REVISITING CAUSAL REASONING IN LANGUAGE MODELS THROUGH CONTROLLED SYNTHETIC WORLDS

Abhirath Sangala[1] **Vineeth N. Balasubramanian**[1*] **Amit Sharma**[1*]
[1]Microsoft Research, India

ABSTRACT

Evidence on whether LLMs can reason causally remains mixed, partly because existing benchmarks either allow retrieval-based shortcuts from pretraining or rely on in-context synthetic stories that are weakly aligned with how models acquire knowledge. We present a controlled synthetic-world benchmark that mirrors LLMs' training setting: we generate a causal world with a known DAG structure and Boolean mechanisms, textualize it into observations, and fine-tune LLMs before evaluating them on three task families (simple/forward prediction, L1 associational reasoning, and L2 interventional reasoning). Unlike prior benchmarks, our framework provides training observations from a structural causal model, enabling identification of specific causal reasoning abilities as the training dataset mix is changed. Across experiments, models learn individual causal mechanisms and can generalize to shifted distributions when some examples from those distributions are seen during training. However, they struggle to compose novel causal chains, generalize to new scenario structures, and transfer knowledge across related tasks. These results suggest that current LLMs internalize local causal information without forming an accurate internal causal model. Our results help explain prior mixed findings: current LLMs, trained on large and diverse training data, can achieve improved performance on many benchmarks, but systematic generalization beyond seen distributions remains limited.

1 INTRODUCTION

There is growing interest in using Large Language Models (LLMs) as causal experts in critical, high-value applications such as clinical effect estimation (Dhawan et al., 2024; Ma et al., 2025), disease causal discovery (Xu et al., 2025), scientific gene regulatory network inference (Afonja et al., 2024), root cause analysis in production systems (Chen et al., 2024a; Ahmed et al., 2023), and automated policy evaluation (Verma et al., 2025). This naturally leads us to the question: can LLMs reason causally? Several works attempt to answer this question within the context of different tasks, e.g. causal discovery (Kiciman et al., 2023; Jiralerspong et al., 2024; Ban et al., 2023), pairwise causal judgment (Jin et al., 2023a), and interventional and counterfactual reasoning (Jin et al., 2023b; Chen et al., 2024b). However, the evidence from these works is mixed. Moreover, these works do not capture the way LLMs learn by training on text describing causal processes. Most causal reasoning benchmarks test in-context reasoning abilities (Jin et al., 2023b; Chen et al., 2024b; Jin et al., 2023a), whereas many downstream uses rely on LLMs reasoning about causal information acquired through pre-training. At its core, reasoning about causal structure reduces to *multi-step deductive inference* over learned rules—a setting where the logical reasoning literature has already identified compositional generalization as a fundamental challenge (Clark et al., 2020; Tafjord et al., 2021; Saparov et al., 2023).

We propose a fresh approach to this question by constructing a controlled synthetic causal world, fine-tuning LLMs on text describing it, and testing them on a variety of reasoning tasks. Unlike prior in-context benchmarks, our setup tests reasoning over causal knowledge *ingested through training*. Unlike story-based benchmarks, it eliminates commonsense shortcuts and provides full *symbolic ground truth*, enabling verification of not only answers but the reasoning process itself.

*Equal advising. Correspondence to: vineeth.nb@microsoft.com, amshar@microsoft.com.

Finally, we probe whether LLMs develop *coherent internal causal representations*—a question largely unexplored in prior work. We find that:

1. Individual pairwise causal mechanisms are learned well, but LLMs fail to compose novel causal chains from them.
2. Generalization succeeds for novel values within familiar graph structures but breaks down for novel graph structures.
3. Increased exposure to diverse data from the world improves reasoning performance in many cases. However, internal causal representations are incoherent. Models may provide the right answer but with a wrong explanation.
4. Cross-task transfer fails even between related tasks.

Our findings help explain the mixed evidence present in the literature where LLMs perform rather well at certain causal reasoning tasks but fail at others.

Connection to logical reasoning. At its core, our task is multi-step deductive inference over learned boolean functions. RuleTaker (Clark et al., 2020) and PrOntoQA-OOD (Saparov et al., 2023) show that depth coverage in training is necessary but not sufficient for compositional generalization over inference chains—a failure our depth and subgraph results corroborate in a causal setting. ProofWriter (Tafjord et al., 2021) further demonstrates that faithful proof generation is a distinct challenge from answer accuracy, paralleling our finding that correct answers can co-occur with incorrect causal explanations. Our causal framing extends this logical reasoning literature by additionally testing whether internalized knowledge transfers across structurally different task formats.

2 Related Work

Early results suggested LLMs possess meaningful causal reasoning abilities: Kiciman et al. (2023) found that LLMs could produce correct causal arguments across a range of pairwise tasks, and subsequent work successfully leveraged LLMs as priors for causal graph discovery (Jiralerspong et al., 2024; Ban et al., 2023). However, this optimism has been challenged: Zečević et al. (2023) argue that LLMs can "talk causality" without genuinely reasoning causally, Yang et al. (2024) show many benchmark successes can be explained by domain knowledge retrieval, and Chi et al. (2024) provide evidence that LLMs perform only shallow causal reasoning that degrades on fresh benchmarks. The consensus thus remains mixed. We identify three gaps in existing evaluations that make it difficult to isolate genuine causal reasoning. First, nearly all benchmarks provide causal structure *in-context*—as explicit graphs, rules, or premises—rather than testing *internalized* causal knowledge acquired during training, which is just as relevant (if not more) to how LLMs are deployed as causal experts in practice. Second, many benchmarks rely on real-world causal content, allowing LLMs to exploit parametric commonsense knowledge rather than perform genuine causal reasoning (Yang et al., 2024; Chi et al., 2024); story-based benchmarks such as GLUCOSE (Mostafazadeh et al., 2020) and e-CARE (Du et al., 2022) further lack *symbolic ground truth* for explanations, making it impossible to distinguish faithful reasoning from post-hoc rationalization. Third, *cross-task coherence* has been largely untested: existing benchmarks evaluate each task in isolation, but a genuine internal causal model would naturally support knowledge transfer across structurally different tasks. Our framework addresses all three.

Benchmarks for causal reasoning in LLMs. CLadder (Jin et al., 2023b) evaluates causal reasoning aligned with Pearl's causal hierarchy, CausalBench (Chen et al., 2024b) evaluates several causal tasks independently, Corr2Cause (Jin et al., 2023a) tests whether LLMs can distinguish causation from correlation, and CausalARC (Maasch et al., 2026) samples reasoning tasks from fully specified SCMs to test in-context causal reasoning under distribution shift. While these benchmarks rigorously test reasoning over explicitly provided causal structure, they all supply this structure in-context—leaving open whether LLMs can reason over causal knowledge *internalized* during training. Separately, Zečević et al. (2023) argue that LLMs can "talk causality" without genuinely reasoning causally, and Yang et al. (2024) show that benchmarks using real-world causal content can often be solved via domain knowledge retrieval rather than genuine reasoning. Our design addresses both gaps: by fine-tuning on synthetic causal worlds and providing no causal graph or mechanism information

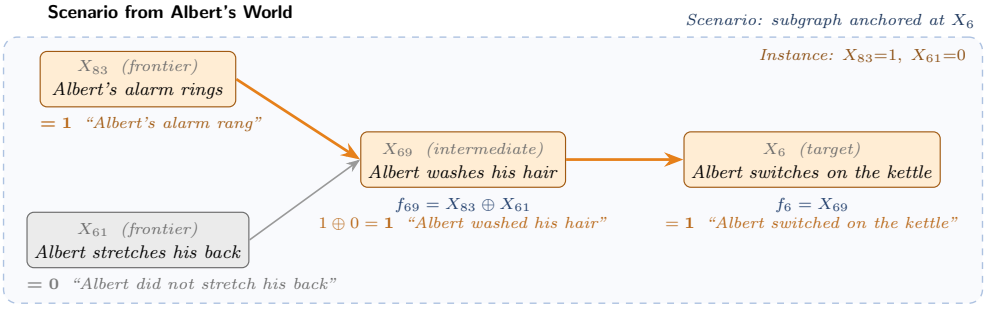


Figure 1: A scenario from Albert's World. Each node is a boolean variable grounded in a natural-language event description. Frontier nodes (X_{83}, X_{61}) are the roots of the subgraph; endogenous nodes (X_{69}, X_6) are determined by mechanisms ($f_{69} = X_{83} \oplus X_{61}$, $f_6 = X_{69}$). Orange nodes and edges carry value 1; gray carry value 0. The *scenario* is the subgraph structure; the *instance* is characterized by a specific value assignment to the frontier variables.

at test time, models must reason over internalized structure about a synthetic world whose causal relationships do not align with real-world commonsense expectations, and cannot be shortcut via pre-training knowledge.

Story-based and commonsense causal benchmarks. GLUCOSE (Mostafazadeh et al., 2020) provides large-scale causal explanations grounded in narratives, and e-CARE (Du et al., 2022) explores explainable causal reasoning from natural text. While these benchmarks use realistic language, they lack symbolic ground truth for explanation verification—it is impossible to definitively determine whether a model's explanation reflects genuine causal reasoning or post-hoc rationalization. Our synthetic world provides complete symbolic ground truth, enabling exact verification of both the logical structure and variable values in generated explanations.

World models and internal representations. Li et al. (2023) demonstrate that a GPT-style model trained to predict legal Othello moves develops an emergent internal representation of the board state. Gurnee & Tegmark (2024) show that LLM activations linearly encode spatial and temporal information, and Nichani et al. (2024) provide theory showing gradient descent can drive transformers to encode causal graph structure in attention. Lampinen et al. (2023) show that imitation learners and language models can learn active causal strategies from passive data, though transferability depends on alignment between training supervision and target causal structure. These results suggest internal world representations can emerge from training on structured data. However, our results indicate a more pessimistic boundary for causal structure: even when individual mechanisms are well-learned, the resulting representations appear to be task-bound and do not support systematic composition or cross-task transfer. Having a representation is not the same as having a transferable, coherent causal model (Nanda et al., 2023; Wang et al., 2024).

3 A Synthetic Causal World Framework

We are interested in understanding how LLMs learn and reason about the causal structure of a world they have been exposed to through fine-tuning. To study this, we need a world whose causal structure is fully known and controllable (enabling exact ground-truth verification of both answers and reasoning traces), and crucially, does not align with commonsense knowledge, so that models cannot leverage pre-training associations to circumvent our tests.

We make this concrete with *Albert's World* (Figure 1): a synthetic world built around the daily events of a fictional character. Each variable represents a boolean event—"Albert's alarm rings," "Albert washes his hair," "Albert switches on the kettle"—and causal mechanisms link them in deliberately arbitrary ways. For example, whether Albert washes his hair is determined by the XOR of whether his alarm rang and whether he stretched his back—a relationship that bears no connection to real-world commonsense. Our approach instantiates this idea as a deterministic structural causal model (SCM)

with 100 boolean variables, textual groundings for each, and randomly composed mechanisms. From this world, we construct three types of textual tasks—Simple observations (forward prediction), L1 observations (associational reasoning), and L2 observations (interventional reasoning)—fine-tune LLMs on them, and test the fine-tuned models in controlled settings. We use Albert's World to illustrate the key concepts throughout this section.

3.1 PRELIMINARIES

At its core, a structural causal model (Pearl, 2009; Peters et al., 2017) specifies two things: (A) the causal dependencies within a system, and (B) the mechanism that determines each variable's value given its direct causes. We restrict ourselves to the deterministic case where every variable takes boolean values.

The causal dependencies are encoded in a directed acyclic graph (DAG) $G = (V, E)$, where each directed edge $(X_i, X_j) \in E$ indicates that X_i is a direct cause of X_j (e.g., the edge $X_{83} \to X_{69}$ in Figure 1). Root variables—nodes without parents—form the *exogenous* set $\mathcal{U}$, determined solely by external factors. Non-root variables form the *endogenous* set $\mathcal{X}$, fully determined by other variables in the system. We write $\mathrm{Pa}(X)$ for the parent set of X. For each endogenous variable $X \in \mathcal{X}$, there exists a deterministic mechanism $f_X : \{0,1\}^{|\mathrm{Pa}(X)|} \to \{0,1\}$ mapping parent values to the value of X (e.g., $f_{69} = X_{83} \oplus X_{61}$ in Figure 1). Given any assignment to the exogenous variables, the values of all endogenous variables are uniquely determined. Finally, each variable is associated with a natural-language *grounding*—a textual event description (e.g., "Albert's alarm rings" for X_{83})—together with tense variants, negated forms, and synonym sets that enable diverse textual realizations.

3.2 INSTANTIATING ALBERT'S WORLD

Causal Graph Generation Our graph G has $|V| = 100$ nodes organized across 10 layers, with an average indegree of 1.2. Edges are distributed as evenly as possible across nodes to avoid extreme degree concentration. This is substantially larger and deeper than existing causal reasoning benchmarks, which typically involve fewer than 10 variables (Jin et al., 2023b;a). The precise sampling algorithm is described in Appendix A.1.

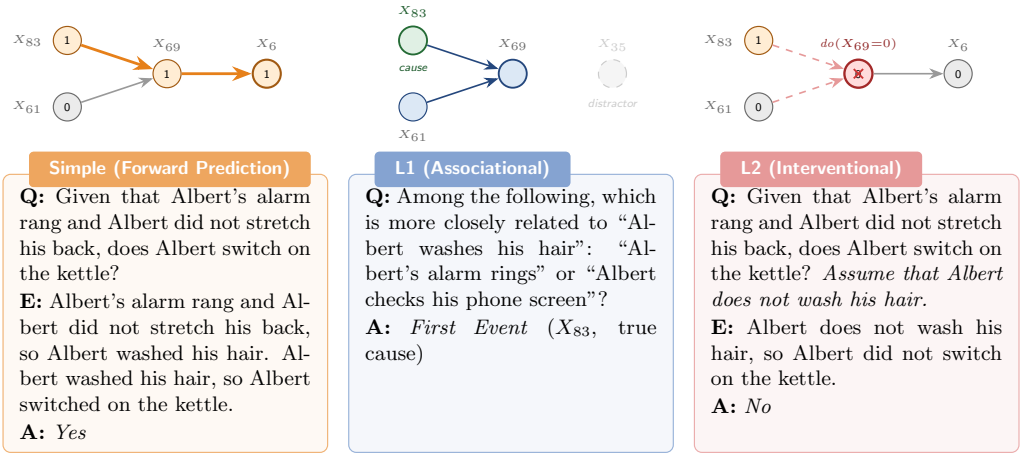
Causal Mechanism Generation Each mechanism f_X is randomly composed from a bank of primitive boolean functions (AND, OR, XOR, NOT, and their variants). Some primitives inherently skew outputs, e.g OR produces True for 3 of 4 input combinations, while AND produces True for only 1 combination. For this reason, we assign carefully tuned sampling weights to ensure balanced output distributions across the graph. This prevents evaluation metrics from being confounded by label imbalance. The full primitive set and weights are in Appendix A.2.2.

Grounding Generation To give the symbolic world a textual surface, every variable is associated with a distinct event description revolving around the day-to-day life of Albert (e.g., "Albert takes the bus," "Albert goes for a walk in the evening"). We prompt an LLM to generate unmistakably distinct present-tense descriptions, then for each generate synonym sets, tense variants (present-continuous, past, future), and negated counterparts (e.g., "Albert doesn't take the bus"). This yields a rich grounding set with multiple synonym sets, multiple tenses, affirmative and negated forms per variable.

Given the graph, mechanisms, and groundings, we enumerate all exogenous value configurations and propagate through the mechanisms to obtain a complete description of the causal world.

3.3 OBSERVATIONS

We formalize information about the causal world as tasks called *observations* (full examples in Appendix A.3). Symbolically, each observation captures a specific fact about the causal structure; these are then realized in natural language via templates and the grounding set (Figure 2). We define three types, each encoding a different aspect of causal knowledge: **Simple observations** capture forward prediction—the target value given frontier values, determined by composing mechanisms. **L1 observations** capture associational structure—which of two candidate events is causally adjacent



Simple (Forward Prediction)

Q: Given that Albert's alarm rang and Albert did not stretch his back, does Albert switch on the kettle?

E: Albert's alarm rang and Albert did not stretch his back, so Albert washed his hair. Albert washed his hair, so Albert switched on the kettle.

A: *Yes*

L1 (Associational)

Q: Among the following, which is more closely related to "Albert washes his hair": "Albert's alarm rings" or "Albert checks his phone screen"?

A: *First Event* (X_{83}, true cause)

L2 (Interventional)

Q: Given that Albert's alarm rang and Albert did not stretch his back, does Albert switch on the kettle? *Assume that Albert does not wash his hair.*

E: Albert does not wash his hair, so Albert did not switch on the kettle.

A: *No*

Figure 2: The three observation types derived from the scenario in Figure 1. **Simple** observations encode forward prediction: given frontier values, derive the target. **L1** observations encode associational structure: identify which of two events is more closely related to a query event. **L2** observations encode interventional reasoning: predict outcomes under a *do*-intervention that overrides a variable's mechanism. Each panel shows the annotated subgraph (top) and the resulting textualized observation (bottom).

to a query event. **L2 observations** capture interventional reasoning—the outcome when a *do*-intervention overrides a variable's natural mechanism. Fine-tuning on observations injects causal knowledge into the model; the same observations can also be used to evaluate causal understanding by requiring the model to produce correct completions. Simple and L2 observations are textualized as question–explanation–answer triples; L1 observations comprise a question and answer only.

3.3.1 Simple Observations

Consider the Simple panel of Figure 2. A simple observation is built from two ingredients. The *scenario* is a connected subgraph of the causal DAG anchored at a target variable—here the subgraph $\{X_{83}, X_{61}\} \to X_{69} \to X_6$, with target X_6. It defines *which* variables and mechanisms are involved. The *instance* is a specific value assignment to the frontier variables of that subgraph—here X_{83}=1, X_{61}=0—which, together with the mechanisms, uniquely determines all remaining values (X_{69}=1, X_6=1). Formally, a simple observation is a tuple $(G_D, v_D, \mathbf{s}_D)$: a connected subgraph G_D of G, a target $v_D \in V_D$, and a consistent assignment $\mathbf{s}_D \in \{0,1\}^{|V_D|}$. Subgraphs are sampled by selecting an endogenous variable and hopping backward through parents, producing scenarios of varying depth.

The textual realization has three parts, visible in the Simple panel of Figure 2: a **question** that states the frontier values in natural language and asks about the target ("*Given that Albert's alarm rang and Albert did not stretch his back, does Albert switch on the kettle?*"), an **explanation** that traces the step-by-step derivation through intermediate variables ("*Albert's alarm rang... so Albert washed his hair. Albert washed his hair, so Albert switched on the kettle.*"), and an **answer** ("*Yes*"). Templates and synonym-set groundings ensure surface-level diversity across realizations of the same symbolic observation.

An important note: the set of all depth-1 simple observations fully describes the causal world. Every mechanism is captured by at least one depth-1 observation, so this information suffices *in principle* to answer any question about the world. This is critical context for interpreting the depth generalization failures and L1 results that follow.

3.3.2 L1 (ASSOCIATIONAL) OBSERVATIONS

As illustrated in the L1 panel of Figure 2, an L1 observation presents a query variable (X_{69}), a *related* variable that is causally adjacent (X_{83}), and a *distractor* that is not (X_{35}). The model must identify the related variable—a task corresponding to the first rung of Pearl's causal hierarchy (Pearl, 2009). No graph structure is provided in context, so success requires the model to leverage causal knowledge internalized during fine-tuning. Unlike simple observations, L1 observations require no multi-step composition, only knowledge of direct adjacency.

3.3.3 L2 (INTERVENTIONAL) OBSERVATIONS

The L2 panel of Figure 2 shows an intervention $do(X_{69}{=}0)$ that overrides the mechanism of X_{69}, severing its incoming edges and forcing its value to 0. The model must predict the target X_6 under this modified graph—the second rung of Pearl's causal hierarchy. The textual format closely mirrors simple observations (question–explanation–answer with an interventional prefix), and this structural similarity turns out to be a critical factor in the model's ability to transfer knowledge from simple to L2 observations.

3.4 UNIQUENESS OF SYMBOLIC EXPLANATIONS

For any simple or L2 observation, the symbolic explanation is unique: every intermediate variable must be computed (or skipped under intervention), and the computation order is constrained by the topological ordering of the subgraph. Variables in the same topological layer may be permuted, but cross-layer reorderings violate dependency constraints. This uniqueness enables precise verification of model-generated explanations, which we exploit in our evaluation metrics (§4.2).

4 EXPERIMENTAL SETUP

4.1 MODELS AND FINE-TUNING

Training data. The training set is composed entirely of textualized observations from Albert's World. The base consists of simple observations; for the L1 and L2 experiments (§5.3, §5.4), we mix in task-specific observations at controlled proportions. Each experiment varies which observations are included and what is held out—details are given alongside the respective results. All training is performed with a standard autoregressive language modeling loss (next-token prediction) over the full observation text.

Models. We fine-tune four models: *Gemma-3-1B*, *Llama-3.2-1B*, *Qwen2.5-1.5B*, and *Qwen3-0.6B*, all adapted using LoRA (Hu et al., 2022) with rank 64, alpha 128, targeting all attention and MLP projection layers. Training uses 3 epochs, effective batch size 32, cosine learning rate schedule (2×10^{-4}), and bfloat16 precision. Full hyperparameters are in Appendix B.

4.2 METRICS AND EXPLANATION VERIFICATION

Answer accuracy. For simple and L2 observations, the fraction of correctly predicted target variable values. For L1, whether the model picks the more causally related variable.

Explanation accuracy. This is where our synthetic setup provides an advantage. Since models are fine-tuned on the grounding set, their generated text uses consistent language that can be parsed back to symbolic form via template matching. We then compare the recovered symbolic explanation to the ground truth along two dimensions: *structure matching* (are sentences in a permissible topological order, referencing the correct parent variables?) and *value matching* (does each variable take the correct value?). Both must match for the explanation to count as correct.

5 RESULTS

We organize our results around three experiments. §5.1 use simple observations to test generalization and explanation quality. §5.3 tests cross-task transfer to associational (L1) tasks. §5.4 tests transfer to

interventional (L2) tasks. Unless noted, we report model-averaged metrics; per-model breakdowns are in the appendix.

5.1 Generalization in Simple Observations

Simple observations let us systematically test four modes of generalization by controlling what is withheld from training. Each mode isolates a different aspect of causal understanding (Figures 3 and 4):

- **Depth generalization**: every simple observation has a *depth*—the number of mechanism applications from frontier to target. We partition observations by depth: train on a contiguous range (e.g., depths 1–3) and test on the remainder (e.g., depths 4–9). We run 6 such train/test splits covering both forward (shallow→deep) and backward (deep→shallow) extrapolation.
- **Subgraph generalization**: every simple observation is associated with a particular subgraph (scenario structure). We vary the *fraction* of distinct subgraphs included in training from 10% to 90% and test on the held-out subgraphs. Crucially, all depth-1 observations—which capture every individual mechanism—are always included, so the model has seen every causal rule; the question is whether it can compose them into novel chains.
- **Observation generalization**: each subgraph can be instantiated with different variable value assignments. We vary the fraction of instances per scenario included in training from 10% to 90%, testing on held-out instances of the *same* scenarios. The scenario structure is familiar; only the specific input values are novel.
- **Unnatural generalization**: for some scenarios, certain frontier value combinations never arise under any exogenous configuration—the world's dynamics simply never produce them—but the mechanisms are well-defined for these inputs. We vary the fraction of these "unnatural" instances in training from 1% to 10%, testing on held-out unnatural instances. This probes whether models can apply known mechanisms to inputs that never naturally occur.

Increasing training set diversity improves performance. From Figures 3 and 4, our first result is that all generalization curves improve with more training data. Thus, having a large and diverse training set helps with generalization.

Models cannot compose known mechanisms into novel causal chains. Subgraph generalization provides the cleanest test of compositional reasoning. Remember: all depth-1 observations are always included in training, so the model has seen every individual causal mechanism in the world. The only question is whether it can compose these known rules into novel chains. It largely cannot. At 10% subgraph coverage, answer accuracy on held-out subgraphs is just 60% and explanation accuracy is 7%; even at 50% coverage, these reach only 90% and 77% (Figure 4a). This is not a case of missing knowledge—it is a failure of composition.

Depth generalization reinforces this picture, though the interpretation is more straightforward: testing on unseen depths presents a clear out-of-distribution setting. When trained on depths 1–3 and tested on depths 4–9, answer accuracy drops to 60% and explanation accuracy collapses to 11% (Table 1, Figure 3). Widening the training range helps—training on 1–7 yields 84% answer accuracy on depths 8–9—but performance never approaches in-distribution levels. Backward generalization (deep→shallow) shows a qualitatively similar pattern.

Observation and unnatural generalization confirm the diagnosis. If the failure were about insufficient world knowledge rather than composition, we would expect models to also struggle when given novel *values* for familiar scenario structures. They do not. Unnatural generalization—testing on frontier value combinations that never naturally arise under any exogenous configuration—is robust even at 1% unnatural instances in training: both answer and explanation accuracy average $\sim$80%, and the curve stays essentially flat across all fractions tested (Figure 4c). Observation generalization shows a similar pattern: starting at $\sim$70% with just 10% of instances per scenario and converging to in-distribution levels by 60–70% coverage (Figure 4b). In both cases, the compositional structure of the explanation is unchanged—the same chain of mechanisms applies, only the specific input values differ. These results show a contrast with subgraph generalization: if values are changed, models are

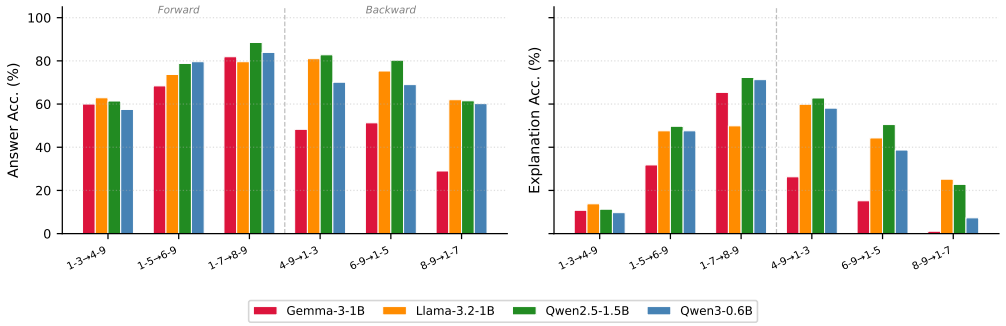


Figure 3: Depth generalization per model: forward (shallow→deep) and backward (deep→shallow). Wider training ranges help, but OOD performance remains well below in-distribution levels for all models.

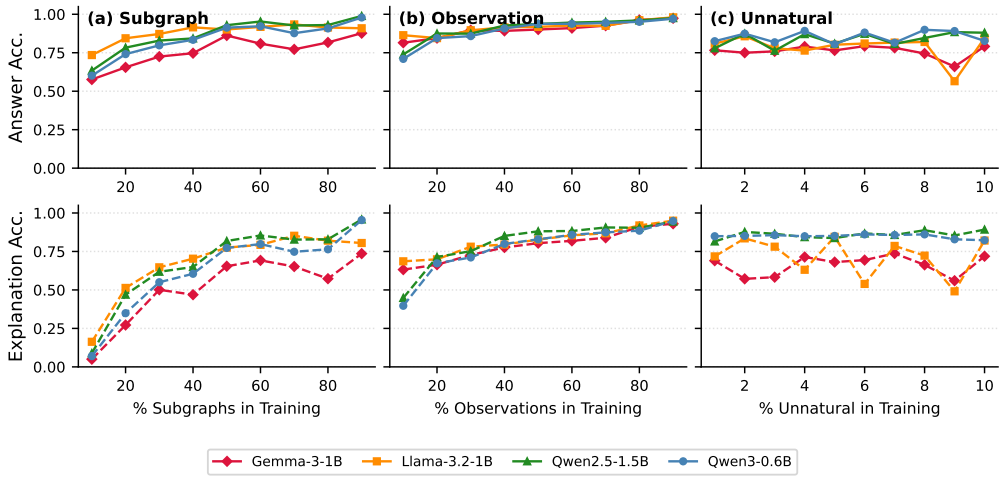


Figure 4: Generalization scaling curves per model. Top row: answer accuracy; bottom row: explanation accuracy. (a,d) Subgraph generalization requires extensive coverage; (b,e) observation generalization converges faster; (c,f) unnatural generalization is robust even at minimal coverage. The contrast between subgraph and the other two modes confirms that *structural composition*, not value generalization, is the bottleneck.

robust; however, if the structure is changed, they fail. This confirms that the bottleneck is *structural composition*, not value generalization or missing mechanism knowledge.

5.2 A Note on Answer–Explanation Decoupling

Something curious happens under distribution shift: models keep getting answers right while their explanations fall apart.

In-distribution, the answer–explanation gap is a modest 7 percentage points. Under distribution shift, it increases to 35: models get the answer right without a correct explanation in 36.9% of OOD cases—nearly as often as they get both right (28.3%). This is the "answer right / explanation wrong" failure mode. If models had acquired a coherent internal causal model, we would expect explanations to degrade alongside answers. Instead, explanations collapse disproportionately, suggesting the behavior may reflect distributional pattern matching rather than step-by-step causal reasoning.

The gap grows systematically with the distance between training and test depths (Table 1): from 3 pp when testing in-distribution on depths 1–3, to 49 pp when extrapolating from depths 1–3 to 4–9. We present the depth generalization case because the evidence is strongest here, though similar patterns are likely present in other generalization modes.

Table 1: Answer–explanation gap across depth generalization settings. Gap = answer accuracy − explanation accuracy (percentage points), aggregated across 4 models.

Train Depth	Eval Depth	Answer (%)	Expl. (%)	Gap (pp)
1–3	1–3 (in)	97.0	94.1	2.9
	4–9 (OOD)	60.5	11.4	49.1
1–5	1–5 (in)	97.3	92.5	4.7
	6–9 (OOD)	75.1	44.1	31.0
1–7	1–7 (in)	95.7	87.8	7.9
	8–9 (OOD)	83.5	64.7	18.8
4–9	4–9 (in)	95.5	88.4	7.2
	1–3 (OOD)	70.6	51.6	19.0
6–9	6–9 (in)	92.3	83.9	8.4
	1–5 (OOD)	69.0	37.0	31.9
8–9	8–9 (in)	85.6	70.9	14.6
	1–7 (OOD)	53.2	14.0	39.2
All in-dist.		**94.8**	**87.5**	**7.3**
All OOD		**65.2**	**29.9**	**35.3**

5.3 L1 Associational Reasoning

Given a variable, the task is to identify which of the two candidates is more closely associated. This requires no multi-step composition—only knowledge of direct adjacency in the causal graph, which is fully covered by the depth-1 simple observations in training. We train on mixtures of simple and L1 observations, varying the L1 proportion from 5% to 70%.

We find that L1 test accuracy hovers at $\sim$50%—indistinguishable from random guessing—across *all* training proportions. Meanwhile, L1 training accuracy reaches 100% and simple observation explanation accuracy remains above 93%. Even though the model does well on predicting the outcome in simple observations and it knows the task format (L1 training accuracy is perfect), it *cannot* bridge the two on unseen L1 instances.

This goes beyond the compositional failures of §5.1. L1 tasks do not require chaining mechanisms—only recognizing direct causal adjacency, which is fully represented in training. The failure suggests that whatever representation the model has is task-bound: it cannot be accessed or leveraged in a different task, even when the underlying knowledge is identical. This points to something closer to a task-specific lookup than a format-independent causal world model.

5.4 L2 Interventional Reasoning

L2 observations use a format structurally similar to simple observations—the same question–explanation–answer structure with an interventional prefix. We train on mixtures of simple and L2 observations, varying both the L2 proportion and the simple observation coverage (10% or 90%).

With limited world knowledge (10% simple observations), L2 answer accuracy reaches $\sim$78% but explanation accuracy plateaus at $\sim$39% (Figure 5a). With extensive world knowledge (90%), performance improves markedly: at 50% L2 proportion, answer accuracy reaches $\sim$97% and explanation accuracy $\sim$88% (Figure 5b). This behavior suggests that L2 reasoning benefits from both world knowledge and task-specific exposure, with format similarity appearing to facilitate knowledge transfer.

Comparison with L1. L1 fails despite world knowledge and task exposure whereas L2 shows substantial improvement. This is an interesting finding that deserves more study. We conjecture that the difference may be format proximity: L2 shares the question–explanation–answer structure with simple observations, differing only by an interventional prefix.

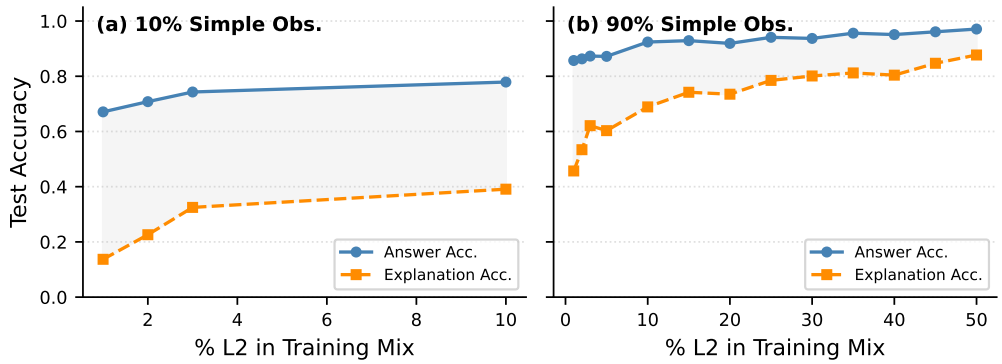


Figure 5: L2 (interventional) performance. (a) With 10% simple observations, limited world knowledge constrains performance. (b) With 90% simple observations, performance improves substantially, reaching $\sim$97% answer and $\sim$88% explanation accuracy at 50% L2.

6 Discussion

Explaining the mixed evidence. We opened this paper by noting that the evidence on LLM causal reasoning is decidedly mixed. Our results suggest a resolution. Models *can* appear to reason causally when two conditions hold: (1) the test format is structurally similar to training, and (2) enough data is provided to cover the relevant distributional surface. Under these conditions—which many existing benchmarks satisfy—performance is impressive. But if we remove either condition, then generalization fails: depth generalization fails, cross-format transfer collapses, and explanations decouple from answers. These results suggest that the capabilities of LLMs may be closer to distributional learning than an underlying causal world model. This explains why studies testing in-format, in-distribution settings find positive results (Kiciman et al., 2023), while more probing evaluations reveal brittleness (Zečević et al., 2023; Yang et al., 2024).

Lack of a cohesive internal causal representation. Individually, each finding has precedent: compositional generalization failures are well-documented (Lake & Baroni, 2018; Keysers et al., 2020; Kim & Linzen, 2020), and answer–explanation divergence has been noted in chain-of-thought work (Turpin et al., 2024; Lanham et al., 2023). But taken together, our dataset provides an opportunity to understand why models fail at generalization, both in observational and interventional settings.

7 Conclusion

Can language models build causal world models? Based on our evidence: not through fine-tuning. We constructed a fully synthetic textual causal world with complete symbolic ground truth, fine-tuned four LLMs on it, and probed their understanding through four generalization axes and two cross-task transfer settings. The results are consistent: models learn individual causal mechanisms reliably but acquire fragmented, task-bound representations rather than a coherent internal world model. They generalize to novel values but not novel structures.

The implication is that apparent causal competence in LLMs may be an artifact of distributional similarity between training and evaluation. When we need genuine causal reasoning—composing known mechanisms in novel ways, transferring causal knowledge across tasks—current fine-tuning falls short. Our benchmark makes a contribution towards understanding *how* and *where* it falls short.

Limitations and Future Work. We study small models (0.6B–1.5B); effects may differ at larger scales. We test a single deterministic boolean SCM—real-world causation involves continuous variables, stochastic mechanisms, and diverse topologies. We test fine-tuning only; pre-training dynamics may produce different internalizations. Replication across multiple graph structures would strengthen our conclusions. Natural extensions include scaling to larger models, testing multiple graph topologies and non-boolean mechanisms, and exploring whether curriculum-based training or neuro-symbolic approaches (Ontañón et al., 2022) can address the composition failures we observe.

REFERENCES

Tejumade Afonja, Ivaxi Sheth, Ruta Binkyte, Waqar Hanif, Thomas Ulas, Matthias Becker, and Mario Fritz. LLM4GRN: Discovering causal gene regulatory networks with LLMs – evaluation through synthetic data generation. *arXiv preprint arXiv:2404.16399*, 2024.

Toufique Ahmed, Supriyo Ghosh, Chetan Bansal, Thomas Zimmermann, Xuchao Zhang, and Saravan Rajmohan. Recommending root-cause and mitigation steps for cloud incidents using large language models. In *IEEE/ACM International Conference on Software Engineering*, 2023.

Taiyu Ban, Lyvyu Chen, Xiangyu Wang, and Huanhuan Chen. Causal inference using LLM-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023.

Runjie Chen, Brian Hou, Dmitri Ustebayev, Arturo Gonzalez, Yiran Li, Ratan Guha, Weiquan Lin, and Pengfei Yin. RCACopilot: Automatic root cause analysis in cloud on-call incident with large language models. In *Proceedings of the European Conference on Computer Systems (EuroSys)*, 2024a.

Yu Neng Chen, Yujie Wu, Tianhao Fan, Yue Ye, Jiahao Gao, Yuchi Zhang, Dongwei Song, and Luo Si. CausalBench: A comprehensive benchmark for causal learning capability of large language models. *arXiv preprint arXiv:2404.06349*, 2024b.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling causal reasoning in large language models: Reality or mirage? In *Advances in Neural Information Processing Systems*, volume 37, pp. 96640–96670, 2024.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *International Joint Conference on Artificial Intelligence*, 2020.

Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul G. Krishnan, and Chris J. Maddison. End-to-end causal effect estimation from unstructured natural language data. In *Advances in Neural Information Processing Systems*, 2024.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Wes Gurnee and Max Tegmark. Language models represent space and time. *International Conference on Learning Representations*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2022.

Zhijing Jin, Jiarui Chen, Yuhuai Gao, Sebastian Borgeaud, Marco Baroni, Max Welling, and Bernhard Schölkopf. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*, 2023a.

Zhijing Jin, Yuen Chen, Felix Leber, Luigi Gresele, Ojasv Kanal, Jiayi Zheng, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, 2023b.

Thomas Jiralerspong, Gael Le Berre, Vedang Bhatt, Satvik Vaidyanath, and Swami Sankaranarayanan Lim. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*, 2024.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Bubere, Daniel Furber, Nikola Kasber, Shikhar Ghelani, et al. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, 2018.

Andrew K. Lampinen, Stephanie C. Y. Chan, Ishita Dasgupta, Andrew J. Nam, and Jane X. Wang. Passive learning of active causal strategies in agents and language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benni Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *International Conference on Learning Representations*, 2023.

Yuchen Ma, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. LLM-driven treatment effect estimation under inference time text confounding. In *Advances in Neural Information Processing Systems*, 2025.

Jacqueline Maasch, John Kalantari, and Kia Khezeli. Causalarc: Abstract reasoning with causal world models, 2026.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Orr Braz, and Alan Ritter. GLUCOSE: Generalized and contextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In *International Conference on Machine Learning*, 2024.

Santiago Ontañón, Joshua Ainslie, Vaclav Cvicek, and Zachary Fisher. Making transformers solve compositional tasks. *arXiv preprint arXiv:2108.04378*, 2022.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Abulhair Saparov, Nesar Saparov, and He He. Testing the general deductive reasoning capacity of large language models using OOD examples. *Advances in Neural Information Processing Systems*, 2023.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of ACL*, 2021.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 2024.

Vishal Verma, Sawal Acharya, Samuel Simko, Devansh Bhardwaj, Anahita Haghighat, Mrinmaya Sachan, Dominik Janzing, Bernhard Schölkopf, and Zhijing Jin. Causal AI scientist: Facilitating causal data science with large language models. In *NeurIPS Workshop on CauScien*, 2025.

Ruoyao Wang, Graham Todd, Ziang Yuan, Rishabh Xiao, Marc-Alexandre Cote, Peter Clark, and Peter Jansen. Can language models serve as text-based world simulators? *arXiv preprint arXiv:2406.06485*, 2024.

Wei Xu, Gang Luo, Weiyu Meng, Xiaobing Zhai, Keli Zheng, Ji Wu, Yanrong Li, Abao Xing, Junrong Li, Zhifan Li, Ke Zheng, and Kefeng Li. MRAgent: An LLM-based automated agent for causal knowledge discovery in disease via Mendelian randomization. *Briefings in Bioinformatics*, 2025.

Linying Yang, Vik Hua, Hamilton Hanna, Katy Gunderson, Andrew Taliaferro, and Rex Ying. A critical review of causal reasoning benchmarks for large language models. *AAAI Workshop on Causal Inference and Machine Learning*, 2024.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023.

A Generation Algorithms

A.1 Sampling a Hierarchical Graph

The causal graph $G = (V, E)$ is constructed as a layered DAG with $|V| = N$ nodes, d topological layers, and a target average indegree k. Acyclicity is guaranteed by construction: edges only point from earlier layers to later layers.

Step 1: Layer assignment. A random permutation of $\{0, \ldots, N-1\}$ is drawn. The first d nodes form a *backbone*—one per layer, ensuring no layer is empty. The remaining $N-d$ nodes are each assigned uniformly at random to one of the d layers. Let $\mathcal{L}_0, \mathcal{L}_1, \ldots, \mathcal{L}_{d-1}$ denote the resulting layer sets.

Step 2: Guaranteed connectivity. For every node in layer $i \geq 1$, one parent is sampled uniformly at random from layer $\mathcal{L}_{i-1}$. This ensures every non-root node has at least one parent and the graph is connected across layers.

Step 3: Integer indegree edges. If $\lfloor k \rfloor \geq 2$, we repeat the following $\lfloor k \rfloor - 1$ additional times: for each layer $i \geq 1$, accumulate all nodes from layers $0, \ldots, i-1$ as candidate parents, and sample one parent per node in $\mathcal{L}_i$ uniformly from this pool. Duplicate edges are silently ignored.

Step 4: Fractional indegree edges. Let $f = k - \lfloor k \rfloor$ be the fractional part. We add $n_{\text{extra}} = \lfloor f \cdot N \rceil$ additional edges: for each, sample a non-root node v uniformly, then sample a parent uniformly from all nodes in layers strictly above v (i.e. layers $0, \ldots, \text{depth}(v)-1$).

With N=100, d=10, and k=1.2, this produces ∼110 edges (90 from Step 2 plus ∼20 from Step 4), giving an average indegree of ∼1.22 over the 90 non-root nodes.

A.2 Causal Mechanism Generation

A.2.1 Generating Causal Mechanisms from Primitive Boolean Functions

Each endogenous variable $X \in \mathcal{X}$ requires a mechanism $f_X : \{0,1\}^{|\text{Pa}(X)|} \to \{0,1\}$. We construct f_X by iteratively composing primitive boolean functions from the bank in Table 2.

Composition procedure. Let the parents of X be $\{P_1, \ldots, P_m\}$ in a random permutation. The mechanism is built left-to-right:

1. Initialize the running expression as $e_1 = P_1$.
2. For each subsequent parent P_i ($i = 2, \ldots, m$), sample a primitive g from the weighted distribution in Table 2 and set $e_i = g(e_{i-1}, P_i)$.
3. The final expression e_m defines f_X.

For nodes with a single parent (m=1), the mechanism is either identity or negation, sampled from the unary operators. This left-fold composition means that for $m > 2$ parents, the resulting expression is a nested binary tree of primitives. For example, with three parents $\{P_1, P_2, P_3\}$ and sampled primitives XOR and AND, the mechanism would be $(P_1 \oplus P_2) \wedge P_3$.

Safe compilation. Mechanism expressions are compiled into vectorized NumPy callables using a recursive-descent parser (not `eval`), supporting the grammar: *expr* → *or_expr*, *or_expr* → *xor_expr* (`or` *xor_expr*)*, *xor_expr* → *and_expr* (`^` *and_expr*)*, *and_expr* → *unary* (`and` *unary*)*, *unary* → `not` *unary* | *primary*.

A.2.2 List of used primitives and their respective weights and other information

Table 2 lists the full set of primitive boolean functions used to construct causal mechanisms as well as their sampling weights. The *Bias* column indicates whether the primitive skews the marginal output distribution toward True (Yes-leaning), toward False (No-leaning), or neither (Balanced), assuming uniform random inputs.

Primitive	Gate Name	Weight	Bias
Binary operators			
$A \wedge B$	AND	0.06	No-leaning
$A \wedge \neg B$	—	0.06	No-leaning
$\neg A \wedge B$	—	0.06	No-leaning
$\neg(A \vee B)$	NOR	0.11	No-leaning
$A \vee B$	OR	0.10	Yes-leaning
$A \vee \neg B$	—	0.08	Yes-leaning
$\neg A \vee B$	$A \Rightarrow B$	0.03	Yes-leaning
$\neg B \vee A$	$B \Rightarrow A$	0.03	Yes-leaning
$\neg(A \wedge B)$	NAND	0.03	Yes-leaning
$\neg A \vee B$	—	0.08	Yes-leaning
$A \oplus B$	XOR	0.12	Balanced
$\neg(A \oplus B)$	XNOR	0.12	Balanced
Unary operators			
A	Identity (A)	0.03	Balanced
B	Identity (B)	0.03	Balanced
$\neg A$	NOT (A)	0.03	Balanced
$\neg B$	NOT (B)	0.03	Balanced

Table 2: Primitive boolean functions and their sampling weights. Weights sum to 1. *Bias* denotes output skew under uniform inputs: No-leaning primitives output True with probability ≤ 0.25; Yes-leaning primitives with probability ≥ 0.75; Balanced primitives with probability 0.50.

A.3 Observation Examples

Below we provide complete examples of each observation type, showing both the symbolic form and the textualized realization.

A.3.1 Simple Observation Example

Symbolic form. Consider the subgraph $(X_{83}, X_{61}) \to X_{69} \to X_6$ from Figure 1, with target X_6. The frontier is $\{X_{83}, X_{61}\}$ and the mechanisms are $f_{69} = X_{83} \oplus X_{61}$, $f_6 = X_{69}$. For the instance X_{83}=1, X_{61}=0: $X_{69} = 1 \oplus 0 = 1$, then $X_6 = 1$.

Textualized form. Using question template "Given that {CAUSES}, does {EFFECT}?" and cause-first explanation templates:

Question: *Given that Albert's alarm rang and Albert did not stretch his back, does Albert wash his hair?*

Explanation: *Albert's alarm rang and Albert did not stretch his back, so Albert switched on the kettle. Albert switched on the kettle, hence Albert washed his hair.*

Answer: *Yes.*

A.3.2 L1 OBSERVATION EXAMPLE

Symbolic form. Query variable: X_{69}. Related (adjacent) variable: X_{83}. Distractor (non-adjacent): X_{35}.

Textualized form. Using the template "Among the two following events which one is more closely related to {TARGET}: {OPTION_1} or {OPTION_2}?":

Question: *Among the two following events which one is more closely related to "Albert switches on the kettle": "Albert reads the newspaper" or "Albert's alarm rings"? Answer with 'First Event' or 'Second Event'.*

Answer: *Second Event.*

A.3.3 L2 OBSERVATION EXAMPLE

Symbolic form. Same subgraph as the simple example, but with intervention $do(X_{69}{=}0)$, which severs incoming edges to X_{69} and forces its value to 0 regardless of its parents. Under this intervention: $X_6 = f_6(X_{69}) = 0$.

Textualized form. Using question template "Given that {CAUSES}, does {EFFECT}. Assume appropriate interventions have been made so that {INTERVENTIONS}.":

Question: *Given that Albert's alarm rang and Albert did not stretch his back, does Albert wash his hair. Assume appropriate interventions have been made so that Albert did not switch on the kettle.*

Explanation: *It was intervened that Albert did not switch on the kettle. Albert did not switch on the kettle, hence Albert did not wash his hair.*

Answer: *No.*

A.4 TEMPLATES FOR GENERATING TEXTUALIZED OBSERVATIONS

All observations are textualized using randomly selected templates from the sets below. Placeholders are filled with groundings drawn from synonym sets and appropriate tense variants. `{CAUSES}` is filled with past-tense groundings of frontier variables, `{EFFECT}` with simple-present-tense grounding of the target, `{TARGET}`, `{OPTION_1}`, `{OPTION_2}` with primary-tense groundings, and `{INTERVENTIONS}` with primary-tense groundings of intervened variables.

A.4.1 SIMPLE OBSERVATION TEMPLATES

Question templates.

1. "Given that {CAUSES}, does {EFFECT}?"
2. "Does {EFFECT}, given that {CAUSES}?"

Explanation templates (cause-first, multi-cause).

1. "{CAUSES}, so {EFFECT}."
2. "{CAUSES}, therefore {EFFECT}."
3. "{CAUSES}, thus {EFFECT}."
4. "{CAUSES}, hence {EFFECT}."
5. "{CAUSES}, which led to {EFFECT}."
6. "{CAUSES}, resulting in {EFFECT}."
7. "{CAUSES}, as a result {EFFECT}."

Explanation templates (effect-first, multi-cause).

1. "{EFFECT}, because {CAUSES}."
2. "{EFFECT}, since {CAUSES}."
3. "{EFFECT}, which happened because {CAUSES}."
4. "{EFFECT}, this is because {CAUSES}."

For single-cause steps, only "so/hence" (cause-first) or "because/since" (effect-first) variants are used. The explanation order (cause-first or effect-first) is chosen uniformly at random for each step independently.

A.4.2 L1 Observation Templates

1. "Among the two following events which one is more closely related to "{TARGET}": "{OPTION_1}" or "{OPTION_2}"? Answer with 'First Event' or 'Second Event'."
2. "Out of the two following events: "{OPTION_1}" and "{OPTION_2}", which one is more closely related to "{TARGET}"? Answer with 'First Event' or 'Second Event'."

The order of the related and distractor variables in the option slots is randomized.

A.4.3 L2 Observation Templates

Question templates.

1. "Given that {CAUSES}, does {EFFECT}. Assume appropriate interventions have been made so that {INTERVENTIONS}."
2. "Does {EFFECT}, given that {CAUSES}. Assume appropriate interventions have been made so that {INTERVENTIONS}."
3. "Assuming interventions have been made so that {INTERVENTIONS}, does {EFFECT} given that {CAUSES}?"

Explanation templates are identical to those used for simple observations. The explanation begins by stating the intervened value(s), then proceeds with the same cause-first or effect-first derivation from the intervention point to the target.

B Fine-tuning and Evaluation Details

All models are fine-tuned using supervised fine-tuning (SFT) with low-rank adaptation (LoRA). We use the same hyperparameter configuration across all four models and all experiments unless otherwise stated. Table 3 summarizes the training hyperparameters and Table 4 summarizes the LoRA configuration.

Evaluation. During evaluation, we generate completions greedily (`do_sample`=False, temperature $= 0.1$) with a maximum of 512 new tokens and a batch size of 32. Generated textual explanations are parsed back into symbolic form by template matching against the grounding set, after which they are compared to the unique ground-truth symbolic explanation.

Models. We fine-tune four models: *Gemma-3-1B*, *Llama-3.2-1B*, *Qwen2.5-1.5B*, and *Qwen3-0.6B*. All models are loaded in bfloat16 precision with Flash Attention 2 where supported. Each model is fine-tuned independently per experiment configuration, producing a separate LoRA adapter.

Hardware. All experiments are run on a single NVIDIA GPU. Training a single model–configuration pair typically requires one GPU.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2×10^{-4}
LR scheduler	Cosine
Weight decay	0.01
Warmup steps	100
Training epochs	3
Per-device batch size	32
Gradient accumulation steps	1
Effective batch size	32
Max sequence length	Dynamic (no truncation)
Precision	bfloat16
Gradient checkpointing	True
Seed	42

Table 3: SFT training hyperparameters.

Hyperparameter	Value
Rank (r)	64
Alpha (α)	128
RSLoRA	True
Target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`
Dropout	0.05

Table 4: LoRA adapter configuration.