	0.248
Qwen2.5-7B	0.246	0.246	0.467	0.764	0.207	0.904	0.877	0.026
Qwen2.5-14B	0.746	0.772	0.650	0.845	0.311	0.991	0.982	0.009
Qwen2.5-32B	0.763	0.772	0.783	0.868	0.328	1.000	1.000	0.000
Qwen2.5-70B	0.833	0.851	0.717	0.845	0.308	1.000	1.000	0.000
Llama3.2-3B	0.228	0.263	0.400	0.690	0.071	0.637	0.685	-0.048
Llama3.1-8B	0.263	0.281	0.445	0.731	0.096	0.758	0.789	-0.031
Llama3.3-70B	0.877	0.939	0.717	0.891	0.284	1.000	0.965	<u>0.035</u>

Table 5: Per-model performance on **BossFight**.

provides no information about whether triangle -> circle exists. However, triangle started moving after circle was moved — and since there was no other cause for triangle’s movement (no intervention on triangle, no other moving shapes), **the most parsimonious explanation is that circle caused triangle to move. This strengthens the belief in circle -> triangle.** The triangle -> circle direction is unaffected because the circle’s movement is fully explained by the intervention itself."

Failure analysis: This example showcases the two failure modes together: 1) the model failed to recognize the acyclicity of the causal structure entails the two directions are exclusive. and 2) despite recognizing that the evidence supports Circle → Triangle, it fails to identify the correct update to the prior belief.

D.3 Evidence Against Insufficient Prompting

We closely examine whether the prevalent failure to integrate prior belief in an artifact from insufficient prompting. The following observations serve as evidence against insufficient instructions.

1. **Keyword Anchor:** To examine whether models explicitly reason in terms of belief states, we searched each model’s raw reasoning output on belief updating tasks for five anchor phrases: current belief, prior belief, initial belief, plausible, and implausible. The results reveal striking variation across model families, independent of general instruction-following abilities. DeepSeek and Qwen models use at least one anchor in virtually ev-

ery response (94-100%), with plausible/implausible and current belief being the dominant terms; for instance, deepseek-70b and qwen25-14b invoke current belief in 94% and 92% of responses, respectively. Llama models occupy a middle tier (71-95%), relying primarily on plausible/implausible while using explicit belief-state terms more sparingly. The Claude models (56-63%) show moderate anchor usage driven almost entirely by plausible/implausible, rarely employing the more precise current belief or initial belief framings (11%). The GPT models diverge most sharply: GPT-5-mini uses anchors in only 27% of responses, and GPT-5.2 in just 10%, suggesting these models either reason more implicitly or produce more compressed outputs that forgo explicit belief-state language. These statistics prove that frontier models bypass genuine integration of given prior beliefs for reasons other than ambiguous prompting.

2. **Explicit Reasoning about Priors:** We observe explicit reasoning traces about priors regardless of the correctness of the final answer that would be absent if there was indeed insufficient prompting. Following are example excerpts from different models:

"...this would mean that the initial belief about 'environment effects' causing 'boss debuffs' is incorrect, thus contradicting it. So, the new evidence contradicts the initial belief..." (DeepSeek 70B)

"...This is consistent with a causal link cir-

Metric A	Metric B	r	p_{adj}	Sig.
<i>Causal Prior</i>				
Prior (relaxed)	Tracing (case)	+0.887	< 0.001	***
Prior (relaxed)	Updating	+0.748	< 0.001	***
Prior (relaxed)	Agency (accuracy)	+0.637	< 0.001	***
Prior (relaxed)	Agency (gap)	-0.116	0.519	ns
<i>Forward State Tracing</i>				
Tracing (case)	Updating	+0.794	< 0.001	***
Tracing (case)	Agency (accuracy)	+0.708	< 0.001	***
Tracing (case)	Agency (gap)	-0.091	0.611	ns
<i>Belief Updating</i>				
Updating	Agency (accuracy)	+0.581	< 0.001	***
Updating	Agency (gap)	+0.018	0.951	ns
<i>Error-type specificity (Causal Prior)</i>				
Over-rejection rate	Tracing (case)	-0.829	< 0.001	***
Over-rejection rate	Updating	-0.733	< 0.001	***
Over-rejection rate	Agency (gap)	+0.071	0.696	ns
<i>Agency</i>				
Agency (accuracy)	Agency (gap)	-0.603	< 0.001	***

Table 6: Inter-primitive correlation structure pooled across 14 models and 3 domains ($n = 42$). Pearson correlations with Benjamini–Hochberg correction. Agency (accuracy) = agent-framed accuracy; Agency (gap) = observer minus agent accuracy. Over-rejection rate = rate of *impossible* predictions on non-impossible items. *** $p_{adj} < 0.001$; ns = not significant.

	Case A	Case B
Claim	High physician screentime → high diagnostic errors	High EMR usage → high diagnostic errors
Observation	Screentime: low; Errors: low	EMR usage: low; Errors: low
Ground truth	irrelevant	irrelevant
Prediction	plausible (wrong)	impossible (wrong)
Reasoning	“the observation is plausible because it suggests a scenario where low physician screentime is associated with low diagnostic errors [...] It is also not irrelevant because it provides some insight into the potential relationship by showing that when one factor is low, the other is also low.”	“both the cause (usage of a new EMR software) and the effect (diagnostic errors) are observed to be low, this contradicts the premise of the causal statement that expects a high cause to lead to a high effect.”

Table 7: Llama 3.3-70B produces opposite errors on structurally identical evidence (both cases: cause and effect absent, ground truth *irrelevant*). Case A over-accepts by invoking a real-world monotonic association; Case B over-rejects by misinterpreting the absence of the cause as a contradiction.

cle -> triangle, which contradicts the prior belief that circle -> triangle is implausible...
(Claude 4.6 Sonnet)

"We start with the prior belief that 'circle -> triangle' is implausible and 'triangle -> circle' is plausible..." (DeepSeek-v3.2-Chat)

"...which could indicate that 'diagnostic errors' might indeed influence 'usage of a new EMR software' in some way. This is contrary to the initial implausible belief, suggesting our initial belief might need revision based on this new evidence..." (Llama3 70B)

"...it must be influencing 'diagnostic errors' in some way, even though the current belief considers this implausible..." (Qwen2.5 7B)

E Discussions on Construct Validity

Because CausalPrism’s central contribution is a decomposition, two questions must be settled before its diagnostic claims can be trusted: whether the ground-truth labels are correct, and whether the four primitives are genuinely distinct measurement axes rather than one capability under four names.

Label correctness. Every label is a deterministic function of the input scenario under DAG semantics. It carries no annotation subjectivity and admits no inter-rater disagreement; we release the full generation logic and code so any label can be independently recomputed.

Within-tier rank reversals. A single-factor account makes a falsifiable prediction: within a

Algorithm 1 Intervention Sequence Generation (Agency Sensitivity)

Require: DAG G , initial states S_0 , cause node c , effect node e , step cap $K=20$

Ensure: Intervention sequence $\mathcal{A}$ (each entry stores action, pre-state, post-state)

```
1:  $\mathcal{A} \leftarrow []$ ;  $S \leftarrow S_0$ ;  $k \leftarrow 0$ 
2: while ( $S[c] = \text{ACTIVE} \vee S[e] = \text{ACTIVE}$ )  $\wedge$   $k < K$  do
3:    $k \leftarrow k + 1$ 
4:    $T \leftarrow \{u \in \text{Anc}(c) \cup \text{Anc}(e) \mid S[u] = \text{ACTIVE}\}$ 
5:   if  $S[c] = \text{ACTIVE}$  then
6:      $T \leftarrow T \cup \{c\}$ 
7:   end if
8:   if  $S[e] = \text{ACTIVE}$  then
9:      $T \leftarrow T \cup \{e\}$ 
10:  end if
11:   $R \leftarrow \{v \in T \mid \forall p \in \text{Pa}(v), S[p] \neq \text{ACTIVE}\}$ 
12:  if  $R = \emptyset$  then
13:    break
14:  end if
15:   $r \leftarrow \min(R)$ 
16:  Record ( $\text{do}(r \leftarrow \text{INACTIVE})$ ,  $S$ ) in  $\mathcal{A}$ 
17:   $S \leftarrow \text{SIMDEACTIVATE}(G, S, r)$ ; set recorded post-state to  $S$ 
18: end while
19: Record ( $\text{do}(c \leftarrow \text{ACTIVE})$ ,  $S$ ) in  $\mathcal{A}$ 
20:  $S \leftarrow \text{SIMACTIVATE}(G, S, c)$ ; set recorded post-state to  $S$ 
21: return  $\mathcal{A}$ 
```

fixed performance stratum, models should order consistently across primitives. The data violate this. Claude 4.6 Sonnet attains the highest Belief Updating accuracy of any model (0.613) yet trails both GPT models on Causal Prior and Forward State Tracing, while the GPT models approach ceiling on Prior and Tracing but fall to 0.52–0.53 on Updating. Agency framing is orthogonal in a stronger sense: the observer–agent gap correlates near zero with every other primitive across all three domains. Prior, Tracing, and Updating do covary — expected for tasks that share basic scenario-parsing demands (Appendix C.2) — but this covariation leaves substantial structure (bottlenecks, ceilings, framing effects) that no single factor reproduces.

Agency framing sensitivity as a distinct diagnostic axis. While raw Agency accuracy co-varies with general capability, the framing gap does not:

it correlates near-zero with the other primitives across all three narrative domains (full correlation matrix in Appendix C.2). In weaker models where the framing gap is prominent, the confusion breakdown suggests that framing effects are primarily due to the observer advantage and the agent failures. Such sensitivity appears to be family-dependent: DeepSeek shows the largest and most consistent observer advantage (mean gap 9.7%, positive in all domains), while Llama and Qwen exhibit smaller mean gaps with mixed directions.

Together, these patterns show the primitives are correlated but not collapsible: ceiling effects on some primitives coexist with bottlenecks on others, and framing sensitivity dissociates from accuracy across the full model spectrum.

F Prompting Strategy

F.1 Prompt Development Summary

We iteratively refined prompts through multiple rounds of pilot experiments, focusing on removing semantic ambiguity in task instructions and label definitions while avoiding “gotcha” phrasing. The initial system prompts, specifying the rules and dynamics of the hypothetical environment, were developed according to the original CausalLink framework. Revisions primarily targeted (i) clearer formalization of the simulated world dynamics, (ii) consistent intervention semantics across primitives (interventions as transient inputs rather than overrides of causal rules), and (iii) output schemas that better reflect the intended inference operation.

Causal Prior Prompt updates progressively reduced ambiguity in label meanings and strengthened background assumptions. We clarified how an observation should be interpreted (as a snapshot drawn from a state consistent with the world’s deterministic causal rules, with probabilistic initialization across worlds), refined the task framing and label set to better distinguish contradiction vs. weak support vs. irrelevance, and renamed task elements (e.g., “hypothesis” to “causal statement”) for conceptual precision.

Belief Updating Revisions emphasized alignment with the causal-prior setup and made intervention semantics explicit. We clarified that the underlying causal structure is acyclic and unconfounded, and that interventions act as inputs that do not override internal causal rules. We also revised outputs from categorical link judg-

ments (e.g., yes/no/uncertain) to change-based update labels (e.g., *strengthen/contradict/no impact*), and relabeled prior states (e.g., *plausible/neutrallimplausible*) to reduce interpretive ambiguity.

Forward State Tracing. Prompt updates focused on improving the transparency of the simulation dynamics. We made graph semantics explicit (adjacency lists indicating directed causal influence) and stated the propagation rule directly (a node is moving iff it is intervened to move or at least one parent is moving), while reiterating that interventions are transient inputs. Later refinements further clarified the meaning of “override” to prevent confusion between overriding rules vs. overriding states.

Agency Sensitivity Revisions standardized background assumptions to match the other primitives and strengthened compliance cues in the interactive setting. We clarified the same causal-world assumptions (acyclicity, no confounders, intervention semantics) and explicitly instructed models to execute only the prescribed interventions, since the interaction protocol requires repeating the provided JSON action to proceed.

F.2 Prompt Templates

We provide the prompt templates used in our experiments as reference. The prompts are written in Jinja2 templates for easy integration into the system pipeline.

F.2.1 ShapeWorld

Causal Prior

You are analyzing causal relationships in a hypothetical world of shapes. How the movements of shapes causally influence each other is governed by deterministic causal rules. When entering this world, you observe a snapshot of the world, which includes the set of moving shapes and the set of static shapes. This snapshot represents a state that is valid under the world's internal causal rules.

****The Task:****
Analyze the informational value of the Observation relative to the Causal Claim.

****Observation:****
Moving shapes: {{ active_vars }}
Static shapes: {{ inactive_vars }}

****Causal Statement:****
The movement of {{ cause_var }} causes {{ effect_var }} to move.

****Label Definitions:****

*** **impossible**:** The observation contradicts the causal statement.
*** **plausible**:** The observation provides a basis to postulate the causal statement.
*** **irrelevant**:** The observation is uninformative regarding the causal statement.
*** **certain**:** The observation provides deductive evidence for the causal statement.

****Question:****
Which label is the most suitable to describe the relationship between the observation and the causal claim?

****Output Format:****
Respond with a valid JSON object strictly following this schema:

```
...
{
    "reasoning": "Your step-by-step logic",
    "answer": "impossible" | "plausible" |
    "irrelevant" | "certain"
}
...
```

Forward State Tracing

You are a causal engine in a hypothetical of shapes. Your goal is to determine the final state of shaper after an intervention.

The movements of shapes follow deterministic causal rules. There are no hidden confounders.

****The Physics (Laws of the Universe):****
1. ****Connectivity:**** The causal graph is defined by the adjacency list below. When an edge from shape A to shape B is present, it means that the movement of shape A causes the movement of shape B.
2. ****State Logic:**** A shape is ****MOVING**** if and only if:
* It is directly intervened on to move.
* OR at least one of its parents in the graph is MOVING.
3. ****Default:**** If neither condition is met, the shape is ****STATIC****.

****The System:****
*** **Causal Graph:**** {{ causal_graph_text }}
*** **Pre-Intervention State:****
* Moving: {{ active_vars }}
* Static: {{ inactive_vars }}

****The Intervention:****
Interventions in this world are transient ****inputs**** and does not override the internal causal rules.

"Move" (Excitatory):
* Applies a force to the shape to initiate movement.

"Stop" (Inhibitory):
* Removes the shape's internal momentum.
* If a shape is being moved by an ****active parent****, you cannot hold the shape static.

****Intervention:**** {{ intervention_text }}

****Task:****
Determine the final state of every shape after the intervention propagates through the graph.

****Output Format:****
Respond with a valid JSON object following the schema below.
```json  
{  
 "reasoning": "Briefly explain the causal chain triggered by the intervention.",  
 "final\_state": {  
 {% for var in vars %}  
 "{{ var }}": "moving" | "static",  
 {% endfor %}  
 }  
}  
```

Belief Updating

You are an expert in causal logic. You are updating the given prior beliefs about a hypothetical world of shapes based on an ****Intervention Experiment****. The movements of shapes follow deterministic causal rules. A shape moves if it is directly intervened or if another shape's movement causes it to move. The underlying causal graph is acyclic, meaning there are no feedback loops in the causal structure. There are no hidden confounders.

****The Intervention:****
Interventions in this world are transient ****inputs**** and does not override the internal causal rules.

"Move" (Excitatory):
* Applies a force to the shape to initiate movement.

****"Stop" (Inhibitory):****
* Removes the shape's internal momentum.
* If a shape is being moved by an ****active parent****, you cannot hold the shape static.

****State 1: Pre-Intervention****
* Moving: {{ prior_active_vars }}
* Static: {{ prior_inactive_vars }}

****Current Belief:****
* {{ var_1 }} -> {{ var_2 }}: ****{{ prior_12 }}****
* {{ var_2 }} -> {{ var_1 }}: ****{{ prior_21 }}****

****Intervention:**** {{ intervention_text }}

****State 2: Post-Intervention Observation****
* Moving: {{ post_active_vars }}
* Static: {{ post_inactive_vars }}

Does this new evidence ****strengthen****, ****contradict****, or have ****no impact**** on the belief?

```json  
{ "reasoning": "Explain your rationale step by step here",  
  
 "{{ var\_1 }} -> {{ var\_2 }}": "strengthen" |  
 "contradict" | "no impact",  
 "{{ var\_2 }} -> {{ var\_1 }}": "strengthen" |  
 "contradict" | "no impact"}  
```

"{{ var_2 }} -> {{ var_1 }}": "strengthen" |
"contradict" | "no impact"}
```

### Agency Sensitivity - Observer

You are an expert in causal logic. You are observing a sequence of events in a hypothetical world of shapes and inferring the causal relationships between the movements of shapes.

The movements of shapes follow deterministic causal rules. A shape moves if it is directly intervened or if another shape's movement causes it to move. The underlying causal graph is acyclic, meaning there are no feedback loops in the causal structure. There are no hidden confounders.

**\*\*Intervention:\*\***  
Interventions in this world are transient **\*\*inputs\*\*** and does not override the internal causal rules.

"Move" (Excitatory):  
\* Applies a force to the shape to initiate movement.

**\*\*"Stop" (Inhibitory):\*\***  
\* Removes the shape's internal momentum.  
\* If a shape is being moved by an **\*\*active parent\*\***, you cannot hold the shape static.

Observe the following sequence of events, and answer the question:  
Does the movement of {{ cause\_var }} cause the movement of {{ effect\_var }}?

**\*\*Initial setup:\*\*** {{ initial\_setup }}  
**\*\*Observation history:\*\***  
{%- for step in obs\_history %}  
{{ step }}  
{%- endfor %}

**\*\*Output Format:\*\***  
Respond with a valid JSON object strictly following this schema:

```json  
{
 "reasoning": "Your step-by-step logic",
 "answer": "Yes" | "No"
}
```

### Agency Sensitivity - Agent

You are an expert in causal logic. You are conducting a series of interventions in a hypothetical world of shapes to infer the causal relationships between the movements of shapes.

The movements of shapes follow deterministic causal rules. A shape moves if it is directly intervened or if another shape's movement causes it to move. The underlying causal graph is acyclic, meaning there are no feedback loops in the causal structure. There are no hidden confounders.

```

Intervention:
Interventions in this world are transient
inputs and does not override the internal
causal rules.

"Move" (Excitatory):
* Applies a force to the shape to initiate
movement.

"Stop" (Inhibitory):
* Removes the shape's internal momentum.
* If a shape is being moved by an **active
parent**, you cannot hold the shape static.

At each step, you will be given the current
state of the world as a result of the previous
intervention, and the next intervention you
should perform.
Do not devise your own interventions. Deviating
from the given instructions will result in a
failed experiment.

After you are done with the interventions,
answer the question:
Does the movement of {{ cause_var }} cause the
movement of {{ effect_var }}?

Initial setup: {{ initial_setup }}

Your next intervention: {{ action }} the {{
target_name }}.

Output Format:
Respond with the JSON object below. Replying
with the provided intervention allows you to
interact with the world:

```json
{
  "shape": "{{ target_name }}",
  "action": "{{ action }}"
}
```

The current status of the shapes: {{
current_status }}

Your next intervention is to {{ action }} the
{{ target_name }}.

Output Format:
Respond with the JSON object below. Replying
with the provided intervention allows you to
interact with the world:

```json
{
  "reasoning": "Your reasoning about the
experiments so far",
  "shape": "{{ target_name }}",
  "action": "{{ action }}"
}
```

The current status of the shapes: {{
current_status }}

You have completed all interventions necessary
to answer the question:

```

Does the movement of {{ cause\_var }} cause the
movement of {{ effect\_var }}?

**\*\*Output Format:\*\***  
Respond with a valid JSON object strictly
following this schema:

```

```json
{
  "reasoning": "Your step-by-step logic",
  "answer": "Yes" | "No"
}
```

```

## F.2.2 HealthCare

### Causal Prior

You are analyzing causal relationships in a
hypothetical healthcare system simulation. How
the factors of this system causally influence
each other is governed by deterministic causal
rules.

When entering this simulation, you observe a
snapshot of the system, which includes the
current values of all factors. This snapshot
represents a state that is valid under the
simulation's internal causal rules.

**\*\*The Task:\*\***  
Analyze the informational value of the
Observation relative to the Causal Claim.

**\*\*Observation:\*\***  
High: {{ active\_vars }}  
Low: {{ inactive\_vars }}

**\*\*Causal Statement:\*\***  
High {{ cause\_var }} causes high {{ effect\_var
}}.

**\*\*Label Definitions:\*\***  
\* **\*\*impossible\*\***: The observation contradicts
the causal statement.  
\* **\*\*plausible\*\***: The observation provides a
basis to postulate the causal statement.  
\* **\*\*irrelevant\*\***: The observation is
uninformative regarding the causal statement.  
\* **\*\*certain\*\***: The observation provides
deductive evidence for the causal statement.

**\*\*Question:\*\***  
Which label is the most suitable to describe
the relationship between the observation and
the causal claim?

**\*\*Output Format:\*\***  
Respond with a valid JSON object strictly
following this schema:

```

```
{
  "reasoning": "Your step-by-step logic",
  "answer": "impossible" | "plausible" |
"irrelevant" | "certain"
}
```

```

### Forward State Tracing

You are a causal engine in a simulation of a hypothetical healthcare system. Your goal is to determine the final state of factors after an intervention.

How the factors of this system causally influence each other is governed by deterministic causal rules, represented by an acyclic causal graph. There are no hidden confounders.

```
The Causal Rules of the System:
1. **Connectivity:** The causal graph is defined by the adjacency list below. When an edge from factor A to factor B is present, it means that a high value of A causes a high value of B (and a low value of A causes a low value of B).
2. **State Logic:** The value of a factor can be toggled to high or low through a regime switch as an intervention:
 * A factor's value is high if it is directly toggled from low to high.
 * OR at least one of its parents in the graph is high.
3. **Default:** If neither condition is met, the value of the factor is low.
```

```
The System:
* **Causal Graph:** {{ causal_graph_text }}
* **Pre-Intervention State:**
 * high: {{ active_vars }}
 * low: {{ inactive_vars }}
```

```
The Intervention:
Interventions in this simulation are transient **inputs** and does not override the internal causal rules.
```

```
"Increase" (Excitatory):
* Applies a regime switch to change the value of a factor from low to high.
```

```
"Decrease" (Inhibitory):
* Attempts to use the regime switch to change the value of a factor from high to be low.
* If a factor's value is high due to an **active parent**, you cannot force the factor's value to be low.
```

```
Intervention: {{ intervention_text }}
```

```
Task:
Determine the final state of every factor after the intervention propagates through the graph.
```

```
Output Format:
Respond with a valid JSON object following the schema below.
```json
{
  "reasoning": "Briefly explain the causal chain triggered by the intervention.",
  "final_state": {
    {% for var in vars %}
    "{{ var }}": "high" | "low",
    {% endfor %}
  }
}
```

Belief Updating

You are an expert in causal logic. You are updating the given prior beliefs about a simulation of a hypothetical healthcare system based on an ****Intervention Experiment****. How the factors of this system causally influence each other is governed by deterministic causal rules, represented by an acyclic causal graph (no feedback loops). There are no hidden confounders.

```
**The Intervention:**
Interventions in this world are transient **inputs** and does not override the internal causal rules.
```

```
"Increase" (Excitatory):
* Applies a regime switch to change the value of a factor from low to high.
```

```
"Decrease" (Inhibitory):
* Attempts to use the regime switch to change the value of a factor from high to be low.
* If a factor's value is high due to an **active parent**, you cannot force the factor's value to be low.
```

```
**State 1: Pre-Intervention**
* High: {{ prior_active_vars }}
* Low: {{ prior_inactive_vars }}
```

```
**Current Belief:**
* {{ var_1 }} -> {{ var_2 }}: **{{ prior_12 }}**
* {{ var_2 }} -> {{ var_1 }}: **{{ prior_21 }}**
```

```
**Intervention:** {{ intervention_text }}
```

```
**State 2: Post-Intervention Observation**
* High: {{ post_active_vars }}
* Low: {{ post_inactive_vars }}
```

Does this new evidence ****strengthen****, ****contradict****, or have ****no impact**** on the belief?

```
```json
{ "reasoning": "Explain your rationale step by step here",
```

```
"{{ var_1 }} -> {{ var_2 }}": "strengthen" | "contradict" | "no impact",
```

```
"{{ var_2 }} -> {{ var_1 }}": "strengthen" | "contradict" | "no impact"
```
```

Agency Sensitivity - Observer

You are an expert in causal logic. You are observing a sequence of events in a simulation of a hypothetical healthcare system. How the factors of this system causally influence each other is governed by deterministic causal rules, represented by an acyclic causal graph (no feedback loops). There are no hidden confounders.

The value of a factor is either high or low. A factor's value is high if it is directly intervened or if at least one of its parents in

the graph is high. Otherwise, the value of the factor is low.

****The Intervention:****
Interventions in this world are transient
****inputs**** and does not override the internal causal rules.

"Increase" (Excitatory):
* Applies a regime switch to change the value of a factor from low to high.

"Decrease" (Inhibitory):
* Attempts to use the regime switch to change the value of a factor from high to be low.
* If a factor's value is high due to an ****active parent****, you cannot force the factor's value to be slow.

Observe the following sequence of events, and answer the question:
Does high {{ cause_var }} cause high {{ effect_var }}?

****Initial setup:**** {{ initial_setup }}
****Observation history:****
{%- for step in obs_history %}
{{ step }}
{%- endfor %}

****Output Format:****
Respond with a valid JSON object strictly following this schema:

```
```json
{
 "reasoning": "Your step-by-step logic",
 "answer": "Yes" | "No"
}
```
```

Agency Sensitivity - Agent

You are an expert in causal logic. You are conducting a series of interventions in a simulation of a hypothetical healthcare system. How the factors of this system causally influence each other is governed by deterministic causal rules, represented by an acyclic causal graph (no feedback loops). There are no hidden confounders.

The value of a factor is either high or low. A factor's value is high if it is directly intervened or if at least one of its parents in the graph is high. Otherwise, the value of the factor is low.

****The Intervention:****
Interventions in this world are transient
****inputs**** and does not override the internal causal rules.

"Increase" (Excitatory):
* Applies a regime switch to change the value of a factor from low to high.

"Decrease" (Inhibitory):
* Attempts to use the regime switch to change the value of a factor from high to low.
* If a factor's value is high due to an

****active parent****, you cannot force the factor's value to low.

At each step, you will be given the current state of the simulation as a result of the previous intervention, and the next intervention you should perform. Do not devise your own interventions. Deviating from the given instructions will result in a failed experiment.

After you are done with the interventions, answer the question:
Does high {{ cause_var }} cause high {{ effect_var }}?

****Initial setup:**** {{ initial_setup }}

****Your next intervention:**** {{ action }} the {{ target_name }}.

****Output Format:****
Respond with the JSON object below. Replying with the provided intervention allows you to interact with the simulation:

```
```json
{
 "factor": "{{ target_name }}",
 "action": "{{ action }}"
}
```
```

The current status of the factors: {{ current_status }}

Your next intervention is to {{ action }} the {{ target_name }}.

****Output Format:****
Respond with the JSON object below. Replying with the provided intervention allows you to interact with the simulation:

```
```json
{
 "reasoning": "Your reasoning about the experiments so far",
 "factor": "{{ target_name }}",
 "action": "{{ action }}"
}
```
```

The current status of the factors: {{ current_status }}

You have completed all interventions necessary to answer the question:
Does high {{ cause_var }} cause high {{ effect_var }}?

****Output Format:****
Respond with a valid JSON object strictly following this schema:

```
```json
{
 "reasoning": "Your step-by-step logic",
 "answer": "Yes" | "No"
}
```
```

...

F.2.3 BossFight

Causal Prior

You are analyzing causal relationships in a hypothetical action game. How the game elements of this game causally influence each other is governed by deterministic causal rules. When entering this game, you observe a snapshot of the game, which includes the current values of all game elements. This snapshot represents a state that is valid under the game's internal causal rules.

****The Task:****

Analyze the informational value of the Observation relative to the Causal Claim.

****Observation:****

On: {{ active_vars }}

Off: {{ inactive_vars }}

****Causal Statement:****

The activation of {{ cause_var }} causes {{ effect_var }} to be on.

****Label Definitions:****

* ****impossible****: The observation contradicts the causal statement.

* ****plausible****: The observation provides a basis to postulate the causal statement.

* ****irrelevant****: The observation is uninformative regarding the causal statement.

* ****certain****: The observation provides deductive evidence for the causal statement.

****Question:****

Which label is the most suitable to describe the relationship between the observation and the causal claim?

****Output Format:****

Respond with a valid JSON object strictly following this schema:

...

```
{
  "reasoning": "Your step-by-step logic",
  "answer": "impossible" | "plausible" |
  "irrelevant" | "certain"
}
```

...

Forward State Tracing

You are a causal engine in an action game for a boss fight. Your goal is to determine the final state of game elements after an intervention.

How the game elements causally influence each other is governed by deterministic causal rules, represented by an acyclic causal graph. There are no hidden confounders.

****The Causal Rules of the System:****

1. ****Connectivity:**** The causal graph is defined by the adjacency list below. When an edge from game element A to game element B is present, it means that the activation of game

element A causes the activation of game element B.

2. ****State Logic:**** The value of a game element can be toggled to on or off through a switch as an intervention:

* A game element's value is on if it is directly toggled from off to on.

* OR at least one of its parents in the graph is on.

3. ****Default:**** If neither condition is met, the value of the game element is off.

****The System:****

* ****Causal Graph:**** {{ causal_graph_text }}

* ****Pre-Intervention State:****

* on: {{ active_vars }}

* off: {{ inactive_vars }}

****The Intervention:****

Interventions on the game elements are transient ****inputs**** and does not override the game's internal causal rules.

"Activate" (Excitatory):

* Set the value of a game element from off to on.

"Deactivate" (Inhibitory):

* Attempts to use the switch to turn off a game element.

* If a game element is on due to an ****active parent****, you cannot force the game element's value to be off.

****Intervention:**** {{ intervention_text }}

****Task:****

Determine the final state of every game element after the intervention propagates through the graph.

****Output Format:****

Respond with a valid JSON object following the schema below.

```json

```
{
 "reasoning": "Briefly explain the causal
 chain triggered by the intervention.",
 "final_state": {
 {% for var in vars %}
 "{{ var }}": "on" | "off",
 {% endfor %}
 }
}
```

...

#### Belief Updating

You are an expert in causal logic. You are updating the given prior beliefs about the mechanics of a hypothetical action game based on an **\*\*Intervention Experiment\*\***. How the game elements causally influence each other is governed by deterministic causal rules, represented by an acyclic causal graph (no feedback loops). There are no hidden confounders.

**\*\*The Intervention:\*\***

Interventions on the game elements are transient **\*\*inputs\*\*** and does not override the

game's internal causal rules.

"Activate" (Excitatory):

\* Set the value of a game element from off to on.

"Deactivate" (Inhibitory):

\* Attempts to use the switch to turn off a game element.

\* If a game element is on due to an **active parent**, you cannot force the game element's value to be off.

**State 1: Pre-Intervention**

\* On: {{ prior\_active\_vars }}

\* Off: {{ prior\_inactive\_vars }}

**Current Belief:**

\* {{ var\_1 }} -> {{ var\_2 }}: **prior\_12**

\* {{ var\_2 }} -> {{ var\_1 }}: **prior\_21**

**Intervention:** {{ intervention\_text }}

**State 2: Post-Intervention Observation**

\* On: {{ post\_active\_vars }}

\* Off: {{ post\_inactive\_vars }}

Does this new evidence **strengthen**, **contradict**, or have **no impact** on the belief?

```json

{ "reasoning": "Explain your rationale step by step here",

"{{ var_1 }} -> {{ var_2 }}": "strengthen" | "contradict" | "no impact",

"{{ var_2 }} -> {{ var_1 }}": "strengthen" | "contradict" | "no impact"

```

### Agency Sensitivity - Observer

You are an expert in causal logic. You are observing a sequence of events in a hypothetical action game to discover the hidden game mechanics. How the game elements causally influence each other is governed by deterministic causal rules, represented by an acyclic causal graph (no feedback loops). There are no hidden confounders.

The game element is either on or off. A game element is on if it is directly activated or if at least one of its parents in the graph is on. Otherwise, the value of the game element is off.

**The Intervention:**

Interventions on the game elements are transient **inputs** and does not override the game's internal causal rules.

"Activate" (Excitatory):

\* Set the value of a game element from off to on.

"Deactivate" (Inhibitory):

\* Attempts to use the switch to turn off a game element.

\* If a game element is on due to an **active**

**parent**, you cannot force the game element's value to be off.

Observe the following sequence of events, and answer the question:

Does the activation of {{ cause\_var }} cause {{ effect\_var }} to be on?

**Initial setup:** {{ initial\_setup }}

**Observation history:**

{%- for step in obs\_history %}

{{ step }}

{%- endfor %}

**Output Format:**

Respond with a valid JSON object strictly following this schema:

```json

{
 "reasoning": "Your step-by-step logic",
 "answer": "Yes" | "No"
}

```

### Agency Sensitivity - Agent

You are an expert in causal logic. You are conducting a series of interventions in a hypothetical action game to discover the hidden game mechanics. How the game elements causally influence each other is governed by deterministic causal rules, represented by an acyclic causal graph (no feedback loops). There are no hidden confounders.

The game element is either on or off. A game element is on if it is directly activated or if at least one of its parents in the graph is on. Otherwise, the value of the game element is off.

**The Intervention:**

Interventions on the game elements are transient **inputs** and does not override the game's internal causal rules.

"Activate" (Excitatory):

\* Set the value of a game element from off to on.

"Deactivate" (Inhibitory):

\* Attempts to use the switch to turn off a game element.

\* If a game element is on due to an **active parent**, you cannot force the game element's value to be off.

At each step, you will be given the current state of the game as a result of the previous intervention, and the next intervention you should perform. Do not devise your own interventions. Deviating from the given instructions will result in a failed experiment.

After you are done with the interventions, answer the question:

Does the activation of {{ cause\_var }} cause {{ effect\_var }} to be on?

**Initial setup:** {{ initial\_setup }}

**\*\*Your next intervention:\*\*** {{ action }} the {{ target\_name }}.

**\*\*Output Format:\*\***

Respond with the JSON object below. Replying with the provided intervention allows you to interact with the game:

```
```json
{
  "game_element": "{{ target_name }}",
  "action": "{{ action }}"
}
```
```

The current status of the game elements: {{ current\_status }}

Your next intervention is to {{ action }} the {{ target\_name }}.

**\*\*Output Format:\*\***

Respond with the JSON object below. Replying with the provided intervention allows you to interact with the game:

```
```json
{
  "reasoning": "Your reasoning about the experiments so far",
  "game_element": "{{ target_name }}",
  "action": "{{ action }}"
}
```
```

The current status of the game elements: {{ current\_status }}

You have completed all interventions necessary to answer the question:  
Does the activation of {{ cause\_var }} cause {{ effect\_var }} to be on?

**\*\*Output Format:\*\***

Respond with a valid JSON object strictly following this schema:

```
```json
{
  "reasoning": "Your step-by-step logic",
  "answer": "Yes" | "No"
}
```
```

## G Generative Assistance in Authorship

### G.1 Writing

We used ChatGPT (v5.2) and Gemini 3.1 Pro as writing assistants. Specifically, they were used for:

- **Literature Search:** The assistants were used to surface potentially relevant papers. All cited works were selected, verified, and cited by the authors.
- **Language editing:** The assistants were used

to proofread and polish author-written text (e.g., grammar, clarity, and tone) throughout the manuscript.

- **Draft scaffolding:** The assistants were used to produce outline-level drafts for Section 1 (Introduction), Section 2 (Related Work), Appendix A (Extended Related Work), and Section 5 (Conclusion), based on author-provided background literature and the study's methods and results. Generated text served only as a structural scratchpad and was rewritten by the authors.

The AI assistants were **not** used to generate original research ideas or claims.

### G.2 Coding

We used Claude 4.6 Sonnet and Gemini 3.1 Pro for coding assistance. The core logic and base classes were implemented by the authors. The assistants were used to draft documentation, regex patterns, fully specified algorithm implementations, and standard scripts for evaluation and plotting. All generated code was reviewed by the authors to ensure correctness.