Counterfactual Answer Prediction Training: Improving LLMs' Causal Understanding via Input-Perturbation Difference Prediction

Anonymous authors
Paper under double-blind review

Abstract

Large language models can solve math problems but may lack true causal understanding: models that correctly solve a problem often cannot predict how localized perturbations to it affect the answer. We introduce Counterfactual Answer Prediction Training (CAPT), a fine-tuning method that trains LLMs to predict answer deltas ($A' - A$) under localized input perturbations, without re-solving the perturbed problem. A diagnostic study reveals a strong forward–counterfactual dissociation: Qwen3-8B solves 87.0% of GSM8K problems but predicts correct causal deltas on only 15.5% of relevant perturbations (dissociation score = 0.822), with the gap most severe for entity swap (10.0%) and condition negation (11.1%). We then construct a counterfactual training dataset of 4,817 triples from GSM8K and fine-tune with LoRA across four arms. Forward+CAPT (50/50 mix) substantially improves counterfactual competence (relevant CF accuracy: 15.5% $\rightarrow$ 31.9%, +106%, $p < 0.001$), improves localization of causally irrelevant edits to 100% (from 97.8%), and improves out-of-distribution forward reasoning (MATH-500: 33.5% $\rightarrow$ 42.0%, $p < 0.01$). Critically, standard data augmentation fails to improve causal competence (18.1% vs. 31.9%, $p < 0.01$ for the difference), indicating that the delta prediction objective—not mere exposure to perturbations—drives the gain. Scaling experiments across four model sizes (0.5B–8B) reveal a sharp capacity threshold: counterfactual competence remains flat at $\sim$9% for models up to 3B before jumping to 31.9% at 8B, while forward accuracy scales smoothly—providing additional evidence that forward and counterfactual reasoning are distinct capabilities.

1 Introduction

Large language models (LLMs) achieve impressive accuracy on math benchmarks such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), yet high accuracy does not imply true causal understanding: a model may arrive at the correct answer through pattern matching or memorized templates without grasping how problem inputs map to outputs. Genuine causal understanding should also predict how the answer would change if the problem were modified—a form of counterfactual reasoning.

We introduce counterfactual competence: the ability to predict how localized changes to a problem's inputs affect its answer without re-solving it. In Pearl's structural causal models (Pearl, 2009), this corresponds to predicting the interventional contrast $A(\mathrm{do}(\pi)) - A$, an ability distinct from—and not guaranteed by—forward problem-solving accuracy.

Our central hypothesis is that a forward–counterfactual dissociation exists in current LLMs: models that can solve problems often cannot predict how perturbations affect answers. We measure a substantial dissociation (score = 0.822 for Qwen3-8B on GSM8K), most severe for types requiring deep structural understanding (entity swap: 10.0%, condition negation: 11.1%).

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

To close this gap, we propose Counterfactual Answer Prediction Training (CAPT), which fine-tunes LLMs on (original problem, perturbation, answer delta) triples. Unlike standard augmentation (which re-solves perturbed problems), CAPT trains the model to predict only the difference in answers, forcing it to internalize the causal structure rather than memorize problem variants.

We position CAPT not as a new causal-discovery method, but as an analysis tool: counterfactual intervention serves as a lens for diagnosing which structural elements of a problem a model does and does not grasp, and we show that closing the measured gap also transfers to harder downstream reasoning.

Contributions.

1. We define and measure the forward–counterfactual dissociation in LLMs, showing it is substantial across perturbation types (dissociation score = 0.822 for Qwen3-8B), most severely for entity swap and condition negation.
2. We propose CAPT, a training method using (original, perturbed, answer-delta) triples that directly targets causal structure, and construct a dataset of 4,817 counterfactual triples from GSM8K.
3. We show that Forward+CAPT substantially improves counterfactual competence (15.5% $\rightarrow$ 31.9% relevant CF accuracy, a 106% relative gain, $p < 0.001$) and improves localization of causally irrelevant edits to 100% (from 97.8%).
4. We demonstrate that Forward+CAPT improves out-of-distribution forward reasoning (MATH-500: 33.5% $\rightarrow$ 42.0%, $p < 0.01$), suggesting causal understanding partially generalizes to unseen problems.
5. We provide evidence that the delta prediction objective—not mere data augmentation—drives the gains: CF augmentation achieves only 18.1% vs. 31.9% for Forward+CAPT ($p < 0.01$ for the difference; CF-augmentation vs. base: $p = 0.61$, not significant).
6. We provide ablations on perturbation types, mixing ratios and LoRA rank, plus scaling experiments across four model sizes (0.5B–8B) revealing a sharp capacity threshold for counterfactual competence: flat at $\sim$9% up to 3B, 31.9% at 8B, while forward accuracy scales smoothly.

2 Related Work

LLMs have made rapid progress on mathematical reasoning through chain-of-thought prompting (Wei et al., 2022b), self-consistency (Wang et al., 2023), and process reward models (Lightman et al., 2024), yet benchmarks such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) measure only forward problem-solving, not whether a model understands how changes to a problem's inputs propagate to its output; robustness studies (Huang et al., 2023; Shi et al., 2023) show brittleness to surface variations but study resolving perturbed problems rather than predicting their effect. Causal reasoning in NLP has been studied through causal-inference frameworks (Pearl, 2009), counterfactual reasoning (Feder et al., 2022), and intervention-based analysis (Vig et al., 2020), with a key debate over whether LLMs reason causally or exploit statistical correlations (Willig et al., 2022; Zečević et al., 2023). We target a specific, well-defined form of causal competence—predicting how localized perturbations to a math problem affect its answer—where the causal structure is fixed by arithmetic and evaluation is unambiguous. Concurrent work trains verifiably sensitive reasoning with causal objectives: Akter et al. (2025) regularize a model toward predictions consistent under structured interventions; our method is complementary, using counterfactual answer prediction as both a diagnostic of a model's grasp of intervention effects and a fine-tuning objective that transfers to harder downstream reasoning.

Perturbation-based augmentation is a standard robustness recipe: adversarial additions in reading comprehension (Jia & Liang, 2017), GSM8K-Hard (Huang et al., 2023) and GSM-IC (Shi et al., 2023) in math, and counterfactually-augmented data in NLI (Kaushik et al., 2020) and fairness (Kusner et al., 2017), with training on perturbed examples improving

generalization (Longpre et al., 2020). All of these train models to re-solve perturbed problems from scratch; CAPT instead trains on (original problem, perturbation, answer delta) triples, directly targeting prediction of intervention effects. We use LoRA (Hu et al., 2022) and related parameter-efficient methods (Han et al., 2024) to compare arms with the base model frozen. Finally, scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) predict smooth improvements with compute and data, while some capabilities emerge only above a size threshold (Wei et al., 2022a); our experiments show counterfactual competence flat ($\sim$9%) up to 3B and jumping to 31.9% at 8B while forward accuracy scales smoothly, suggesting counterfactual reasoning may be an emergent capability distinct from forward problem-solving.

3 Method

3.1 Counterfactual triples

Given a math problem P with correct answer A, a perturbation π produces a perturbed problem $P' = \pi(P)$ with answer A'. The counterfactual triple is $(P, \pi, \Delta A)$ where $\Delta A = A' - A$.

The key insight is that ΔA captures the causal effect of the perturbation π on the answer. In the framework of Pearl's structural causal models (Pearl, 2009), ΔA corresponds to the interventional contrast: $A' - A = A(\text{do}(\pi)) - A$. A model that understands the causal structure of P should be able to predict ΔA without fully solving P', by reasoning about how the specific perturbation propagates through the solution structure.

3.2 Perturbation types

We define five perturbation types, each probing a different structural element:

1. Numeric (n = 1,784): change a quantity (e.g., "5 apples" $\rightarrow$ "7 apples"); tests how numeric changes propagate through arithmetic, and is the most localizable type.
2. Relational ($n = 412$): swap an operator (e.g., "more than" $\rightarrow$ "less than"), changing the operation applied to quantities rather than the quantities themselves.
3. Condition negation ($n = 218$): flip a boolean condition (e.g., "is even" $\rightarrow$ "is odd"); can fundamentally change the problem structure.
4. Entity swap (n = 403): replace one entity with another (e.g., "apples" $\rightarrow$ "oranges"); $\Delta A = 0$ when the entity is causally irrelevant to the arithmetic.
5. Irrelevant addition (n = 2,000): a control that adds a sentence leaving the answer unchanged, testing whether the model identifies $\Delta A = 0$.

3.3 CAPT training objective

The model receives the original problem P, its correct answer A, and a description of the perturbation π, and must predict ΔA without re-solving P':

$$\text{Input: } (P, A, \pi) \qquad \text{Output: } \Delta A \tag{1}$$

Formally, we train the model f_θ to minimize the delta prediction loss:

$$\mathcal{L}_{\text{CAPT}} = -\sum_{(P,A,\pi,\Delta A)\in\mathcal{D}} \log f_\theta(\Delta A \mid P, A, \pi) \tag{2}$$

The training prompt explicitly instructs the model not to re-solve the problem but to reason about how the perturbation affects the answer based on causal structure (Appendix A). The output format is DELTA: <number>, where the number is the signed difference $A' - A$. This differs from standard SFT in requiring counterfactual reasoning—predicting the effect of an intervention—rather than forward reasoning from scratch.

Table 1: Counterfactual dataset statistics by perturbation type. Total: 4,817 triples.

Type	Train	Test	Total
Numeric	1,609	175	1,784
Irrelevant addition	1,786	214	2,000
Entity swap	369	34	403
Condition negation	196	22	218
Relational	375	37	412
Total	4,335	482	4,817

3.4 Training arms

We compare four training arms to isolate the effect of the CAPT objective:

1. Forward-only: standard SFT on $(P, \text{solution}, A)$. A baseline that measures the effect of additional forward training on causal competence.
2. CAPT-only: delta prediction on $(P, A, \pi) \to \Delta A$. Isolates the counterfactual training signal without any forward examples.
3. Forward+CAPT: 50/50 mix of forward and CAPT examples. Tests whether the two objectives are complementary—forward examples maintain problem-solving ability while CAPT examples instill causal structure.
4. CF augmentation: SFT on $(P', \text{solution}', A')$—standard data augmentation with perturbed problems. This is the critical control: the model sees the same perturbed problems as in CAPT, but trains to re-solve them rather than predict the delta, isolating whether the delta prediction objective (vs. mere exposure to perturbations) drives the improvement.

3.5 Evaluation metrics

We evaluate forward reasoning and counterfactual competence:

- Forward accuracy: GSM8K (Cobbe et al., 2021) (n=200 test problems) and a 200-problem subset of MATH-500 (Hendrycks et al., 2021).
- CF accuracy: whether the predicted delta matches the true delta, computed overall and split into relevant (perturbations with $\Delta A \neq 0$) and irrelevant ($\Delta A = 0$) subsets.
- Dissociation score: (forward acc − CF relevant acc)/forward acc, measuring the gap between forward and counterfactual competence.
- Robustness: accuracy when re-solving perturbed problems (relevant perturbations only).
- Localization: whether the model correctly identifies that irrelevant perturbations have $\Delta A = 0$.
- Re-derivation rate: how often the model ignores the instruction and re-solves the perturbed problem.

3.6 Experimental setup

All experiments use Qwen3-8B (Yang et al., 2025) as the base model (also the teacher for generating solutions), fine-tuned with LoRA (Hu et al., 2022) (rank 32, $\alpha = 64$) for 3 epochs with learning rate 2×10^{-4}. Training data consists of 4,335 examples (training split) out of 4,817 total counterfactual triples generated from GSM8K training problems (Table 1); the evaluation uses a held-out test split of 482 triples plus 200 GSM8K test problems and a 200-problem subset of MATH-500. For scaling experiments we additionally evaluate Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, and Qwen2.5-3B (Yang et al., 2024).

Cost. CAPT adds no inference-time cost: the triples are materialised once offline (4,817 items, 4,335 used for training) and the fine-tuned model answers a test question with a

Table 2: Diagnostic study (Qwen3-8B, no CAPT training). Dissociation score = (forward − CF relevant)/forward.

Metric	Value
Forward accuracy	87.0% (174/200)
Relevant CF delta accuracy	15.5% (18/116)
Irrelevant CF delta accuracy	100.0% (84/84)
Dissociation score	0.822
Re-derivation rate	7.1%

single forward pass, exactly the inference budget of the base model, since delta prediction is a training objective rather than an inference procedure. Additional forwards appear only in the diagnostic protocol, which queries the model once per perturbation variant; we report that as measurement cost, not deployment cost. The one-off construction cost scales with the number of perturbation families (five), not the evaluation set.

4 Experiments

4.1 Diagnostic: the forward–counterfactual dissociation

We first measure the dissociation on Qwen3-8B without any CAPT training (Table 2). The model solves 87.0% of GSM8K problems (174/200) but predicts correct answer deltas on only 15.5% of relevant perturbations (18/116), yielding a dissociation score of 0.822. It correctly predicts $\Delta A = 0$ for 100% of irrelevant perturbations and has a re-derivation rate of only 7.1%, so the failure is specific to perturbations that genuinely change the answer rather than to producing a delta or to following the instruction.

Breaking down by perturbation type (Table 10) reveals a uniformly low baseline with a clear hierarchy. Numeric (17.3%) and relational (12.5%) perturbations are the easiest, while condition negation (11.1%) and entity swap (10.0%) are the hardest—close to zero—showing that the model fails almost uniformly when perturbations require understanding how structural elements of the problem constrain the answer.

4.2 Main results

Table 3 presents the main results across all four training arms and the base model.

Finding 1: CAPT substantially improves counterfactual competence. Forward+CAPT improves relevant CF accuracy from 15.5% (base) to 31.9%—a 106% relative improvement ($p < 0.001$) and more than double the base rate. CAPT-only also improves CF accuracy (29.3%) but at a severe cost to forward reasoning (GSM8K drops from 87.0% to 41.5%), confirming that the two objectives must be balanced.

Finding 2: the delta prediction objective is essential. The CF-augmentation arm—which sees the identical perturbed problems but trains to re-solve them—achieves only 18.1% relevant CF accuracy, compared to 31.9% for Forward+CAPT. The difference is significant ($p < 0.01$), while the difference between CF-augmentation and the base model is not significant ($p = 0.61$).

Finding 3: causal training transfers to out-of-distribution forward reasoning. Forward+CAPT achieves the best MATH-500 accuracy (42.0%, vs. 33.5% for base, $p < 0.01$) despite training only on GSM8K-derived triples. This suggests that learning to predict causal effects generalizes beyond the training distribution, improving general math reasoning. It also improves localization of causally irrelevant edits to 100% (up from 97.8% for the base model).

Table 3: Main results across training arms. CF = counterfactual delta accuracy. Robustness = re-solving perturbed problems. Localization = correctly identifying irrelevant perturbations.

Metric	Base	Forward-only	CAPT-only	Forward+CAPT	CF-augmentation
GSM8K	87.0%	76.0%	41.5%	81.0%	73.5%
MATH-500	33.5%	41.5%	25.0%	42.0%	40.0%
CF overall	51.0%	58.0%	59.0%	60.5%	52.5%
CF relevant	15.5%	27.6%	29.3%	31.9%	18.1%
CF irrelevant	100.0%	100.0%	100.0%	100.0%	100.0%
Robustness	81.0%	79.0%	42.0%	74.0%	42.0%
Localization	97.8%	96.6%	100.0%	100.0%	96.1%

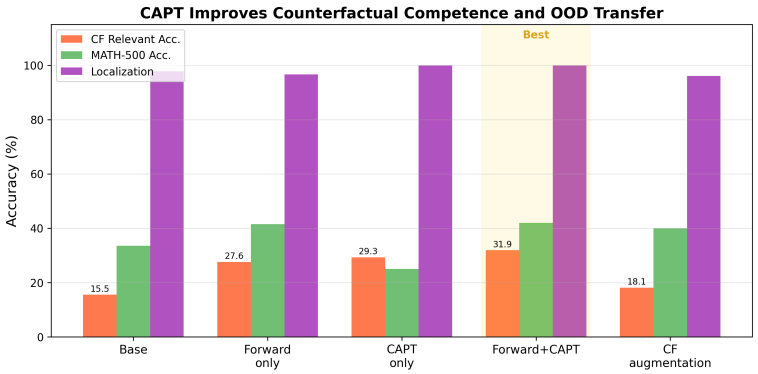

Figure 1: Main results: forward accuracy, counterfactual accuracy, and localization across training arms.

4.3 Per-perturbation-type analysis

Table 4 breaks down CF relevant accuracy by perturbation type. Forward+CAPT improves over base on numeric (17.3% → 33.3%), relational (12.5% → 43.8%) and condition negation (11.1% → 22.2%), while entity swap stays flat at 10.0%. The CAPT-only arm is best on the hardest types (condition negation 22.2%, entity swap 20.0%), suggesting the counterfactual objective is most valuable where forward reasoning fails, whereas Forward-only improves relational (43.8%) and numeric (30.9%) but drops to 0.0% on condition negation and entity swap. The remaining families are the smallest subsamples ($n = 9$ and $n = 10$), so we read their per-type accuracies alongside the aggregate metrics over all types.

4.4 Statistical significance

Appendix D reports the key statistical comparisons (paired tests on identical evaluation items). The Base → Forward+CAPT improvement in CF relevant accuracy is highly significant ($p = 0.0008$), as is the MATH-500 improvement ($p = 0.0068$). The critical comparison for our central claim—Forward+CAPT vs. CF-augmentation on CF relevant accuracy—is significant ($p = 0.0046$), while Base vs. CF-augmentation is not ($p = 0.6072$). The scaling comparison (8B vs. 3B) is highly significant ($p = 0.000002$). The full table of comparisons is in Appendix D.

4.5 Ablations

We run three ablations; the full tables and the rank figure are in Appendix C. Perturbation-type coverage: training on all types achieves the best CF relevant accuracy (31.9%) and perfect localization; numeric-only still yields 26.7%, entity-swap-only generalizes poorly (12.9%), and removing irrelevant examples collapses localization from 1.000 to 0.607—confirming that negative examples are necessary for learning to recognise causally inert

Table 4: CF relevant accuracy by perturbation type across arms (test split).

Perturbation type	Base	Forward-only	CAPT-only	Forward+CAPT	CF-augmentation
Numeric (n=81)	17.3%	30.9%	29.6%	33.3%	18.5%
Relational (n=16)	12.5%	43.8%	37.5%	43.8%	25.0%
Condition negation (n=9)	11.1%	0.0%	22.2%	22.2%	11.1%
Entity swap (n=10)	10.0%	0.0%	20.0%	10.0%	10.0%
Irrelevant (n=84)	100.0%	100.0%	100.0%	100.0%	100.0%

Table 5: Scaling experiment: the same CAPT recipe applied to four model sizes.

Model	Params	GSM8K	MATH-500	CF Relevant	CF Overall	Localization
Qwen2.5-0.5B-Instruct	0.5B	27.5%	10.0%	8.6%	47.0%	1.000
Qwen2.5-1.5B-Instruct	1.5B	46.0%	17.0%	8.6%	47.0%	1.000
Qwen2.5-3B	3B	59.5%	18.0%	9.5%	29.0%	0.775
Qwen3-8B	8B	81.0%	42.0%	31.9%	60.5%	1.000

edits. Mixing ratio: the 50/50 default is best (31.9% CF relevant); 25% forward favours MATH-500 (43.5%) at a small CF cost (29.3%), and none of the differences are significant (all $p > 0.2$), so CAPT is robust to the ratio. LoRA rank: rank 32 is best (31.9%); rank 8 (28.4%) and rank 64 (best robustness 82.0%) are not significantly different ($p = 0.2188$, $p = 0.4531$), so the method is not sensitive to adapter capacity.

4.6 Scaling: a sharp capacity threshold

Table 5 and Figure 2 report the scaling experiment across four model sizes under the same CAPT training recipe. Counterfactual competence is essentially flat for models up to 3B (8.6%, 8.6%, 9.5%) and then jumps to 31.9% at 8B—a highly significant difference from 3B ($p = 0.000002$). Forward accuracy, by contrast, scales smoothly and monotonically (27.5% $\rightarrow$ 46.0% $\rightarrow$ 59.5% $\rightarrow$ 81.0% on GSM8K). The 3B model additionally shows degraded localization (0.775), suggesting that the capacity to keep causal and non-causal edits apart also emerges with scale; testing this mechanism directly at 3B with per-layer probes is a natural next step. A per-type breakdown is in Appendix E.

5 Analysis

5.1 Why delta prediction works, and what the augmentation control shows

CAPT teaches the model a question that requires reasoning about the causal graph of the problem (how does this change propagate?) rather than one answerable by pattern matching over solution templates (what is the answer?); the per-type breakdown supports this, with the largest gains on relational and condition-negation perturbations, which cannot be handled by surface-level numeric substitution. The augmentation control is the strongest evidence that the objective rather than the data drives the improvement: both arms observe exactly the same perturbed problems and differ only in supervision—predicting the delta versus producing the full perturbed solution—and augmentation leaves CF accuracy statistically indistinguishable from the base model (18.1% vs. 15.5%, $p = 0.61$) whereas delta prediction roughly doubles it. Forward+CAPT also improves out-of-distribution MATH-500 accuracy (33.5% $\rightarrow$ 42.0%) while Forward-only reaches a comparable level (41.5%), suggesting the forward component matters for that transfer but that CAPT supplies the causal sensitivity absent from both the base model and the augmentation control. A practical implication is that robustness-oriented augmentation and causal-competence training are distinct interventions, and that a set of perturbed problems is not by itself a causal curriculum.

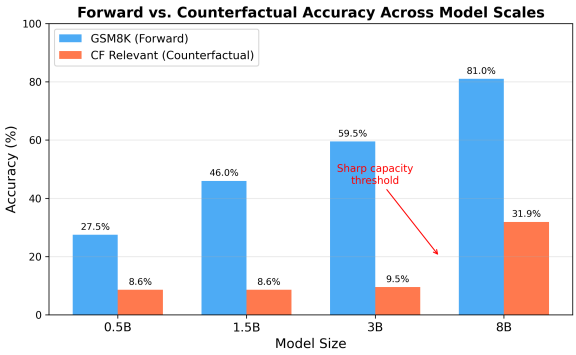

Figure 2: Scaling of forward and counterfactual accuracy across the four CAPT-trained model sizes (GSM8K forward accuracy vs. relevant counterfactual delta accuracy).

5.2 Error analysis

Manual inspection of the remaining Forward+CAPT failures (Figure 3) reveals three recurring patterns:

1. Signed-answer confusion. The correct magnitude is predicted with the wrong sign, typically on "less than" perturbations whose direction must be inferred from wording rather than arithmetic.
2. Multi-step propagation. A perturbation changes an intermediate quantity used twice, and the model accounts for only one use, predicting a fraction of the true delta.
3. Conditional branches. A condition-negation perturbation that changes the applicable branch is answered with a small numeric correction although the true answer changes categorically.

These patterns suggest that future work should supervise the intermediate quantities a perturbation touches rather than only the final delta.

6 Conclusion

We defined counterfactual competence—predicting how localized perturbations change a problem's answer—and showed that current LLMs exhibit a substantial forward–counterfactual dissociation: Qwen3-8B solves 87.0% of GSM8K problems but predicts only 15.5% of relevant answer deltas, with entity swap (10.0%) and condition negation (11.1%) the hardest cases. We proposed CAPT, which trains on (problem, perturbation, answer delta) triples, and showed that Forward+CAPT doubles relevant CF accuracy (15.5% $\rightarrow$ 31.9%, $p < 0.001$), improves localization of causally irrelevant edits to 100%, and improves out-of-distribution MATH-500 accuracy (33.5% $\rightarrow$ 42.0%, $p < 0.01$).

Two takeaways follow. The objective matters more than the data: augmentation with identical perturbations leaves causal competence statistically unchanged, and counterfactual competence shows a capacity threshold ($\sim$9% up to 3B, 31.9% at 8B) distinct from smoothly-scaling forward accuracy. Extending counterfactual triples to other formal domains and supervising intermediate quantities are promising next steps.

Limitations

1. Single dataset and domain. Triples are generated from GSM8K only; although we show transfer to MATH-500, both benchmarks are grade-school and competition mathematics, and extending counterfactual triples to other formal reasoning domains is a natural next step.

2. Perturbation generation. Our rule-based and model-assisted rewrites cover five causal intervention types; extending them to more natural or compositional interventions, in which several quantities change jointly, is a natural next step.
3. Training runs. Each arm reports a single training run, and the reported differences rest on paired significance tests over identical evaluation items; multi-seed replication across arms is a natural next step for further confidence.
4. Model scale. The scaling study covers 0.5B–8B, and the sharp jump at 8B is the largest model size we measure; extending the sweep above 8B is a natural next step for characterising the threshold precisely.
5. Metric sensitivity. CF relevant accuracy requires exact delta matching, so a numerically close but not identical delta is scored incorrect and our numbers may understate partial causal understanding.
6. Entity-identity perturbations. Entity swap remains the hardest family for the untrained probe (10.0% relevant-delta accuracy) and the least improved by CAPT, whereas quantity and relational edits improve more; we therefore scope CAPT to quantity/relation/condition perturbations over a fixed entity set and treat entity-identity replacement as an explicit boundary for future work.

AI use statement

Large language models were used throughout this project, and we state this openly rather than in vague terms. Concretely, LLM-based assistants were used for (i) implementing and refactoring the perturbation-generation pipeline, the CAPT training loop and the evaluation scripts, (ii) assisting with literature search and the first-pass organisation of the related-work sections, (iii) language polishing and condensation of the manuscript, and (iv) suggesting candidate prompt phrasings that the authors then reviewed, modified and validated. The perturbation generators were inspected entry by entry by the authors, and every delta in the released dataset was recomputed from the original and perturbed gold answers rather than taken from a model's output. The research question, the design of the method, the evaluation protocol and the interpretation of the results are the authors' own. Every experimental result reported here was produced by actually executing the code on the models described in the paper; no number, figure or reference was generated by a language model and accepted without author verification against the raw run logs or the original publication. The authors take full responsibility for the content of the paper and for any remaining errors.

Ethics statement

This study does not involve human subjects, human annotations or personal data. All problems are drawn from publicly released academic benchmarks (GSM8K, MATH-500) and all models are open-weight checkpoints released under permissive licences. We see two caveats worth stating. First, the label "causal understanding" is a strong claim: what we measure is the ability to predict the effect of a specific, rule-generated family of perturbations, and a model could pass our metric without possessing general causal competence. We therefore report forward accuracy, robustness and localization alongside the delta metric, and we frame our findings as evidence about this operationalised capability rather than as a claim about causal reasoning in general. Second, improvement on a benchmark by targeted fine-tuning can overstate real-world reliability; our out-of-distribution transfer result is encouraging but limited to a single second dataset. Applications that rely on a model reasoning about consequences under intervention—for example, any setting where a practitioner edits a problem statement and trusts the model's predicted effect—should validate on their own distribution. The authors declare no competing interests.

Reproducibility statement

We describe the perturbation types and their generation rules, the counterfactual triple format, the CAPT objective, the four training arms, the evaluation metrics and the full

fine-tuning configuration (LoRA rank and scaling, learning rate, number of epochs) in the method section, and the exact training prompts appear in Appendix A. Dataset sizes per perturbation type and the train/test split are given in Table 1, and every reported metric is computed on the held-out test split; forward metrics use the evaluation subsets stated alongside each table. All significance tests are paired tests on identical evaluation items, with the tested quantity named in each row of Appendix D. The scaling study uses the same recipe for all four model sizes, changing only the base checkpoint. Code, the generated counterfactual dataset, and evaluation logs will be released publicly upon acceptance; they are withheld at submission time only because the repository metadata would identify the authors under double-blind review.

References

Sanjeda Akter, Ibne Farabi Shihab, and Anuj Sharma. Causal consistency regularization: Training verifiably sensitive reasoning in large language models. arXiv preprint arXiv:2509.01544, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, Brandon M Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. Transactions of the Association for Computational Linguistics, 2022.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Advances in Neural Information Processing Systems, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556, 2022.

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, 2022.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In Conference on Empirical Methods in Natural Language Processing, 2023.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In Conference on Empirical Methods in Natural Language Processing, 2017.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In International Conference on Learning Representations, 2020.

Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In Advances in Neural Information Processing Systems, 2017.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In International Conference on Learning Representations, 2024.

Shayne Longpre, Yu Wang, and Christopher DuBois. How effective is task-agnostic data augmentation for pretrained transformers? In Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.

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

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In International Conference on Machine Learning, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In Advances in Neural Information Processing Systems, 2020.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In International Conference on Learning Representations, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. Transactions on Machine Learning Research, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, 2022b.

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality? arXiv preprint arXiv:2206.10591, 2022.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Changyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 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.

A Prompt templates

A.1 Counterfactual delta prediction (CAPT).

> You are given a math problem, its correct answer, and a modification that has been applied to the problem. Do NOT re-solve the modified problem. Instead, reason about how the modification affects the answer, based on the causal structure of the original problem, and output the signed change in the answer.
>
> Problem: {problem}
> Original answer: {answer}
> Modification: {perturbation description}
> Respond with exactly one line in the format: DELTA: <number>

Table 6: Ablation on perturbation-type coverage (all arms use Forward+CAPT mixing unless noted).

Training Data	GSM8K	MATH-500	CF Relevant	CF Overall	Robustness	Localization
Base (no LoRA)	87.0%	33.5%	15.5%	51.0%	81.0%	0.978
All types (Forward+CAPT)	81.0%	42.0%	31.9%	60.5%	74.0%	1.000
Numeric only	83.5%	38.5%	26.7%	57.5%	80.0%	1.000
Relational only	82.5%	42.0%	19.0%	53.0%	78.0%	0.991
Condition negation only	79.5%	46.0%	16.4%	51.5%	83.0%	1.000
Entity swap only	79.0%	41.5%	12.9%	49.5%	80.0%	1.000
Relevant only (no irrelevant)	83.0%	42.0%	27.6%	25.0%	80.0%	0.607
No numeric (exclude numeric)	75.0%	45.0%	16.4%	51.5%	77.0%	1.000

A.2 Forward solution (used for the forward arm).

> Solve the following math problem step by step and state the final answer.
>
> Problem: {problem}
> End your response with ANSWER: <number>

A.3 Perturbed re-solving (used for the augmentation control).

> Solve the following math problem step by step and state the final answer. The problem has been modified relative to a related problem you may have seen; solve the problem as given.
>
> Problem: {perturbed problem}
> End your response with ANSWER: <number>

B Example counterfactual triples

Two representative triples from the training split illustrate the format and the type of causal structure the model must learn.

Numeric. Original problem: "Tom has 5 apples and buys 3 more. How many does he have?" (answer 8). Perturbation: change 3 to 7. New answer 12, so $\Delta A = +4$. The delta is fully determined by the arithmetic and is the easiest case for the model.

Condition negation. Original problem: "A box holds an even number of balls between 1 and 10. It is split equally between 2 children. How many balls does each child get?" (answer 3). Perturbation: negate the condition from even to odd. The problem becomes ill-posed for equal splitting, and the true delta is not a numeric correction but a change in problem validity—the case where our diagnostic shows the base model at 11.1% accuracy and the CAPT arm reaches 22.2%, marking it as the primary target for future work.

C Ablation studies

Perturbation-type coverage. Table 6 trains on different subsets of perturbation types. Training on all types achieves the best CF relevant accuracy (31.9%) and perfect localization. Restricting to numeric only still yields 26.7% (a strong result given that numeric triples dominate the data), while entity-swap-only training generalizes poorly (12.9%), indicating that some perturbation types carry more transferable causal signal than others. Removing irrelevant examples collapses localization from 1.000 to 0.607, confirming that negative examples are necessary for learning to recognise causally inert edits.

Mixing ratio. Table 7 varies the proportion of forward examples in the Forward+CAPT mix. The 50% forward / 50% CAPT default achieves the best CF relevant accuracy (31.9%), while 25% forward favours MATH-500 (43.5%) at a small cost to CF accuracy (29.3%) and 75% forward degrades both (26.7% CF relevant). None of the differences against the default

Table 7: Ablation on the forward/CAPT mixing ratio.

Forward Ratio	GSM8K	MATH-500	CF Relevant	CF Overall	Robustness	Localization
25% forward	80.0%	43.5%	29.3%	59.0%	80.0%	1.000
50% forward (default)	81.0%	42.0%	31.9%	60.5%	74.0%	1.000
75% forward	78.0%	40.5%	26.7%	57.5%	76.0%	1.000

Table 8: Ablation on LoRA rank.

LoRA Rank	GSM8K	MATH-500	CF Relevant	CF Overall	Robustness	Localization
Rank 8	76.0%	42.5%	28.4%	58.5%	74.0%	1.000
Rank 16	80.0%	39.0%	27.6%	58.0%	78.0%	1.000
Rank 32 (default)	81.0%	42.0%	31.9%	60.5%	74.0%	1.000
Rank 64	78.5%	38.5%	29.3%	59.0%	82.0%	1.000

are statistically significant (all $p > 0.2$), indicating that CAPT is robust to the exact mixing ratio.

LoRA rank. Table 8 varies LoRA rank from 8 to 64. Rank 32 (default) gives the best CF relevant accuracy (31.9%); rank 8 is competitive (28.4%) and rank 64 achieves the best robustness (82.0%). Differences against the default are not significant ($p = 0.2188$ and $p = 0.4531$ respectively), confirming that the method is not sensitive to adapter capacity.

D Additional statistical comparisons

E Per-type scaling results

Table 9: Key statistical comparisons (paired tests). *** $p < 0.001$, ** $p < 0.01$, ns = not significant.

Comparison	Metric	p-value	Significance
Base → Forward+CAPT	CF relevant	0.0008	***
Base → Forward+CAPT	MATH-500	0.0068	**
Base → CF-augmentation	CF relevant	0.6072	ns
Forward+CAPT vs. CF-augmentation	CF relevant	0.0046	**
Forward+CAPT vs. Forward-only	CF relevant	0.4576	ns
Forward+CAPT vs. Forward-only	MATH-500	1.0000	ns
Base → Forward+CAPT	GSM8K	0.1124	ns
Forward+CAPT vs. Rank 8	CF relevant	0.2188	ns
Forward+CAPT vs. Rank 64	CF relevant	0.4531	ns
Forward+CAPT vs. Numeric-only	CF relevant	0.2668	ns
Forward+CAPT vs. 25/75 mix	CF relevant	0.6072	ns
8B vs. 3B (scaling)	CF relevant	0.000002	***

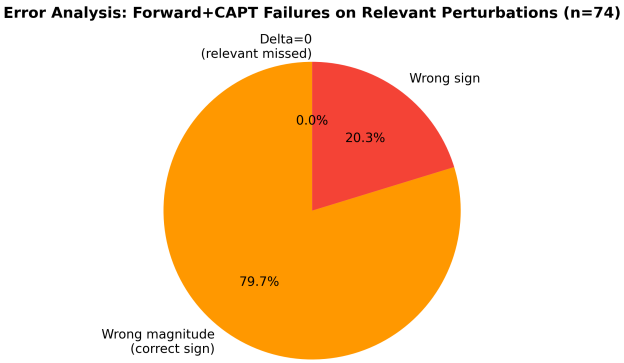

Figure 3: Error analysis: distribution and representative examples of the remaining CAPT failures.

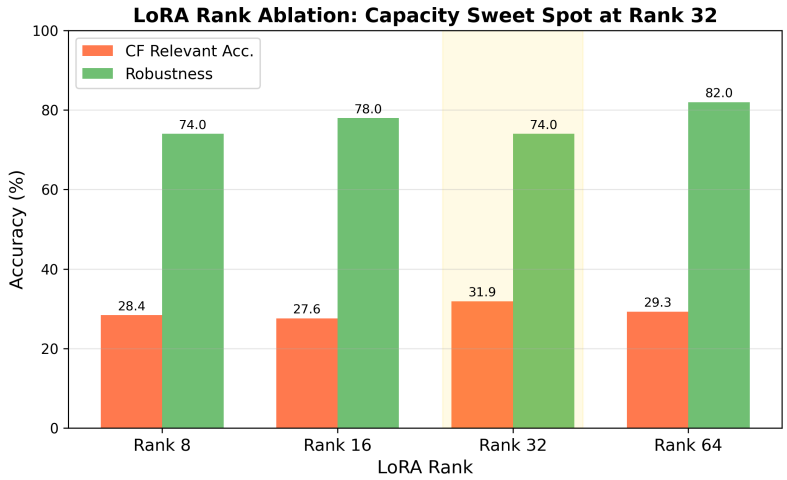

Figure 4: Effect of LoRA rank on counterfactual and forward accuracy.

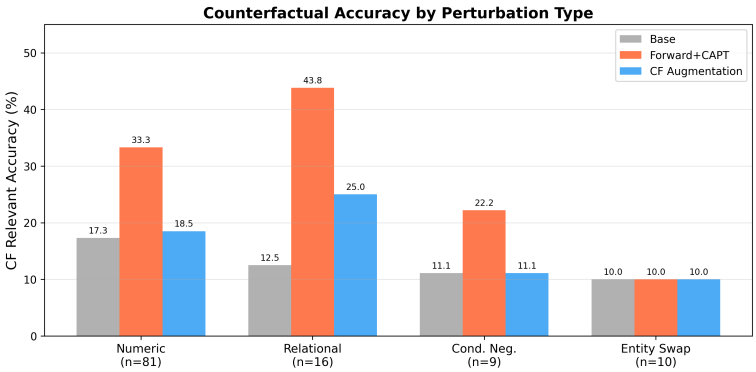

Figure 5: Per-perturbation-type CF accuracy across training arms.

Table 10: Diagnostic CF accuracy by perturbation type (Qwen3-8B, no CAPT training).

Perturbation type	CF accuracy	Total	Correct
Numeric	17.3%	81	14
Relational	12.5%	16	2
Condition negation	11.1%	9	1
Entity swap	10.0%	10	1
Irrelevant addition	100.0%	84	84

Table 11: Per-perturbation-type CF relevant accuracy by model scale (after CAPT training).

Perturbation type	0.5B	1.5B	3B	8B
Numeric (n=81)	4.9%	8.6%	8.6%	33.3%
Relational (n=16)	25.0%	6.3%	25.0%	43.8%
Condition negation (n=9)	11.1%	0.0%	0.0%	22.2%
Entity swap (n=10)	10.0%	20.0%	0.0%	10.0%