SELECTIVE ENFORCEMENT OF ORDER-INVARIANT CAUSAL REASONING IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

If we accept the statements *(A causes B, B causes C)*, then conclusions we draw from these relations should not depend on the order of presentation. The reordered sequence *(B causes C, A causes B)* describes the same causal graph and should therefore yield identical downstream judgments. We refer to this requirement as *order-invariant causal consistency*. Prior work has shown that language models violate this requirement in a variety of contexts, particularly when asked to reason about hypothetical outcomes.

We introduce a methodology for selective enforcement of causal constraints in language models, and apply it to this problem. We first construct a narrowly targeted diagnostic – the Textual Causal Invariance Test (TCIT) – to isolate failures of order-invariant consistency. We then apply a lightweight training procedure that penalizes order-dependent preferences and reinforces order-invariant reasoning.

Implemented on the open-weight Phi-3 model, this intervention raises TCIT accuracy from 59% (modestly above chance) to 98%, without degrading performance on a suite of regression tests. Furthermore, we demonstrate zero-shot transfer to the natural-language CLadder benchmark, yielding statistically significant improvements specifically on Rung-3 (counterfactual) causal reasoning tasks, with no degradation on lower causal rungs.

These results demonstrate that violations of order-invariant causal consistency can be isolated and corrected through targeted enforcement of a single structural constraint. More broadly, they suggest that selectively enforcing well-defined causal principles may provide a practical path toward improving causal reasoning in language models.

1 INTRODUCTION

When a language model processes a sequence of causal statements, users expect it to extract the underlying causal graph and reason from that structure. Consider the following two descriptions:

> **Version 1:** "Rainfall causes flooding. Flooding causes road closures. Road closures cause traffic delays."
> **Version 2:** "Road closures cause traffic delays. Rainfall causes flooding. Flooding causes road closures."

Both texts describe the identical causal chain: Rainfall $\rightarrow$ Flooding $\rightarrow$ Road Closures $\rightarrow$ Traffic Delays. Given the hypothetical query "Suppose there was no rainfall. Would there be traffic delays?" the answer must be "No" regardless of which version is presented. More formally, under Pearl's structural causal model framework (Pearl, 2009), such queries are evaluated on the underlying graph structure, which remains invariant to the order of its linguistic presentation. We term this requirement *order-invariant causal consistency*.

Transformer-based language models (LMs) struggle with causal reasoning in general (Liu et al., 2025). Failures persist across model families and prompting strategies, supporting the contention that LMs act as "causal parrots" rather than internalizing causal structure (Zečević et al., 2023). However, not all causal reasoning failures are created equal. Some are knowledge-driven, based in missing, conflicting, or misinterpreted world knowledge. Others arise in the application of explicit

causal structure: even when the causal relations are clearly specified, the model may apply them incorrectly or inconsistently. Order-invariance violations fall into this second group.

We maintain that consistent application of explicit causal structure is a foundational requirement for robust causal reasoning. However, existing causal reasoning benchmarks evaluate broad performance, conflating multiple sources of error including knowledge conflicts, graph complexity, and linguistic ambiguity. While these benchmarks provide useful assessments of capability, they offer limited guidance for identifying and correcting specific structural inconsistencies.

In this paper, we introduce a deliberately narrow diagnostic and training methodology based on *selective enforcement* of a specific structural constraint. We isolate a single requirement — order-invariant causal consistency — and construct a controlled diagnostic test to expose violations. We then apply a targeted training procedure that directly optimizes the model's preference between correct and incorrect continuations, aligning the objective with the diagnostic signal rather than relying on standard token-wise supervised fine-tuning. The resulting trained model achieves near-perfect performance on this diagnostic and improves performance on the broader CLadder benchmark (Jin et al., 2023).

Order-invariant consistency is far from the only component of causal reasoning. Yet by isolating this constraint and enforcing it through targeted training, we demonstrate that our selective enforcement approach can correct a structural failure and yield measurable improvements on broader causal reasoning tasks. More broadly, our results suggest that selective enforcement of well-defined causal constraints may offer a practical path toward improving causal reasoning in language models.

2 Structural Failure Modes and Selective Enforcement

2.1 Knowledge-Driven versus Structural Failure Modes

Knowledge-related errors occur when a model lacks relevant facts, misapplies background knowledge, or fails to override entrenched priors. Such effects are especially visible in specialized domains. Lee et al. (2025) show that language models struggle with scientifically validated causal relationships in economics and health, and Yamin et al. (2024; 2025) demonstrate that models often default to stored priors when contextual premises conflict with parametric knowledge.

However, failures persist even when the causal relations are explicitly specified and world knowledge is minimized. Synthetic benchmarks using semantically neutral entities show that models can misapply or inconsistently apply provided causal structure (Chen et al., 2025; Joshi et al., 2024). In such cases, the problem is not missing knowledge, but instability in applying structural constraints.

We do not claim that these sources of error are cleanly separable within model representations. Rather, the distinction drives our methodology: isolating settings in which causal structure is explicit allows targeted intervention on specific structural inconsistencies.

2.2 Selective Enforcement

We define *selective enforcement* as a methodology for correcting specific structural inconsistencies in language models. Rather than optimizing broad task performance, selective enforcement isolates a single, explicitly defined constraint, operationalizes it as a diagnostic test, and applies training pressure directly aligned with the diagnostic signal. The goal is not to solve causal reasoning in general, but to correct a precisely defined failure mode under clearly-defined conditions. In the present work, we employ this methodology on the constraint of order-invariant causal consistency.

3 Related Work

Recent work has shown that language models violate structural invariances: their responses to semantically equivalent inputs change when those inputs are reordered or rephrased. Existing causal benchmarks measure overall performance on broad task suites, but aggregate scores mix together distinct sources of error – including world knowledge, structural inconsistency, and linguistic ambiguity – making it difficult to isolate any one failure mode.

Our work differs in focus. Rather than measuring broad causal competence, we isolate a single structural invariance and introduce a training procedure designed specifically to enforce it.

3.1 Order Sensitivity in LLMs

Language models are known to produce inconsistent outputs when semantically equivalent inputs are presented in different orders. Yoon et al. (2025) document this as a general failure mode across diverse tasks and models. In causal reasoning tasks, order sensitivity takes specific forms. Joshi et al. (2024) find that LLMs apply a *position heuristic*: when event X is consistently mentioned before event Y in training text, models infer that X causes Y based on mention order alone, independent of any actual causal information. Yamin et al. (2024) similarly demonstrate performance degradation under permutations of causal facts in chains.

3.2 Causal Reasoning Benchmarks

Recent benchmarks evaluate counterfactual reasoning in language models with varying emphases on complexity, naturalism, and diagnostic clarity. CLadder (Jin et al., 2023) tests natural-language reasoning across Pearl's causal hierarchy (association, intervention, counterfactuals); CounterBench (Chen et al., 2025) provides comprehensive evaluation using synthesized statements (nonsense nouns) across diverse graph topologies and finds that the majority of failures occur at the *inference step* — correctly extracting the graph but operating on it incorrectly — rather than in graph comprehension; Yamin et al. (2025) evaluate whether LLMs can selectively integrate contextual counterfactual premises with parametric knowledge.

Our contribution, the Textual Causal Invariance Test (TCIT), prioritizes diagnostic clarity over scope: we test a single semantic property (order-invariance) on the simplest possible graphs (4-node chains) to isolate this failure mode without confounds from graph complexity or world knowledge.

3.3 Methods for Improving Counterfactual Reasoning

Prior work includes a variety of methods for improving counterfactual reasoning.

Prompting strategies. Chain-of-thought prompting (Jin et al., 2023) and structured reasoning templates aim to elicit better counterfactual reasoning without parameter updates. These methods are model-agnostic and can improve performance on many tasks.

Fine-tuning. Yamin et al. (2025) fine-tune LLMs in an attempt to address a knowledge-based failure mode: whether models can selectively combine parametric knowledge with contextual counterfactual premises (e.g., "If Paris were in Italy, where would the Eiffel Tower be?"). A brief fine-tuning experiment delivered only limited and inconsistent performance improvements.

3.4 Relation to Causal Inference, Representation Learning

We assume the causal graph is known and given textually, and focus on whether language model *outputs* respect its semantics. This is distinct from causal inference (Hernán & Robins, 2024), which estimates effects from data, and from causal representation learning (Schölkopf et al., 2021), the discovery of high-level causal variables from unstructured input data. Our focus on order invariance is also different from environment-invariance (Peters et al., 2016; Arjovsky et al., 2019): rather than stability across distributional shifts, we require consistency across surface-form revisions that *should not* affect semantic interpretation. While IRM-inspired penalties enforce invariance by penalizing environment-dependent gradient directions, our constraint is sufficiently explicit to admit a simpler behavioral approach: direct supervision over all surface-form realizations of the same causal structure.

4 Method: The TCIT Diagnostic Test

We introduce the Textual Causal Invariance Test (TCIT), a diagnostic benchmark for measuring order-invariant counterfactual revision in language models. TCIT consists of two variants, sharing

TCIT-F (factual direction preference)	TCIT-H (hypothetical intervention)
World semantics: An event occurs if and only if its direct cause occurs. There are no other causes. Blorf causes Tward. Tward causes Quux. Quux causes Zant. Continuations: - *Correct*: Therefore, Blorf causes Zant. *Incorrect*: Therefore, Zant causes Blorf.	World semantics: An event occurs if and only if its direct cause occurs. There are no other causes. Blorf causes Tward. Tward causes Quux. Quux causes Zant. Suppose Blorf did not occur. Continuations: - *Correct*: Therefore, Zant did not occur. *Incorrect*: Therefore, Zant occurred.

Table 1: Canonical TCIT-F and TCIT-H formats. We score by log-probability preference between the two fixed continuations under teacher forcing (no generation), and we test invariance by permuting the order of the same causal clauses. Modern LLMs score near-perfectly on TCIT-F but frequently fail on TCIT-H.

identical causal structure, differing only in query type. The baseline TCIT-F tests a model's response to *factual* causal questions, while TCIT-H tests premises under hypothetical intervention. Modern language models perform near-perfectly on TCIT-F, but make frequent errors on TCIT-H.

4.1 Design

TCIT employs four simplifications in order to prioritize diagnostic clarity over semantic breadth:

Synthetic Entity Names. Variables use nonsense words (e.g., *Blorf*, *Tward*, *Quux*) to eliminate world knowledge confounds, forcing models to derive causal relationships from explicit text rather than semantic priors.

Deterministic Causal Semantics. All relationships are deterministic ($A = 1 \Rightarrow B = 1$), eliminating probabilistic confounds and isolating structural counterfactual inference. We test whether models respect *declared* causal structure, not uncertainty handling.

Minimal Graph Complexity. All graphs are simple 4-node chains ($A \to B \to C \to D$), eliminating topological complexity as a confound and ensuring failures reflect order-sensitivity rather than graph reasoning difficulty. This also makes ground truth labels mechanically obvious.

Explicit Order Permutations. For each chain, we generate all $K!$ clause order permutations while holding graph structure fixed, creating matched examples that differ only in surface form to directly measure order-invariance violations.

These choices result in a useful tool for isolating order-sensitivity in counterfactual reasoning with minimal confounds.

4.2 Prompt Construction

Both TCIT-F and TCIT-H use identical ordering perturbations to test order-invariance. For each item, we generate all $K!$ permutations of clause order:

- *Blorf causes Tward. Tward causes Quux. Quux causes Zant.* (sequential)
- *Quux causes Zant. Tward causes Quux. Blorf causes Tward.* (reverse)
- *Tward causes Quux. Blorf causes Tward. Quux causes Zant.* (permuted)
- ... (all other orderings)

All permutations describe the same causal graph $\varphi(t) = (A \to B \to C \to D)$ and should yield identical answers to the query (whether factual direction preference in TCIT-F or hypothetical outcome in TCIT-H). A model satisfying order-invariant causal reasoning will produce consistent predictions across all permutations; violations manifest as **prediction flips** when only the presentation order changes.

4.3 TCIT-F: FACTUAL CAUSAL ENTAILMENT

TCIT-F serves as a **control benchmark**. After presenting the causal graph and a factual direction preference query, we score the model's preference between two fixed continuations that state opposite causal directions for a queried pair (Table 1).

High accuracy on TCIT-F indicates that the model can extract causal structure from text when reasoning requirements are minimal. If a model fails TCIT-F, there is no point testing TCIT-H; the causal information is not even being encoded.

As we show in Tables 2 and 5, modern language models perform near-perfectly on TCIT-F. Our frozen Phi-3-mini achieves 98.15% (Method 1, teacher-forced, n=2000; Table 5) and 99.0% under forced-choice scoring (n=200; Table 2); GPT-4o scores perfectly under both methods. The failure occurs in TCIT-H, where compositional counterfactual reasoning is required.

4.4 TCIT-H: HYPOTHETICAL INTERVENTION

TCIT-H tests hypothetical interventions under the same causal structures. We append a hypothetical premise (e.g., "Suppose X did not occur."), and score the model's preference between two fixed continuations about whether an outcome event occurred (Table 1). This requires revision and multi-hop composition under the declared world semantics. TCIT-H uses intervention-style hypothetical premises in a deterministic chain; it is not claimed to instantiate full Pearlian counterfactual inference with abduction.

4.5 METRICS AND EVALUATION CRITERIA

For open-weight models, TCIT performance is scored using a *log-probability preference* criterion under teacher forcing. That is, the model is presented with two candidate continuations, only one of which is the valid expression. The log-probability of the incorrect continuation is subtracted from the log-probability of the correct continuation. This is equivalent to scoring the model's preference for the correct continuation *relative to the incorrect continuation*.

More precisely, for each prompt x with candidate continuations y^+ (correct) and y^- (incorrect), we compute the log-likelihood margin

$$m(x) = \log P(y^+ \mid x) - \log P(y^- \mid x).$$

An ordering is counted as correct if $m(x) > 0$; ties ($m(x) = 0$) are treated as incorrect. We then aggregate results and report the following metrics: (i) **positive rate**, the fraction of orderings with $m(x) > 0$; (ii) **mean margin**, the average value of $m(x)$; (iii) **within-item standard deviation**, measuring variability of margins across orderings of the same item; and (iv) **flip rate**, defined as the fraction of orderings whose predicted label disagrees with that of a fixed reference ordering for the same underlying item (the canonical forward ordering).

TCIT-H chance-level performance corresponds to a positive rate of approximately 0.5; deviations from this baseline are therefore meaningful indicators of order sensitivity or invariance. "Perfect" performance corresponds to a positive rate of 1.0.

A flip rate of 0% indicates perfect order-invariance (predictions depend only on causal structure, not presentation). A flip rate of 5% means 1 in 20 causal chains yields inconsistent predictions across orderings—a rate that, while seemingly modest as an error rate, constitutes a serious semantic consistency violation. Order-invariance is a hard constraint: just as commutativity requires $2 + 3 = 3 + 2$ with probability 1.0, causal semantics requires $P(Y \mid do(X), t) = P(Y \mid do(X), \pi(t))$ whenever $\varphi(t) = \varphi(\pi(t))$. Any violation represents a category error distinct from mere inaccuracy.

For closed-weight models (e.g., GPT-4o, Claude) where full model access is unavailable, we use alternative scoring methods that approximate the log-probability preference criterion. These methods are detailed in Section A.2 and are indicated in result tables via footnote symbols (Tables 2 and 3).

4.6 TCIT RESULTS ON SELECTED MODELS

We summarize TCIT performance across a small set of representative closed-weight models and a Phi-3 baseline in Table 2 (TCIT-F) and Table 3 (TCIT-H). TCIT-F is near ceiling for all evaluated

Model	n	pos. rate ↑	flip rate ↓	mean margin ↑
Phi-3-mini-4k-instruct (baseline)	200	0.990[‡]	0.011	0.577
claude-3-5-haiku-20241022	200	0.992[§]	0.008	1.718
gpt-4o	200	1.000[‡]	0.000	9.476
gpt-4-turbo	50	1.000[‡]	0.000	10.091
gpt-4o-mini	200	1.000[‡]	0.000	12.497

Table 2: **TCIT-F (factual) results on selected models.** Each row reports the most faithful available scoring method: [†] teacher-forced continuation logprob; [‡] forced-choice direct (next-token logprobs for A/B); [§] forced-choice sampling (choice frequencies under stochastic decoding). Positive rate is the fraction of evaluated orderings with margin $m(x) > 0$ (ties treated as incorrect); flip rate measures disagreement against a fixed reference ordering for the same item.

Model	n	pos. rate ↑	flip rate ↓	mean margin ↑
gpt-4o-mini	200	0.477[‡]	0.086	0.630
Phi-3-mini-4k-instruct (baseline)	200	0.612[§]	0.215	0.240
gpt-4o	200	0.693[‡]	0.007	6.265
gpt-4-turbo	50	0.846[‡]	0.000	6.076

Table 3: **TCIT-H (hypothetical) results on selected models.** Same reporting conventions as Table 2. TCIT-H is scored via preference for the correct counterfactual continuation under order perturbations; higher positive rate and lower flip rate indicate stronger order-invariant counterfactual revision. Note: the Phi-3 baseline row uses Method 3 ([§] sampling-based, $n = 200$), while the primary TCIT-H results in Table 4 use Method 1 (teacher-forced logprob, 2000 items). The difference in flip rate (0.215 here vs. 0.052 in Table 4) reflects this methodological difference rather than a discrepancy in underlying model behavior; Table 9 shows that positive rate and rank ordering agree closely across methods.

models, indicating that factual direction preference is largely preserved under forced-choice evaluation. In contrast, TCIT-H reveals substantial variation: GPT-4o and GPT-4-turbo remain above chance with low flip rates, while GPT-4o-mini falls *below* chance and exhibits higher order sensitivity; the Phi-3 baseline is only modestly above chance with comparatively high flip rate. Notably, even the near-state-of-the-art GPT-4o model chooses the incorrect continuation more than 15% of the time.

5 Method: TCIT-Aligned Training

The diagnostic results in Section 4.6 establish a specific failure mode. TCIT-F shows that causal direction information is available for factual queries, while TCIT-H exposes a breakdown of order-invariant hypothetical revision. Taken together, these findings suggest that this failure arises not from missing causal signal, but from how that signal is expressed and composed under intervention. Our goal is therefore to enforce a constraint that will induce the model to exploit extant causal direction information in order to respect order invariance.

We align training directly with the TCIT diagnostic by optimizing a preference objective over fixed continuations (Burges et al., 2005; Rafailov et al., 2023) derived from the specified causal graph. Adaptation is performed using parameter-efficient LoRA modules (Hu et al., 2022), leaving base model weights frozen.

5.1 Graph-Aligned Preference Objective

Each training instance consists of (i) a short narrative describing a length-4 causal chain, rendered as three clauses, and (ii) two fixed candidate continuations: a *correct* continuation y^+ consistent with

the chain's intervention semantics and an *incorrect* continuation y^- that contradicts it. The objective is to increase the model's relative preference for y^+ over y^-.

Let x denote the prompt. We define the log-likelihood margin

$$m(x) = \log P_\theta(y^+ \mid x) - \log P_\theta(y^- \mid x),$$

and optimize the standard margin-based preference loss (Burges et al., 2005)

$$\mathcal{L}_{\text{pref}}(x; \theta) = \text{softplus}(-m(x)),$$

with log-likelihoods computed under teacher forcing. This objective directly optimizes the relative margin used for TCIT evaluation; standard token-wise SFT maximizes $\log P(y^+ \mid x)$ alone and does not explicitly penalize high probability assigned to y^-.

To encourage order invariance, we apply order augmentation: for each underlying chain we train on $K = 3$ clause-order realizations (forward, reversed, and one random permutation) with identical labels (y^+, y^-). Evaluation uses all 10 orderings per chain (Appendix A), including those not seen during training.

A secondary TCIT-F factual objective is included as an anchor to preserve baseline causal direction performance; exact task-mixture weights are in Appendix B.

5.2 Training Setup

We fine-tune `Phi-3-mini-4k-instruct` using LoRA adapters (rank 16, $\alpha = 32$, dropout 0.05) applied to attention projection layers, with all base parameters frozen.

Training data consist of synthetic 4-node chains ($A \to B \to C \to D$) with disjoint entity pools across splits. Interventions are applied at the root node (A), with queries always targeting the leaf (D). Examples are balanced over outcome polarity: half present an all-on baseline (do(A=0), correct: "D did not occur") and half an all-off baseline (do(A=1), correct: "D occurred"). We choose this minimal intervention as the simplest test of order-invariant revision; we leave interventions at internal nodes and mixed-target queries to future work. Full hyperparameters and data-quality checks are in Appendix B.

6 Results

6.1 TCIT-H and TCIT-F

We evaluate order-invariant hypothetical revision using TCIT-H under the log-probability preference criterion defined in Section 4.5. In addition to the primary positive-rate metric, we report flip rate, mean margin, and within-item variability to characterize both accuracy and sensitivity to narrative order.

Table 4 reports TCIT-H results for the frozen baseline model and for TCIT-Aligned Training. The frozen baseline exhibits a positive rate only modestly above chance, together with nontrivial order sensitivity.

In contrast, TCIT-Aligned Training yields a dramatic increase in positive rate together with a substantial reduction in flip rate. The resulting performance is near ceiling and stable across evaluation seeds, indicating that the model consistently prefers the graph-consistent continuation regardless of surface order. Importantly, this improvement is achieved without introducing new causal information, but by constraining how existing causal signal is expressed under counterfactual intervention.

To assess whether our intervention degrades factual causal knowledge, we evaluate both the baseline and trained models on TCIT-F, which tests factual causal entailment under the same causal structures but without hypothetical intervention. As expected given the high baseline (98.15%), TCIT-F performance is near ceiling for both models; the trained model improves marginally to 99.96% (Table 5), confirming that the targeted training does not erode basic causal direction knowledge.

6.2 Zero-shot Transfer to CLadder

To assess whether TCIT-aligned training transfers beyond synthetic TCIT prompts, we evaluate on CLadder (Jin et al., 2023), a natural-language benchmark spanning Pearl's causal hierarchy: Rung 1

Model	pos. rate ↑	flip rate ↓	mean margin ↑	within-item std ↓
Phi-3 (frozen baseline)	0.5911	0.0522	-0.094957	0.109170
Phi-3 + TCIT-Aligned Training	**0.98483** ± 0.00080	**0.01579** ± 0.00108	**0.72589** ± 0.00630	**0.12007** ± 0.00165

Table 4: TCIT-H results (2000 items, 10 orderings per item). Positive rate measures fraction of orderings with positive log-probability margin (1.0 = perfect, 0.5 = chance on balanced binary task). Flip rate measures disagreement with a canonical reference ordering for the same underlying item (Section 4.5). The TCIT-Aligned row reports the mean over five evaluation seeds $\{1, 2, 3, 4, 42\}$.

Model	pos. rate ↑	flip rate ↓	mean margin ↑	within-item std ↓
Phi-3 (frozen baseline)	0.9815	0.0186	0.46963	0.16516
Phi-3 + TCIT-Aligned Training	**0.99955** ± 0.00022	**0.00045** ± 0.00022	**1.22202** ± 0.00330	**0.20222** ± 0.00180

Table 5: TCIT-F results (2000 items, 10 orderings per item). TCIT-F serves as a control benchmark for factual causal direction preference (no counterfactual intervention). The TCIT-Aligned row reports the mean over five evaluation seeds $\{1, 2, 3, 4, 42\}$.

(association), Rung 2 (intervention), and Rung 3 (counterfactual). We evaluate both the frozen baseline and TCIT-aligned Phi-3 on the standard balanced CLadder dataset and the nonsense-words variant, pooling results for Rung 3 ($n = 7{,}632$) to increase statistical power.

As shown in Table 6, TCIT-aligned training yields a statistically significant improvement of $+2.24$pp on **Rung 3 (counterfactual)** questions ($p = 0.0053$, 95% CI $[0.67, 3.81]$), with no significant change on Rungs 1–2 (association and intervention). Results are consistent across variants: +2.27pp on the balanced set ($p = 0.047$) and +2.21pp on the nonsense-words set ($p = 0.050$).

For context, Table 7 reports published CLadder results from Jin et al. (2023) alongside our Phi-3 results. This comparison is indicative rather than controlled, as evaluation protocols and prompting strategies differ. Notably, the TCIT-aligned improvement is comparable in magnitude to the Rung 3 gains reported for CausalCoT (Jin et al., 2023), while remaining more targeted in scope.

TCIT-aligned training produces a *selective* improvement on CLadder Rung 3 without affecting lower rungs, providing evidence that correcting a specific structural failure mode transfers to natural-language counterfactual reasoning.

6.3 Robustness and Stress Tests

We re-evaluate TCIT-H across multiple evaluation seeds corresponding to different random draws of entity names and ordering permutations. Results are reported as mean±standard deviation for positive rate and flip rate, and are included in Table 4.

In addition, we run a lightweight regression suite to ensure that TCIT-Aligned Training does not introduce pathological behavior or degrade basic instruction-following capabilities. The suite includes four checks: (A) forced-choice `Yes`/`No` sanity items scored via log-probability preference; (B) lightweight instruction-following format checks (non-empty responses, short factual QA, list compliance, and simple continuations); (C) average negative log-likelihood on short held-out passages; and (D) simple degeneration heuristics (repetition and token diversity) on brief generations. Across all four regression checks, TCIT-Aligned Training matches or improves upon the frozen baseline, with no evidence of instruction-following or text-quality regressions. *(Suite description and summary results are in Appendix C.)*

7 Discussion

7.1 Analysis

The dramatic improvement from TCIT-Aligned Training (59% → 98%) contrasts with limited success in prior fine-tuning (Yamin et al., 2025). We attribute this to the following factors:

Rung	Task	Baseline	TCIT-Aligned	Δ (pp)	p
1	Association	51.15%	51.45%	+0.30	0.736
2	Intervention	57.69%	56.73%	−0.96	0.274
3	Counterfactual	55.04%	57.29%	**+2.24**	**0.0053**

Table 6: CLadder transfer results for Phi-3 (3.8B parameters). Rung 3 combines balanced and nonsense-words variants (n = 7,632); Rungs 1–2 from the same runs ($n \approx$ 6,360 each). Two-sided binomial test; bold indicates $p < 0.01$.

Model	Rung 1	Rung 2	Rung 3
Published benchmarks (Jin et al., 2023):			
GPT-3.5	51.80%	54.78%	50.32%
GPT-4	63.01%	62.82%	60.55%
GPT-4 + CausalCoT	83.35%	67.47%	62.05%
This work (Phi-3, 3.8B parameters):			
Phi-3 Baseline	51.15%	57.69%	55.04%
Phi-3 + TCIT-Aligned	51.45%	56.73%	**57.29%**

Table 7: Indicative comparison to published CLadder results. Phi-3 is substantially smaller than GPT-3.5/4; CausalCoT denotes chain-of-thought prompting with causal scaffolding.

Explicit invariance signal. Training on $K = 3$ orderings per chain with identical labels provides direct gradient pressure: "predictions should not change when only surface order changes." This explicit consistency constraint avoids the shortcut learning observed in broader fine-tuning.

Task simplicity. Order-invariance is a narrower constraint than selective knowledge override or multi-hop inference. The model learns "extract graph, then reason from structure," not "decide when to override knowledge."

Information already present. TCIT-F baseline accuracy (98%) shows local causal signal is accessible. The failure is compositional: applying that information consistently under intervention.

7.2 Limitations

We highlight several limitations of the current work:

Narrow Scope. We study a single semantic property (order-invariance) in a highly controlled setting (deterministic binary chains). Many other semantic consistency requirements exist: transitivity of causation, commutativity of independent interventions, contradiction avoidance when causal statements conflict, etc. To apply our approach to each would require separate diagnostic and enforcement approaches.

Synthetic Setting. TCIT uses artificial entity names and template-generated text. While this enables controlled diagnosis, it limits direct applicability to naturalistic language. Our CLadder evaluation (Section 6.2) provides initial evidence for transfer: a small but statistically significant improvement on counterfactual (Rung 3) questions with no degradation on Rungs 1–2. However, the effect size is modest (+2.24pp), and broader transfer remains to be demonstrated.

Single Model. All experiments use Phi-3-mini. While order-sensitivity failures appear general based on prior work (Yamin et al., 2025), the specific effectiveness of TCIT-Aligned Training may be architecture-dependent.

Deterministic Assumption. Real causal reasoning often involves probabilistic relationships, confounders, and selection bias. Our framework assumes away these complications for diagnostic clarity.

No Guarantees Beyond Training Distribution. We encourage order-invariance through exposure to multiple orderings during training, not through an explicit architectural constraint. Behavior on unseen surface variations (novel paraphrases, longer chains, different causal connectives) is not guaranteed.

Further Ablations. The current results do not report comparisons against: (i) supervised fine-tuning on correct continuations without a pairwise preference loss; (ii) order augmentation alone, without the TCIT-F anchor task; or (iii) alternative preference objectives such as DPO (Rafailov et al., 2023) or a positive margin target ($\epsilon > 0$). Disentangling these factors is necessary to determine which components are load-bearing, and we identify these ablations as future work.

7.3 Future Work

Immediate priorities include ablations to isolate the contribution of pairwise preference loss, order augmentation, and the TCIT-F anchor task. The procedure should be validated on additional open-weight models (e.g. Llama), and broader transfer should be evaluated on additional benchmarks (e.g., CounterBench, knowledge-conflict scenarios). Longer-term, we plan to extend the diagnostic-and-enforce methodology to richer causal structures (conjunctions, confounders, longer chains) and to other semantic invariances such as transitivity and commutativity of independent interventions.

8 Conclusion

We introduced *selective enforcement*: a methodology for diagnosing and correcting structural failures in causal reasoning, and applied it to enforce *order-invariant causal consistency*. We constructed a diagnostic test (TCIT) to isolate the failure, and a graph-aligned preference objective to correct it.

This intervention dramatically reduced order-dependent behavior, yielding near-ceiling performance on the diagnostic without degrading factual causal judgments. In addition, the resulting model exhibited a selective, statistically significant improvement on CLadder's counterfactual (Rung 3) questions, with no regression on lower causal rungs. These findings indicate that at least some causal reasoning failures arise from inconsistent application of explicit structure rather than from missing causal information.

Order-invariant consistency is only one component of causal reasoning. However, by isolating and enforcing this constraint under controlled conditions, we demonstrate that targeted correction of a structural failure is both feasible and effective. More broadly, this work suggests that selectively enforcing well-defined causal invariances may provide a practical and modular approach to improving structural reasoning reliability in language models.

Reproducibility Statement

The TCIT benchmark, CLadder evaluation code, training scripts, and all synthetic datasets used in this paper are available in the anonymized repository at `https://anonymous.4open.science/r/tcit-benchmark-E2FE`. The repository includes the data synthesis pipeline (allowing full regeneration of TCIT-F and TCIT-H splits with disjoint entity pools), the trainer implementation with the exact prompt and continuation formats shown in Appendix B.3, and the evaluation harness used for all reported metrics. All hyperparameters are reported in Appendix B. Primary TCIT results are averaged over five random seeds ($\{1, 2, 3, 4, 42\}$) with per-seed standard deviations reported. The LoRA adapter weights for the trained checkpoint used in this paper will be released alongside the code upon de-anonymization.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL `https://arxiv.org/abs/1907.02893`.

Chris J.C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005.

Yuefei Chen, Vivek K. Singh, Jing Ma, and Ruixiang Tang. CounterBench: A benchmark for counterfactuals reasoning in large language models. *arXiv preprint arXiv:2502.11008*, 2025.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman and Hall/CRC, 2 edition, 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. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. LLMs are prone to fallacies in causal inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10553–10569, 2024.

Donggyu Lee, Sungwon Park, Yerin Hwang, Hyoshin Kim, Hyunwoo Oh, Jungwon Kim, Meeyohng Cha, Sangyoon Park, and Jihee Kim. Benchmarking LLM causal reasoning with scientifically validated relationships. *arXiv preprint arXiv:2510.07231*, 2025.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. Large language models and causal inference in collaboration: A comprehensive survey. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7683–7699, 2025.

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

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78 (5):947–1012, 2016.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.18290.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Khurram Yamin, Shantanu Gupta, Gaurav Ghosal, Zachary Lipton, and Bryan Wilder. Failure modes of LLMs for causal reasoning on narratives. *arXiv preprint arXiv:2410.23884*, 2024.

Khurram Yamin, Gaurav Ghosal, and Bryan Wilder. LLMs struggle to perform counterfactual reasoning with parametric knowledge. In *ICML 2025 Workshop on Scaling Up Intervention Models*, 2025. arXiv:2506.15732.

Soyoung Yoon, Dongha Ahn, Youngwon Lee, Minkyu Jung, HyungJoo Jang, and Seung-won Hwang. RoToR: Towards more reliable responses for order-invariant inputs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

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. arXiv:2308.13067.

Variant	Unique chains	Total examples
TCIT-F (factual)	2000	20,000
TCIT-H (hypothetical)	2000	20,000
Total	2000	40,000

Table 8: TCIT dataset statistics. Both variants use 4-node deterministic causal chains with 10 orderings per chain.

A TCIT Details

A.1 Dataset statistics

TCIT is a controlled diagnostic benchmark rather than a comprehensive evaluation suite. Table 8 summarizes the dataset composition.

Each unique causal chain is presented in 10 different orderings. The orderings include forward, reverse, and additional permutations of the three causal clauses. All orderings of the same chain share the same ground-truth label, enabling direct measurement of order-invariance violations.

Training vs. evaluation. During training, we use $K = 3$ orderings per chain (forward, reverse, and one random permutation) to encourage consistency. During evaluation, we test against 10 orderings per chain to more thoroughly probe order sensitivity, including orderings not seen during training.

A.2 Scoring Methodologies

TCIT evaluation uses different scoring methods depending on model access. For open-weight models, we use teacher-forced log-probability scoring over full continuations. For closed-weight models, we use alternative methods that approximate this criterion. All three methods compute a preference margin $m(x)$ between correct and incorrect continuations, enabling consistent metric reporting across model types.

Method 1: Teacher-forced continuation log-probability (open-weight models). This is the primary method used for open-weight models (e.g., Phi-3). Given a prompt x and two candidate continuations y^+ (correct) and y^- (incorrect), we compute:

$$m(x) = \frac{1}{|y^+|}\sum_{i=1}^{|y^+|} \log P(y_i^+ \mid x, y_{<i}^+) - \frac{1}{|y^-|}\sum_{i=1}^{|y^-|} \log P(y_i^- \mid x, y_{<i}^-),$$

where log-probabilities are computed under teacher forcing (no generation) and we normalize by continuation length to avoid length bias. This method provides the most direct measurement of model preference and requires full access to model logits.

Method 2: Forced-choice direct (closed-weight models with logprob access). For closed-weight models that expose next-token log-probabilities via API (e.g., GPT-4), we format the prompt to present choices as "A: [correct continuation]" and "B: [incorrect continuation]". We then extract the next-token log-probabilities for tokens "A" and "B" directly from the model's logits in a single forward pass:

$$m(x) = \log P(\text{“A”} \mid x) - \log P(\text{“B”} \mid x).$$

This method is deterministic and provides a fast approximation of the preference margin, though it only captures the model's preference at the choice point rather than over the full continuation.

Method 3: Forced-choice sampling (closed-weight models without logprob access). For closed-weight models that only support text generation (e.g., Claude), we use a sampling-based approach. We format the prompt with explicit A/B choices and generate N samples (typically $N = 20$)

Method	Suite	Items	pos. rate	flip rate	mean margin
Method 1†	TCIT-F	50	0.980	0.020	0.469
	TCIT-H	50	0.638	0.082	0.096
Method 2‡	TCIT-F	200	0.990	0.011	0.577
	TCIT-H	96	0.597	0.120	0.194
Method 3§	TCIT-F	200	0.921	0.088	0.706
	TCIT-H	200	0.612	0.215	0.240

Table 9: **Phi-3 baseline results across scoring methods.** Comparison of frozen Phi-3-mini-4k-instruct evaluated using all three scoring methodologies on the same test set (seed 42). Method 1: teacher-forced continuation logprob; Method 2: forced-choice direct (A/B next-token logprobs); Method 3: forced-choice sampling (generation-frequency estimate, $N = 20$, $T = 0.7$). Positive rates and flip rates show close agreement across methods, confirming that Methods 2 and 3 provide faithful approximations of Method 1. Mean margins differ in absolute scale but preserve relative ordering.

with temperature $T = 0.7$. We count the frequency of choices "A" and "B" across samples, then estimate probabilities using Laplace smoothing:

$$\hat{P}(\text{“A”} \mid x) = \frac{\text{count}(\text{“A”}) + 1}{N + 2}, \quad \hat{P}(\text{“B”} \mid x) = \frac{\text{count}(\text{“B”}) + 1}{N + 2},$$

and compute the margin as:

$$m(x) = \log \hat{P}(\text{“A”} \mid x) - \log \hat{P}(\text{“B”} \mid x).$$

This method matches the evaluation protocol used by Anthropic for Claude models and enables fair comparison across closed-weight systems, though it introduces sampling variance.

Method selection and reporting. In result tables (Tables 2 and 3), we indicate the scoring method used for each model via footnote symbols: † for Method 1, ‡ for Method 2, and § for Method 3. For our primary results on Phi-3 (Section 6.1), we use Method 1 throughout.

Empirical comparison across methods. To validate that the three scoring methods yield comparable results, we evaluated the frozen Phi-3 baseline using all three methods on the same test set. Table 9 shows that while absolute values differ slightly (particularly for mean margin, which is sensitive to the scoring scale), the relative performance patterns are consistent across methods. Positive rates and flip rates, which are the primary metrics for order-invariance, show close agreement, confirming that the alternative methods provide faithful approximations of the log-probability preference criterion.

B Training Process and Examples

B.1 Training process

Preference loss and margin parameter. We train with a smooth pairwise ranking loss on the preference margin $m(x)$. A common variant introduces a target margin $\epsilon \geq 0$:

$$\mathcal{L}_{\text{pref}}(x; \theta) = \text{softplus}\big(-(m(x) - \epsilon)\big), \tag{1}$$

and we use $\epsilon = 0$ by default.

Task-mixture objective and schedules. Training uses a weighted mixture of TCIT-H (primary) and TCIT-F (anchor) preference losses:

$$\mathcal{L}(\theta; t) = w_{\text{cf}}(t)\, \mathbb{E}_{x \sim \mathcal{D}_{\text{cf}}}[\mathcal{L}_{\text{pref}}(x; \theta)] \; + \; w_{\text{f}}\, \mathbb{E}_{x \sim \mathcal{D}_{\text{f}}}[\mathcal{L}_{\text{pref}}(x; \theta)]\,. \tag{2}$$

We set $w_{\text{f}} = 0.3$ throughout. The TCIT-H weight $w_{\text{cf}}(t)$ is linearly ramped from 0.5 to 0.6 over the first 1000 optimization steps and then held constant.

Notation. $\mathcal{D}_{\text{cf}}$ and $\mathcal{D}_{\text{f}}$ denote the (synthetic) training datasets for TCIT-H and TCIT-F, respectively, and expectations are over prompts x drawn from those datasets.

B.2 Training procedure and controls (implementation details)

Data generation and order augmentation. Training data are synthesized from length-4 causal chains ($A \to B \to C \to D$) with disjoint entity pools across training, validation, and test splits. For each chain, we generate $K = 3$ surface orderings of the same three causal clauses: forward order, reversed order, and one additional random permutation sampled from the remaining permutations. TCIT-H examples cover counterfactual interventions applied at the chain root (A), with the query always targeting the leaf node (D). Examples are balanced over outcome polarity: half present an all-on baseline (A=1 by default, intervention do(A=0), correct outcome "D did not occur") and half an all-off baseline (A=0 by default, intervention do(A=1), correct outcome "D occurred"). TCIT-F examples provide factual causal entailment as an anchor objective.

Optimization. We train with AdamW (learning rate 10^{-4}, weight decay 0.01) using a cosine schedule with 100 warmup steps. Training is conducted in `bf16`. Key–value caching is disabled for forward passes (`use_cache=False`) to ensure consistent log-probability computation under teacher forcing. Effective batch size is controlled via gradient accumulation; multi-GPU runs use distributed data parallelism.

Data-quality checks. Before training, we validate: (i) disjoint entity pools across train/validation/test splits; (ii) ordering diversity per chain; (iii) balance over TCIT-H case types and intervention polarity; (iv) consistency of ground-truth labels across orderings; and (v) labeling sanity checks for TCIT-F direction preference.

B.3 Training examples (canonical vs. order-perturbed)

We include concrete examples below to show the *exact* formatting used in the trainer implementation: a prompt followed by `Continuations:` and a bullet prefix `-` (the continuation strings y^+ and y^- are appended under teacher forcing and are *not* generated). For each underlying chain, the trainer uses three clause-order realizations: **forward** $[0, 1, 2]$, **reversed** $[2, 1, 0]$, and one **random** permutation sampled from $\{[0, 2, 1], [1, 0, 2], [1, 2, 0], [2, 0, 1]\}$ (excluding forward and reversed).

TCIT-H (Task A, primary): counterfactual intervention. Canonical (forward ordering):

```
World semantics:  An event occurs if and only if its direct
cause occurs.  There are no other causes.
Bon causes Gist.  Gist causes Zab.  Zab causes Joriza.
Suppose Bon did not occur.
Continuations:
-
Correct  y+:  Therefore, Joriza did not occur.
Incorrect  y-:  Therefore, Joriza occurred.
```

Order-perturbed (reversed ordering):

```
World semantics:  An event occurs if and only if its direct
cause occurs.  There are no other causes.
Zab causes Joriza.  Gist causes Zab.  Bon causes Gist.
Suppose Bon did not occur.
Continuations:
-
Correct  y+:  Therefore, Joriza did not occur.
Incorrect  y-:  Therefore, Joriza occurred.
```

Order-perturbed (random ordering; example $[1, 0, 2]$**)**:

```
World semantics:  An event occurs if and only if its direct
cause occurs.  There are no other causes.
Gist causes Zab.  Bon causes Gist.  Zab causes Joriza.
Suppose Bon did not occur.
Continuations:
-
Correct  y+:  Therefore, Joriza did not occur.
Incorrect  y-:  Therefore, Joriza occurred.
```

TCIT-F (Task B, anchor): factual direction preference. In the current trainer, TCIT-F is included as an anchor task (loaded from `task_b_v1.jsonl`) with the same clause-order augmentation; the correct continuation always states the forward causal direction for the underlying chain.

Canonical (forward ordering):

```
World semantics:  An event occurs if and only if its direct
cause occurs.  There are no other causes.
Bon causes Gist.  Gist causes Zab.  Zab causes Joriza.
Continuations:
-
Correct  y+:  Therefore, Bon causes Joriza.
Incorrect  y-:  Therefore, Joriza causes Bon.
```

Order-perturbed (reversed ordering):

```
World semantics:  An event occurs if and only if its direct
cause occurs.  There are no other causes.
Zab causes Joriza.  Gist causes Zab.  Bon causes Gist.
Continuations:
-
Correct  y+:  Therefore, Bon causes Joriza.
Incorrect  y-:  Therefore, Joriza causes Bon.
```

Order-perturbed (random ordering; example $[1, 0, 2]$**)**:

```
World semantics:  An event occurs if and only if its direct
cause occurs.  There are no other causes.
Gist causes Zab.  Bon causes Gist.  Zab causes Joriza.
Continuations:
-
Correct  y+:  Therefore, Bon causes Joriza.
Incorrect  y-:  Therefore, Joriza causes Bon.
```

C Regression Suite

To verify that TCIT-Aligned Training does not introduce pathological behavior, we run a lightweight regression suite (`src/eval/regression_suite.py`) on both the frozen baseline and the trained checkpoint used in this paper. The suite is intended as a guardrail against gross regressions (e.g., answer collapse, format collapse, repetition collapse), not as an optimization target.

The suite includes four groups: (A) forced-choice `Yes`/`No` sanity checks scored via log-probability preference; (B) instruction-following format checks (non-empty responses, simple list compliance, short factual QA); (C) average negative log-likelihood (NLL) on short held-out passages; and (D) simple generation-quality heuristics (repetition and token diversity). We report suite-level summaries in Table 10.

D LLM Usage

Large language models were used in both the development of experiment code and the editing of this paper.

Model	Suite A acc. ↑	Suite B pass ↑	Suite C avg NLL ↓	Suite D rep. rate ↓
Phi-3 (frozen baseline)	0.90	0.62	1.114	0.00147
Phi-3 + TCIT-Aligned Training	0.92	0.62	1.112	0.00000

Table 10: Regression suite summary (see text). Suite A is forced-choice preference accuracy; Suite B is an overall instruction-following pass rate; Suite C reports mean NLL on short passages; Suite D reports a repeated 4-gram rate (lower is better). All numbers are computed from the stored suite outputs under `outputs/regression/`.

Claude was used to assist with the development of experiment code and the analysis of data. ChatGPT was used during the drafting process to improve prose clarity and assist with LaTeX formatting.

All research decisions — experimental design, choice of methodology, interpretation of results, and the framing of claims — were made by the human authors. No LLM-generated text was accepted without review and revision by the authors.