HELPFUL WHEN RIGHT, CONTROLLING WHEN WRONG: CAUSAL DEPENDENCE ON EXTERNAL INTERMEDIATE STATES IN SMALL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Step-level feedback is usually judged by whether it improves a final answer, but this hides whether a model recomputes or simply inherits a supplied intermediate state. We isolate these mechanisms with controlled interventions on arithmetic and symbolic reasoning traces. Each trace has a unique first error and deterministic continuation, so we can change only the external state while measuring final repair, local adoption, and exact downstream propagation. Across Qwen3 and Gemma 3 models from 270M to 4B parameters, correct states sometimes produce large gains. On symbolic logic, they improve repair over localization alone by 58.0 percentage points for Qwen3-1.7B and 18.7 points for Gemma 3 1B. Capable models also propagate false states. Qwen3-4B adopts matched wrong feedback on 51.3% of arithmetic items and 56.7% of logic items. A pre-specified seven-level arithmetic intervention reveals a sharp peak at the true state, alongside persistent family-specific susceptibility to counterfactual states. The results support three descriptive regimes: capacity limited, feedback enabled, and autonomous but feedback susceptible. External step-level feedback can therefore be both helpful and behaviorally controlling.

1 INTRODUCTION

Suppose a model is revising a four-step solution. A verifier says, "Step 2 is wrong," and the model corrects the answer. Now the verifier says, "The correct state at Step 2 is 33," and the model improves again. The second message looks more helpful, but the improvement is ambiguous. Did the model use 33 as evidence and verify it, or did 33 simply take control of the remaining computation? Correct feedback cannot distinguish these possibilities. A matched message containing 32 can.

This question sits between two established findings. Intermediate reasoning and process-level supervision often improve multistep performance (Nye et al., 2021; Wei et al., 2022; Lightman et al., 2023; Uesato et al., 2022), and feedback-based revision can improve answers without retraining (Madaan et al., 2023; Shinn et al., 2023; Gou et al., 2023). Yet self-correction is inconsistent (Huang et al., 2023; Liu et al., 2024; Weng et al., 2023), models sometimes favor a user's claim over truth (Sharma et al., 2023), and their answers do not always depend faithfully on their stated reasoning (Lanham et al., 2023). What is missing is a local causal test. When feedback changes an answer, we need to know whether its content entered the computation and, if so, how far it propagated.

Natural reasoning benchmarks make that test difficult. Solutions can follow several valid paths, the first error is often debatable, and a changed step may not have a unique continuation. We therefore study compact transition systems in which every intermediate state is known. Each candidate contains one identifiable first error, and every possible replacement state determines a unique downstream trace. On the same item, we compare error localization, a correct replacement state, and a matched false state. We score not only the final answer, but also whether the model adopts the supplied state and whether it carries that state consistently through the remaining steps.

Across arithmetic and symbolic logic, two model families, and six sizes from 270M to 4B parameters, a coherent picture emerges. Small models are often unable to execute the repair protocol. At intermediate capability, a correct state can transform performance. More capable models can frequently repair from localization alone, but some remain highly vulnerable to false states. A final seven-level

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.

intervention shows that both 4B models respond most strongly to the exact true value, while differing sharply in their willingness to follow false alternatives. These results turn a familiar claim, that process feedback helps, into a more precise one: the same channel that scaffolds reasoning can also control it.

2 Related Work

Process supervision assigns credit to intermediate reasoning rather than only to final outcomes, and it can outperform outcome supervision in mathematics (Lightman et al., 2023; Uesato et al., 2022). Critique-based systems pursue a related idea at inference time. Self-Refine iterates on model-generated feedback (Madaan et al., 2023), Reflexion carries verbal feedback into later decisions (Shinn et al., 2023), and CRITIC checks outputs with external tools (Gou et al., 2023). Models can also write critiques that help human evaluators notice errors (Saunders et al., 2022). Together, this work establishes that intermediate feedback can be useful.

Usefulness, however, is not the same as reliability. Studies of intrinsic self-correction reach different conclusions depending on prompts, temperature, and evaluation design (Huang et al., 2023; Liu et al., 2024; Weng et al., 2023). Sycophancy provides a complementary warning: a model may align its answer with a user's stated belief rather than with the evidence (Sharma et al., 2023). Chain-of-thought interventions similarly show that reliance on an explicit rationale varies across tasks and model sizes (Lanham et al., 2023). Work on AI feedback and scalable oversight raises the same issue at the system level, where one model's judgment becomes another model's supervision (Bai et al., 2022; Lee et al., 2023; Kenton et al., 2024).

Recent work moves closer to the failure mode studied here. ProgCo treats incorrect feedback as a source of misleading refinement and uses program-based verification to make self-correction more robust (Song et al., 2025). Kim and Khashabi show that a user rebuttal with detailed but invalid reasoning can persuade a model to abandon a correct judgment (Kim & Khashabi, 2025). Zhu et al. find that reasoning failures often begin at a small number of early transitions, after which the trajectory remains locally coherent but globally wrong (Zhu et al., 2026). Together, these studies establish that misleading feedback and early errors matter. They do not isolate whether one externally supplied intermediate state is adopted at a specified step and then propagated through the remaining computation.

Our contribution is this finer causal resolution. We intervene on one known local state while holding the problem, candidate solution, error location, feedback form, and decoding procedure fixed. A matched counterfactual supplies a plausible wrong state in the same form as the correct state. Because every intervention has a unique continuation, we can measure exact adoption at the target step and exact propagation afterward. This separates local computational control from persuasion, generic reconsideration, and coincidental recomputation. Public benchmarks such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) remain valuable measures of broad accuracy, but their diverse solution paths make this dependency difficult to identify. Controlled transition systems sacrifice breadth to make the causal question answerable.

3 Controlled Reasoning Framework

3.1 Problem formulation

Let a reasoning trace be a sequence of intermediate states

$$\tau = (s_0, s_1, \ldots, s_T), \qquad s_t = f_t(s_{t-1}; x),$$

where x is the problem and f_t is the transition applied at Step t. Our controlled tasks make each transition deterministic. Given an erroneous candidate, an external verifier intervenes at its known first error k by supplying a state $\tilde{s}_k$:

$$\mathrm{do}(s_k = \tilde{s}_k), \qquad \tilde{s}_t = f_t(\tilde{s}_{t-1}; x) \quad \text{for } t > k.$$

Let $\hat{\tau}^{(\tilde{s}_k)}$ denote the model's revised trace under this intervention, and let $s_T^\star$ be the gold final state. We distinguish three outcomes:

$$\begin{aligned}
\text{Adopt}_k &= \mathbb{1}\left[\hat{s}_k^{(\tilde{s}_k)} = \tilde{s}_k\right], \\
\text{Propagate}_k &= \mathbb{1}\left[\forall t \geq k,\ \hat{s}_t^{(\tilde{s}_k)} = F_{k:t}(\tilde{s}_k)\right], \\
\text{Repair} &= \mathbb{1}\left[\hat{s}_T^{(\tilde{s}_k)} = s_T^\star\right],
\end{aligned}$$

where $F_{k:t}$ is the deterministic continuation from the intervened state through Step t. Correct feedback can produce all three outcomes and therefore cannot reveal whether the model verified or inherited the supplied state. The matched counterfactual resolves this ambiguity: when $\tilde{s}_k \neq s_k^\star$ and $\text{Propagate}_k = 1$, the model has adopted a false local state and generated its uniquely implied continuation. We treat this event as direct evidence of behavioral causal control over the produced reasoning trace, not as an intervention on internal model representations.

3.2 Why controlled tasks

Our goal is causal identification, not benchmark coverage. Every evaluation item must support four operations that are generally unavailable in public reasoning datasets:

- identify a unique first erroneous step;
- intervene on exactly one local state while holding the rest of the prompt fixed;
- compute the downstream trajectory implied by any supplied state;
- distinguish adoption of the supplied state from independent recovery of the gold answer.

Programmatically generated tasks provide these guarantees and avoid pretraining contamination. The trade-off is reduced ecological breadth, which we address through replication across task form and model family rather than by mixing in uncontrolled benchmark items.

3.3 Arithmetic state chains

Each arithmetic problem starts from an integer and applies two to four additions or subtractions. Operands range from 1 to 9 and starting values from 0 to 20. A candidate trace is generated by corrupting one intermediate state; all later candidate states are then computed from that corrupted value. Consequently, each trace has a unique known first error and a coherent but wrong continuation.

The confirmatory set contains 300 corrupted traces: 100 at each depth 2, 3, and 4. We generated and froze this dataset before confirmatory inference.

3.4 Symbolic rule chains

To test whether the phenomenon depends on arithmetic, we generate multi-hop implication chains over nonce categories. Each item includes a valid chain, a coherent candidate-error branch, a distinct coherent wrong-feedback branch, and two distractor rules. Both the correct and counterfactual supplied categories therefore have executable continuations. A model cannot reject the wrong feedback simply because no downstream rule exists.

The confirmatory symbolic set also contains 300 traces, balanced across depths 2 to 4 and frozen before inference.

3.5 Feedback interventions

For each corrupted trace, the primary experiment evaluates three conditions:

- **Localization only:** the verifier states that Step *k* is wrong and asks the model to correct from that point.
- **Structured correct state:** the verifier identifies the first error, supplies the correct state after Step *k*, and asks the model to recompute later steps.

- **Matched wrong state:** the verifier uses the same authoritative form but supplies a counterfactual local state.

The correct and wrong supplied states each induce a unique continuation. This lets us distinguish four outcomes:

1. **Parse:** the response is a valid structured solution.
2. **Repair:** the final answer equals the gold answer.
3. **Adoption:** the response places the supplied state at the targeted step.
4. **Propagation:** every state from the intervention onward equals the trajectory implied by the supplied state.

Parse failures count as failures for all substantive outcomes. Repair without adoption is possible when a model rejects the feedback and recomputes independently. Adoption without full propagation indicates local copying without consistent execution.

3.6 Signed state intervention

The matched wrong-state experiment establishes that a false state can affect reasoning. To characterize this dependence more finely, we pre-specified a seven-level arithmetic intervention. Let the true state at the targeted step be s_k. We supply

$$\tilde{s}_k = s_k + \delta, \qquad \delta \in \{-4, -2, -1, 0, 1, 2, 4\}.$$

Only the numeral changes across conditions. The item, candidate trace, target step, wording, chat template, and decoding parameters remain fixed. Because all subsequent operations are additions or subtractions, exact propagation shifts the final answer by the same δ. We evaluate all 300 arithmetic items at every offset for Qwen3-4B and Gemma 3 4B, yielding 2,100 responses per model.

4 Experimental Setup

4.1 Models

We evaluate instruction-tuned checkpoints from two families:

- Qwen3: 0.6B, 1.7B, and 4B;
- Gemma 3: 270M, 1B, and 4B.

All evaluations use official Hugging Face checkpoints and chat templates. Model architecture and training details are described in the Qwen3 and Gemma 3 technical reports (Qwen Team, 2025; Gemma Team, 2025). Qwen3's optional thinking mode is disabled to maintain a controlled 256-token response budget. Decoding is greedy with temperature zero. Qwen3 runs use FP16. Gemma 3 runs use FP32 because the evaluation GPU does not support BF16 and direct FP16 conversion produced degenerate all-padding generation in a pre-confirmatory smoke test. No confirmatory item was inspected during this calibration.

All model versions were fixed before confirmatory evaluation.

4.2 Confirmatory sequence

Prompt development and the initial representation ablation were exploratory. We then froze a fresh arithmetic dataset, three primary feedback conditions, decoding parameters, outcomes, and statistical tests before confirmatory inference. The symbolic task was separately pre-specified before its outputs were observed. Gemma 3 cross-family replication was frozen before Gemma saw either confirmatory set. The signed intervention was likewise pre-specified before its outputs were inspected.

4.3 Statistical analysis

Primary comparisons are paired because every condition is evaluated on the same items. We report paired repair-rate differences with percentile 95% confidence intervals from 10,000 paired bootstrap

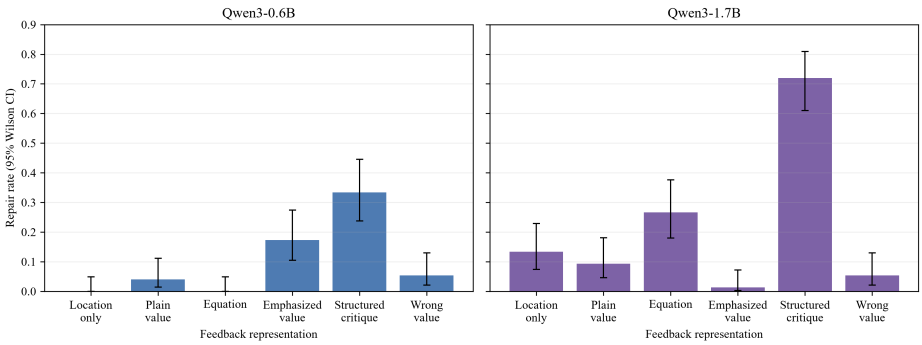


Figure 1: Exploratory feedback-representation ablation on 75 arithmetic items per model. Semantically related presentations of the same local information produce sharply different repair rates. Error bars are 95% Wilson intervals. This result motivated freezing prompt wording before all confirmatory runs.

Table 1: Qwen3 confirmatory repair rates and correct-feedback advantage.

Task	Model	Localization	Correct state	Wrong state	Correct − localization (95% CI)	McNemar p
Arithmetic	0.6B	0.3%	18.0%	3.0%	+17.7 [13.3, 22.0]	3.11e-15
Arithmetic	1.7B	4.3%	67.3%	1.7%	+63.0 [57.3, 68.7]	1.22e-55
Arithmetic	4B	74.3%	60.0%	17.0%	−14.3 [−21.0, −7.7]	5.49e-5
Logic	0.6B	2.0%	5.7%	4.3%	+3.7 [0.7, 6.7]	.0347
Logic	1.7B	14.7%	72.7%	60.0%	+58.0 [52.0, 64.0]	4.83e-44
Logic	4B	79.0%	71.0%	24.0%	−8.0 [−15.0, −1.0]	.0308

samples and exact two-sided McNemar tests. Signed-offset curves use Wilson 95% binomial intervals. For equal-magnitude positive and negative offsets, we use paired McNemar tests to assess directional asymmetry. Results are not filtered on successful parsing.

5 Results

5.1 Feedback wording is an active experimental variable

Figure 1 summarizes the exploratory prompt ablation. Semantically related feedback forms were not interchangeable. Merely naming the error location was often insufficient for small models, whereas an explicit executable state could unlock repair. This observation motivated the confirmatory design and an important methodological decision: all causal comparisons use frozen wording, and the wrong-state control matches the concise state-supplying form rather than introducing additional criticism or explanation.

We treat this prompt sensitivity as a result about interface design, not as permission to select the best prompt after model evaluation. All subsequent claims rely on fresh, frozen data.

5.2 Qwen3 exhibits three feedback regimes

Table 1 summarizes Qwen3 repair. Localization-only performance increases sharply with capacity, from near zero at 0.6B and 1.7B to 74.3% at 4B. Correct-state feedback has its largest effect at 1.7B: a 63.0-point paired gain. At 4B, however, localization alone outperforms structured correct feedback by 14.3 points.

The local metrics reveal why the 4B result should not be read as feedback irrelevance. Qwen3-4B adopts the wrong supplied state on 51.3% of arithmetic items and 56.7% of logic items; exact wrong-branch propagation reaches 44.3% on arithmetic. The model can often correct from localization alone, yet authoritative state feedback still exerts substantial control.

Table 2: Gemma 3 confirmatory repair rates and correct-feedback advantage.

Task	Model	Localization	Correct state	Wrong state	Correct − localization (95% CI)	McNemar p
Arithmetic	270M	0.0%	0.3%	0.0%	+0.3 [0.0, 1.0]	1.0
Arithmetic	1B	0.3%	3.0%	0.0%	+2.7 [0.7, 4.7]	.0215
Arithmetic	4B	53.0%	59.0%	34.0%	+6.0 [−2.0, 14.3]	.168
Logic	270M	0.0%	6.7%	4.3%	+6.7 [4.0, 9.7]	1.91e-6
Logic	1B	0.7%	19.3%	8.3%	+18.7 [14.3, 23.3]	4.09e-16
Logic	4B	58.7%	72.3%	33.7%	+13.7 [6.7, 20.7]	2.15e-4

These results motivate three descriptive regimes rather than a monotonic "more parameters use feedback better" claim:

- **capacity-limited:** the model struggles to execute the output and repair protocol;
- **feedback-enabled:** localization alone rarely succeeds, but a correct executable state unlocks repair;
- **autonomous but feedback-susceptible:** the model can repair independently, while plausible supplied states can redirect its trajectory.

5.3 The pattern transfers beyond arithmetic

The symbolic task removes numerical computation while preserving intervention structure. Qwen3 reproduces the same qualitative transition: a small effect at 0.6B, a 58.0-point correct-feedback advantage at 1.7B, and a negative advantage at 4B because localization-only repair is already high. Wrong-feedback adoption simultaneously grows from 0.3% to 1.3% to 56.7%.

This cross-task replication rules out an explanation based only on copying arithmetic numerals or failing to perform addition. What generalizes is the use of an externally supplied intermediate state in a multi-hop transition system.

5.4 The effect is not specific to Qwen3

Gemma 3 provides a more conservative but still clear replication. Table 2 shows that correct-state feedback significantly improves symbolic repair at 1B and 4B. Gemma 3 4B reaches 72.3% repair with the correct symbolic state, compared with 58.7% under localization and 33.7% with the wrong state. It adopts the wrong symbolic state on 35.0% of items.

Figure 2 places these results in a common cross-task and cross-family view, separating correct-state benefit from wrong-state susceptibility.

The families differ in where the apparent transition occurs. Qwen3 peaks in feedback advantage at its intermediate size and becomes localization-dominant at 4B. Gemma 3 improves more gradually and retains a positive logic advantage at 4B. This is evidence against a universal parameter-count threshold. It is consistent with a task- and family-dependent transition governed by both autonomous competence and trust in external state.

The smallest Gemma requires special care. Its arithmetic parse rate is only 10.3% under correct-state feedback and 6.7% under wrong-state feedback, so near-zero repair is partly an instruction-following floor. We retain these results, as pre-specified, but do not interpret them as a clean estimate of reasoning repair. Figure 3 summarizes the paired correct-state effects and their confidence intervals across all model and task combinations.

5.5 A correctness-gated intervention curve

The seven-level intervention gives the most direct evidence that the supplied state is causally active. Figure 4 shows a sharp peak at $\delta = 0$ for both 4B models.

For Qwen3-4B, supplied-state adoption is 72.7% at the true state. It remains high for false states, ranging from 40.0% to 60.3% across nonzero offsets. Exact propagation is 60.7% at zero and ranges from 32.3% to 54.0% elsewhere. Gemma 3 4B is more selective: adoption is 70.3% at zero but ranges from 10.0% to 18.0% for false values; propagation falls from 45.7% to a range of 4.3% to 11.7%. Table 3 reports the exact values underlying the curves in Figure 4.

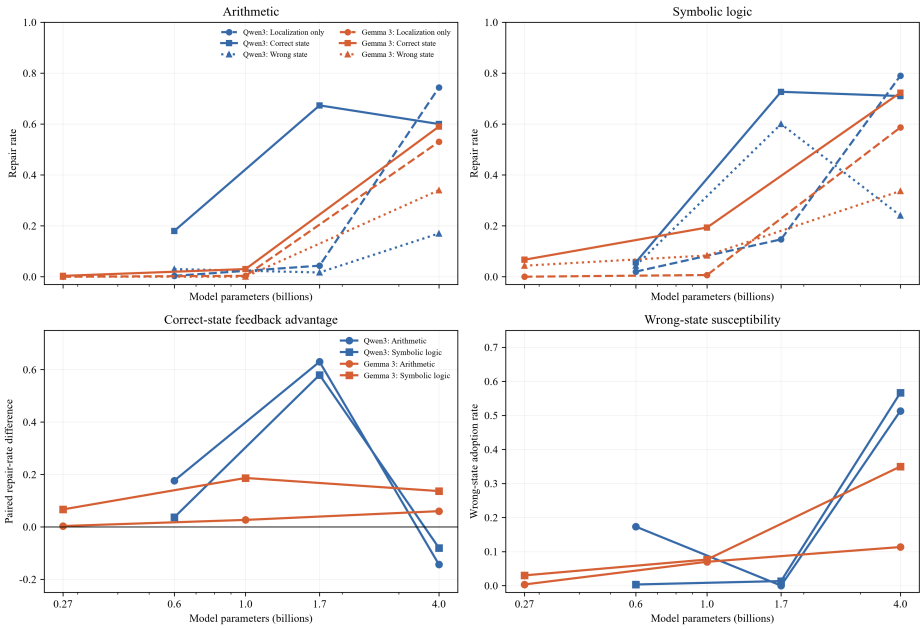


Figure 2: Cross-task and cross-family feedback phase diagram. The top row shows repair under localization, correct-state feedback, and matched wrong-state feedback. The bottom row separates the correct-state repair advantage from susceptibility to wrong-state adoption.

Table 3: Arithmetic signed-state intervention. Values are percentages.

Model	Outcome	−4	−2	−1	0	+1	+2	+4
Qwen3-4B	Adoption	40.0	52.3	46.7	**72.7**	60.3	58.0	47.7
Qwen3-4B	Propagation	32.3	46.7	43.3	**60.7**	54.0	49.0	39.3
Qwen3-4B	Repair	24.3	21.7	21.3	**61.7**	13.7	13.3	18.0
Gemma 3 4B	Adoption	10.0	17.0	16.7	**70.3**	16.3	18.0	13.3
Gemma 3 4B	Propagation	4.3	11.7	9.3	**45.7**	8.7	11.7	7.7
Gemma 3 4B	Repair	37.0	30.3	31.3	**50.0**	32.3	29.7	31.0

The curve is not a smooth monotonic function of absolute numerical distance. Instead, it has a discontinuous correctness advantage: exact truth is treated differently from every tested false value. Beyond that gate, Qwen3 remains broadly compliant with incorrect states, whereas Gemma usually rejects them. Qwen3 also shows signed asymmetry. At magnitude 1, positive offsets are adopted more often than negative offsets (60.3% vs. 46.7%; paired McNemar p=.00043), while repair is lower for the positive offset (13.7% vs. 21.3%; p=.014). Gemma shows no ±1 or ±2 adoption asymmetry.

Taken together, these curves rule out a generic reconsideration effect. An instruction to revise might improve performance whenever a verifier is present, but it does not explain why the supplied numeral reappears at the targeted step, determines the later trace, and produces a sharp peak only when it is true. The model is responding to the content of the external state, not merely to the request to try again.

6 Discussion

Correct feedback helps because the model is willing to condition on it. The wrong-state experiments show the cost of that willingness. At intermediate capability, an external state can bridge a genuine

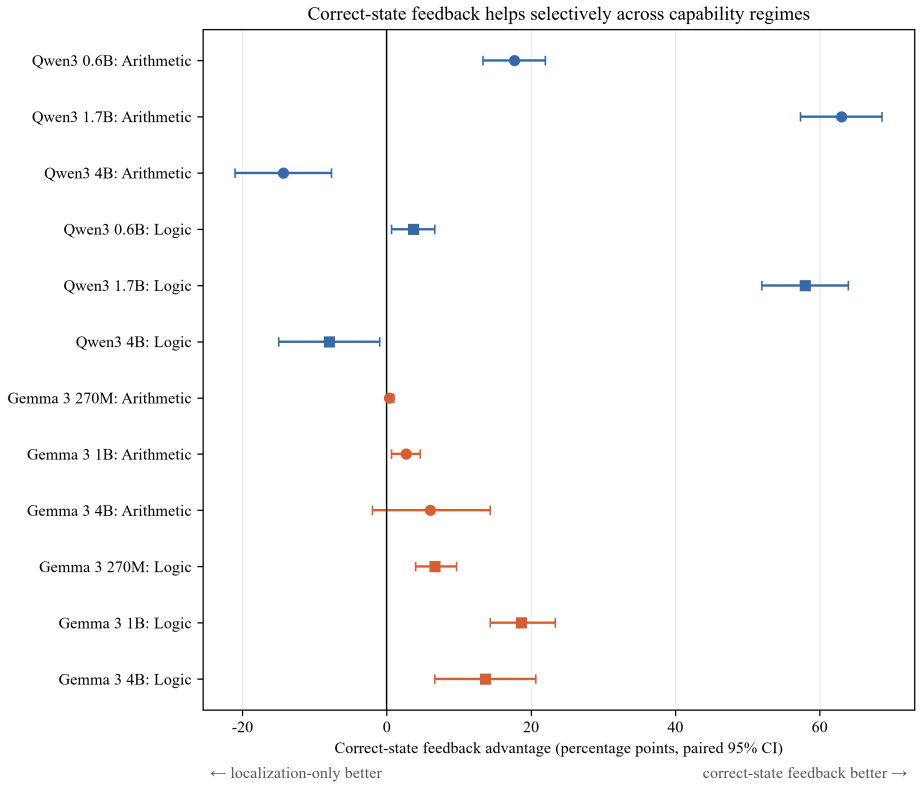


Figure 3: Paired correct-state feedback effects with bootstrap 95% confidence intervals. Large benefits concentrate in particular capability regimes rather than increasing monotonically with parameter count. Negative values indicate that localization-only revision performs better.

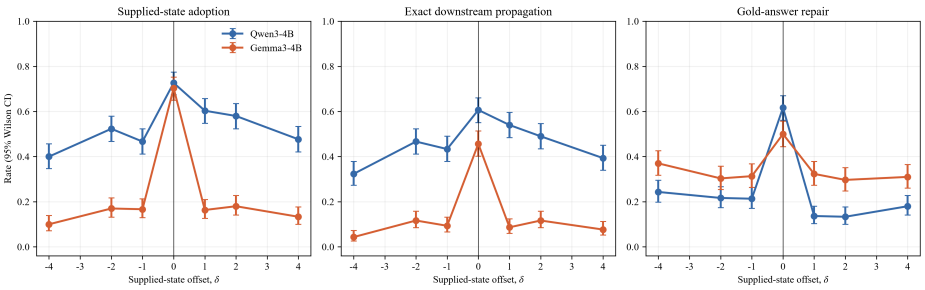


Figure 4: Signed local-state intervention curves for the two 4B models. Only the supplied state is changed. Both families peak sharply at the true state (delta zero), but Qwen3 continues to adopt and propagate false states substantially more often than Gemma 3. Error bars are 95% Wilson intervals.

computational gap. Once autonomous repair becomes reliable, the same state can be redundant or actively harmful. Final accuracy hides this distinction. A system may benefit from a verifier on average and still fail badly on the verifier's rare mistakes.

The contrast between the two 4B models makes this point concrete. Both respond sharply to the exact true state, yet Qwen3 follows false alternatives far more often than Gemma 3. This is not a simple difference in reasoning accuracy. It is a difference in how evidence is weighted. Qwen3 behaves like a more compliant reviser, while Gemma 3 more often preserves its own computation when the supplied state is false. Neither behavior is uniformly better. Compliance is valuable when the verifier is trustworthy; resistance is valuable when it is not.

Capacity therefore changes the role of feedback rather than merely increasing its benefit. Our three regimes are descriptive. They do not imply a universal threshold or a sharp phase transition. The location of the transition differs across the two families, and a denser set of sizes would be needed to estimate it. The stable observation is the tension itself: autonomous computation and responsiveness to external state develop together, but not at the same rate.

This suggests a simple design principle for verifier-guided systems. Localization and state replacement should be separate actions. A verifier can first point to the suspect step and allow the generator to recompute. It should provide a replacement state only when its confidence is high or when the generator cannot repair from localization alone. Evaluation should include matched false feedback and should track local adoption and propagation, not just final accuracy. Otherwise, a system can appear robust while silently transferring control from the generator to an unreliable supervisor.

Controlled tasks are central to this conclusion. We do not seek ecological coverage; we seek causal identification of dependence on externally supplied intermediate states. Here, simplicity is an experimental control. GSM8K and MATH (Cobbe et al., 2021; Hendrycks et al., 2021) measure broad mathematical competence, but they rarely provide a unique first error, a canonical intermediate state, and a deterministic counterfactual continuation. On those tasks, rejecting feedback may simply reflect a different valid derivation. Our cross-task and cross-family replications offer stronger evidence for local causal dependence than an uncontrolled benchmark sweep. Natural-language evaluation remains useful for ecological validity, but it addresses a different question and would complement rather than replace the present design.

7 Limitations

First, the tasks are synthetic transition systems. This is deliberate for causal identification, but the measured rates should not be extrapolated directly to open-ended mathematical proofs, tool-using agents, or human feedback. Second, we evaluate two model families and a maximum size of 4B; larger reasoning-specialized models may verify feedback differently. Third, model family and precision are partially confounded: hardware constraints required FP32 for Gemma 3 and FP16 for Qwen3. Within each family, however, conditions use identical precision, and the central causal comparisons are paired within a model. Fourth, prompts present the verifier as authoritative. Future work should vary source reliability, calibrated confidence, and opportunities for explicit verification. Fifth, exact trace parsing may undercount valid free-form corrections, especially for the smallest model. We report parse rates and count failures conservatively rather than post hoc repairing responses. Finally, the signed intervention covers small integer offsets in additive chains; richer semantic counterfactuals could produce different geometry.

8 Conclusion

Step-level feedback does more than point a model toward a better answer. It can install an intermediate state that controls the rest of the computation. Controlled interventions across arithmetic and symbolic reasoning, Qwen3 and Gemma 3, and multiple model sizes reveal a consistent tension: the capacity to use correct external states arrives with a capacity to propagate false ones. The balance changes with task, size, and family. Reliable verifier-guided reasoning therefore requires evaluating not only whether feedback helps when correct, but how the model behaves when the same channel is wrong.

Reproducibility Statement

The task generators, frozen prompts, intervention definitions, parsing rules, model settings, and statistical procedures are documented in the main text and Appendix A. The supplementary material contains the code and generated datasets needed to reproduce the reported analyses. All confirmatory conditions are retained, including parse failures, and no items are excluded after inference.

AI Use Statement

Generative AI tools were used to assist with brainstorming and refining aspects of the experimental design, drafting and checking portions of the evaluation code, organizing and interpreting experimental results, and editing the manuscript for clarity, structure, and LaTeX formatting. All research questions, experimental decisions, reported results, and scientific claims were reviewed and validated by the authors. The authors independently executed the experiments, verified the evaluation code and statistical outputs, checked relevant literature and citations, and take full responsibility for the final content of the paper. Scientific figures were generated from the verified experimental results; no generative image system was used to create or modify them.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Zhibin Gou, Zhihong Shao, Yeyun Gong, et al. CRITIC: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. Measuring mathematical problem solving with the math dataset. In *NeurIPS*, 2021.

Jie Huang, Xinyun Chen, Swaroop Mishra, et al. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Zachary Kenton, Noah Y. Siegel, Jan Kramar, et al. On scalable oversight with weak llms judging strong llms. *arXiv preprint arXiv:2407.04622*, 2024.

Sung Won Kim and Daniel Khashabi. Challenging the evaluator: Llm sycophancy under user rebuttal. In *Findings of EMNLP*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, et al. RLAIF: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, et al. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Daixuan Liu, Abdallah Nassereldine, Ziming Yang, et al. Large language models have intrinsic self-correction ability. *arXiv preprint arXiv:2406.15673*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, et al. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

William Saunders, Catherine Yeh, Jeff Wu, et al. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Noah Shinn, Federico Cassano, Edward Berman, et al. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2023.

Xiaoshuai Song, Yanan Wu, Weixun Wang, Jiaheng Liu, Wenbo Su, and Bo Zheng. ProgCo: Program helps self-correction of large language models. In *Proceedings of ACL*, 2025.

Jonathan Uesato, Nate Kushman, Ramana Kumar, et al. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Yixuan Weng, Minjun Zhu, Fei Xia, et al. Large language models are better reasoners with self-verification. In *Findings of EMNLP*, 2023.

Wei Zhu, Jian Zhang, Lixing Yu, Kun Yue, and Zhiwen Tang. Dissecting failure dynamics in large language model reasoning. In *Proceedings of ACL*, 2026.

A REPRODUCIBILITY SUMMARY

- Depths: 2, 3, 4; 100 corrupted candidates per depth and task.
- Generation: greedy, temperature 0, 256 new-token cap.
- Primary conditions: localization only, structured correct state, matched wrong state.
- Main outcomes: parse, repair, targeted adoption, exact downstream propagation.
- Main tests: 10,000-sample paired bootstrap confidence intervals and exact two-sided McNemar tests.
- Signed intervention: seven offsets $\times$ 300 items $\times$ two 4B models = 4,200 generations.
- Exploratory and confirmatory outputs are stored separately. Datasets, prompts, outcomes, and statistical tests were frozen before confirmatory inference.

A.1 ROBUSTNESS ACROSS REASONING DEPTH

Table A1 reports a post hoc stratification of the frozen arithmetic results. Each cell contains 100 items. The central family contrast is present at every tested depth: Qwen3-4B propagates matched false states substantially more often than Gemma 3 4B. The pattern is not monotonic in depth, so we treat chain length as a robustness check rather than as an explanatory variable.

Table A1: Exact adoption and propagation by arithmetic-chain depth (%).

Model	Feedback state	Metric	Depth 2	Depth 3	Depth 4
Qwen3-4B	Correct	Adoption	61	71	80
Qwen3-4B	Correct	Propagation	54	56	68
Qwen3-4B	Matched false	Adoption	42	58	54
Qwen3-4B	Matched false	Propagation	41	47	45
Gemma 3 4B	Correct	Adoption	64	69	75
Gemma 3 4B	Correct	Propagation	44	63	69
Gemma 3 4B	Matched false	Adoption	8	10	16
Gemma 3 4B	Matched false	Propagation	1	6	6