WHEN CONCLUSIONS REVERSE: SIGNED VS. MAGNITUDE EFFECTS IN CHAIN-OF-THOUGHT EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As reasoning models are increasingly evaluated by their chain of thought (CoT), perturbation tests are a common way to probe how much the chain affects the model's answer by changing part of it and measuring how the answer shifts. The resulting effect can be scored in two ways: a signed score that keeps both the size and the direction of the change in the reference answer's probability, or a magnitude-only score that keeps only the size. We ask whether this scoring choice alone can change the conclusions drawn from a CoT evaluation when everything else is held fixed. Across about 79,000 perturbation records from nine models and six question-answering datasets, we find that it can. (1) In our main comparison, the two scores rank two models in the opposite order over the 132 questions they share (out of 144 and 150); and (2) they rank a different part of the reasoning first for 40–56% of questions in three of four model–dataset combinations spanning two comparisons, one on a multiple-choice dataset and one on a yes/no dataset. In the broader descriptive analysis, model rankings change in most comparisons, while others show little or no disagreement between the two scores. These results show that signed and magnitude-only scores answer different questions and should be reported separately when both are relevant.

1 INTRODUCTION

Whether a model's chain-of-thought (CoT) actually affects its answer, rather than merely accompanying it, has become central to evaluating reasoning models (Lanham et al., 2023). If the trace does not affect the answer, inspecting it can give a misleading view of what the answer depends on. A common approach is perturbation-based evaluation: modify part of a recorded reasoning trace, sample the model's continuations, and measure how the answer changes (Bogdan et al., 2025; Macar et al., 2026). Such effects can be used to compare how models respond to CoT perturbations (Lanham et al., 2023) or to identify influential reasoning steps (Bogdan et al., 2025; Macar et al., 2026).

For each perturbation record, let p_{ref} denote the proportion of sampled continuations that give the dataset's reference answer. We call $X = p_{\mathrm{ref}}(\mathrm{clean}) - p_{\mathrm{ref}}(\mathrm{perturbed})$ the *signed* score; the corresponding *magnitude-only* score is $|X|$. Under this convention, $X > 0$ means that the perturbation lowers reference-answer support, which we call a harmful effect, while $X < 0$ means that it raises reference-answer support, which we call a helpful effect. These labels refer only to the effect on reference-answer support, not to faithfulness or overall model performance. The question is whether an evaluation should distinguish between changes that lower and raise reference-answer support, or ignore the sign and consider only how large the change is. Signed X keeps this distinction, whereas $|X|$ keeps only the size of the change. Figure 1 illustrates this distinction and two conclusions it can affect.

The distinction between signed and magnitude-only quantities is not itself new, and prior work has shown more broadly that measurement choices can alter interpretability conclusions (Wang et al., 2026; Young, 2026; Edin et al., 2025). Our question is instead empirical, and we test it for matched cross-model ordering and reasoning-span selection:

> *When the perturbation measurements are held fixed, can discarding their direction change the conclusion drawn from them?*

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.

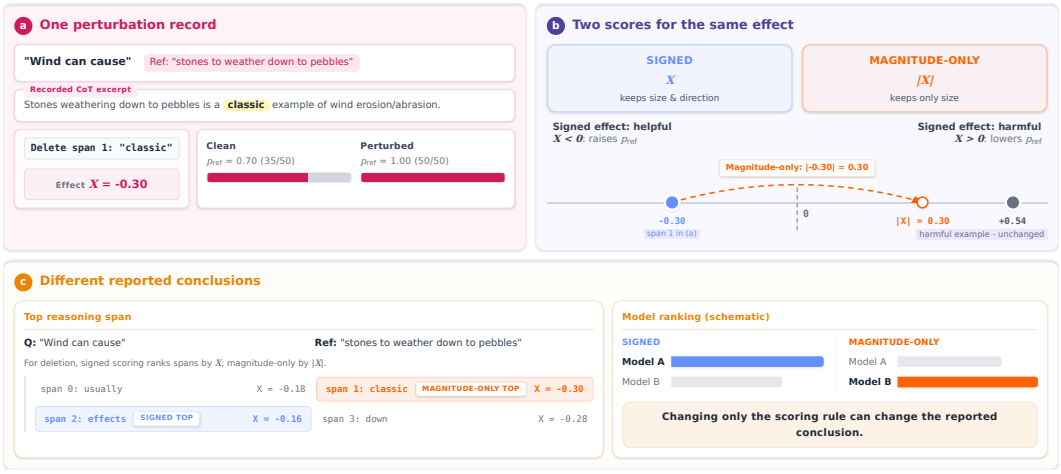


Figure 1: **The same perturbation records can lead to different conclusions depending on how the effect is scored.** **(a)** For a Qwen3.6 OpenBookQA item (Mihaylov et al., 2018), deleting the single-word span "classic" raises reference-answer support from 0.70 to 1.00, giving $X = -0.30$. **(b)** Signed scoring retains -0.30, whereas magnitude-only scoring uses $|X| = 0.30$; a harmful signed effect from another item is shown for contrast. **(c)** For the same item under deletion, with all four perturbation effects helpful, the two scores rank different reasoning spans first; model-ranking bars are schematic.

We find that perturbations often increase reference-answer support rather than reduce it. In the two full OpenBookQA populations from which our main matched comparison is drawn (Mihaylov et al., 2018), helpful effects ($X < 0$) occur in 59–63% of records (Table 3). These populations include only items whose recorded clean-trace answer matches the dataset's reference answer before any perturbation. The models differ much more in how much reference-answer support they gain from helpful effects than in how much they lose from harmful effects, even though helpful effects occur at similar rates. Because signed and magnitude-only scoring combine these effects differently, they can rank the same two models in opposite orders. This ranking reflects how the models respond to perturbations, not their benchmark accuracy.

Both matched comparisons reverse: Gemma-4 versus Qwen3.6 on OpenBookQA and Gemma-2 versus Gemma-4 on BoolQ (Clark et al., 2019); the latter is more sensitive to output termination (continuations that end without an answer). Signed and magnitude-only scoring also rank non-overlapping reasoning spans first for approximately 40–56% of items across perturbation operators in three of the four primary model–dataset populations.

In our reanalysis of the independently released successive-prefix effects from Du et al. (2026), signed and magnitude-only scoring select different top steps in 36.7% of trajectories. In a broader descriptive analysis of roughly 79,000 perturbation records from nine models and six datasets, most matched comparisons also change ordering, although some populations show little difference between signed and magnitude-only scoring.

Our contributions are:

1. **A controlled comparison of signed and magnitude-only scoring.** We apply both scoring rules to the same CoT perturbation measurements and test whether they change which model has the higher average perturbation score or which reasoning span is ranked first.
2. **Evidence that discarding sign can change evaluation conclusions.** Signed and magnitude-only scoring can give different model orderings and rank different reasoning spans first.
3. **An explanation for when the two scores disagree.** The OpenBookQA reversal is driven mainly by differences in how strongly perturbations raise reference-answer support. When most perturbations have little or no effect, the two scores agree much more closely.

2 Related Work

2.1 Perturbation-based evaluation of reasoning traces

A growing literature probes chain-of-thought reasoning by modifying or resampling reasoning content and measuring downstream model behavior. Lanham et al. (2023) truncate, corrupt, paraphrase, or replace CoT content and measure whether the resulting answer changes. At finer granularity, Thought Anchors estimates sentence-level importance by resampling from different points in the reasoning trace and measuring how much the final answer changes (Bogdan et al., 2025), while Thought Branches uses on-policy resampling by letting the model continue from its own intermediate reasoning steps to study the causal influence and resilience of those steps (Macar et al., 2026).

CoT does not always improve answer quality, and the model's final answer does not always depend on the reasoning trace. Liu et al. (2025) identify tasks on which eliciting CoT reduces accuracy, while Lanham et al. (2023) report that early truncation changes the final answer in fewer than 10% of cases across several datasets. These studies use different interventions and estimands from ours, but they show that changes to the reasoning trace do not always make the final answer worse.

Recent work further broadens CoT-faithfulness evaluation through multi-method benchmarks, controlled counterfactual settings, metric meta-evaluation, and optimization studies (Shen et al., 2026; Han et al., 2026; Gur-Arieh et al., 2026; Sun et al., 2026). Our question starts after the primitive perturbation effect has been measured: *can the treatment of its direction alone change the conclusion of the evaluation?*

2.2 Signed, magnitude-only, and measurement-dependent evaluation

Signed and magnitude-only quantities already appear in reasoning research, but previous work differs in what is being measured and in whether both scores are applied to the same realized measurements. TTS uses absolute treatment-effect magnitude for step importance (Zhao et al., 2025), NLDD retains the sign of corruption-induced effects (Ye et al., 2026), Du et al. (2026) define signed rollout-based step advantage but use its magnitude to identify consequential steps, and Step-Saliency aggregates absolute attention–gradient influence (Xu et al., 2026).

The closest paired precedent we identified is Wang et al. (2026), who compare reasoning-token prioritization under signed integrated-gradient scores and their absolute values, obtaining different downstream fine-tuning performance. Their primitive is attribution-based rather than a behavior-derived perturbation effect, and they do not study matched cross-model ordering under signed and magnitude-only scoring.

Broader interpretability work likewise shows that conclusions depend on measurement choices: MIB contrasts high-value with high-magnitude circuit construction (Mueller et al., 2025), while normalized AOPC can alter cross-model faithfulness rankings (Edin et al., 2025). Quantus and MetaQuantus study evaluator specification and reliability (Hedström et al., 2023b;a), while Young (2026) obtains different CoT model rankings from alternative faithfulness measurements applied to the same underlying traces.

Positioning. To our knowledge, prior work establishes several components of this measurement problem but does not compare signed and magnitude-only scoring on the same fixed behavior-derived CoT perturbation records while testing consequences for both matched cross-model ordering and reasoning-span selection.

Table 1 contrasts our setting with the closest precedents. Detailed definitions, qualifications, and a broader comparison are provided in Appendix B.

Table 1: **Closest precedents and their relation to our setting.** Checks indicate properties demonstrated in the original work. Column definitions and qualifications are provided in Appendix B.

Work	Main goal	Behavior-derived	Signed	Magnitude	Both scores same values	Model order	Reasoning-unit selection
Wang et al. (2026)	Select reasoning tokens for fine-tuning	–	✓	✓	✓	–	✓[a]
TTS (Zhao et al., 2025)	Estimate causal step importance	✓	✓	✓	–	–	–
NLDD (Ye et al., 2026)	Measure faithfulness decay and reasoning horizon	✓	✓	–	–	–	–
Du et al. (2026)	Compare judged and rollout-based step importance	✓	✓	✓	–[b]	–	–
Young (2026)	Test measurement-dependent CoT conclusions	–	–	–	–	✓	–
Edin et al. (2025)	Normalize feature-attribution faithfulness across models	✓	–	–	–	✓	–
This paper	**Test consequences of signed vs. magnitude-only scoring**	✓	✓	✓	✓[c]	✓	✓

3 Measurement Setup

We study perturbation-based evaluation of recorded chain-of-thought reasoning traces. For each item, a model produces a completed reasoning trace. We select four predefined perturbation sites (spans) at different positions in the trace and modify one site at a time using masking, deletion, or typed substitution. We then sample continuations from the clean and perturbed contexts.

For non-GSM8K populations, each site is a single selected word near one of four target relative positions in the trace; for GSM8K (Cobbe et al., 2021), the sites are near-equal blocks of consecutive reasoning lines. Our primary matched analyses use OpenBookQA (Mihaylov et al., 2018) and BoolQ (Clark et al., 2019); exact populations and matching rules are given in §4, with full span-construction details in Appendix C.

3.1 Primitive effect and scoring rules

For each item, the protocol contains one shared pool of 50 clean continuations and a separate pool of 50 continuations for each item–operator–span perturbation condition. Let n_{ref} be the number of sampled outputs whose parsed answer matches the dataset's reference answer. We estimate $p_{\text{ref}} = n_{\text{ref}}/50$. The target is the dataset reference answer, not the model's own clean-context answer, and invalid parses remain in the denominator.

For a realized perturbation record, the behavioral effect is $X = p_{\text{ref}}(\text{clean}) - p_{\text{ref}}(\text{perturbed})$. We compare the signed score X with the magnitude-only score $|X|$; $X > 0$ denotes lower reference-answer support after perturbation and $X < 0$ higher support.

The signed score retains the direction of the behavioral change, whereas the magnitude-only score retains only its size. The absolute-value transformation is applied to each perturbation record before aggregation, so magnitude-only summaries average $|X|$, not the absolute value of an averaged signed effect.

Within each model, these two scores are computed from the same realized perturbation records. The questions, reasoning traces, perturbations, sampled outputs, parser, and reference answers are therefore held fixed; only the transformation from X to $|X|$ changes.

3.2 Evaluation endpoints

Matched model ordering. Each complete item contributes 12 perturbation records: four spans under three operators. For each model, we first average the 12 record-level scores within each item and then average across the items shared by both models. We compare the models separately under signed and magnitude-only scoring: $D_S = \bar{X}_a - \bar{X}_b$ and $D_M = \overline{|X|}_a - \overline{|X|}_b$. If D_S and D_M have opposite signs, the two scoring methods rank the models in opposite orders.

Reasoning-span selection. For a given item and perturbation operator, let X_j be the effect of perturbing span j. The two scores select $T_S = \arg\max_j X_j$ and $T_M = \arg\max_j |X_j|$.

Signed scoring prioritizes the span whose perturbation most lowers reference-answer support. If all perturbations raise support, it selects the least helpful one. Magnitude-only scoring instead prioritizes the span that produces the largest change in either direction.

Because two or more spans can receive the same top score, T_S and T_M may contain more than one span. We define *top-span disagreement* as $\mathbb{1}[T_S \cap T_M = \varnothing]$: the two scores disagree only when their top-span sets have no overlap. Any shared top span therefore counts as agreement.

4 Experimental Design

4.1 Primary populations and matched comparisons

The primary analysis uses three model checkpoints: Gemma-2-9B-IT, Gemma-4-12B, and Qwen3.6-35B-A3B; revision status is reported in Appendix C. We evaluate Gemma-2 and Gemma-4 on BoolQ, and Gemma-4 and Qwen3.6 on OpenBookQA.

The four model–dataset populations contain 200, 200, 144, and 150 eligible items, respectively. Eligibility is model-specific: an item enters a model's population only if the answer parsed from its recorded clean trace matches the reference answer and the trace satisfies the usable-trace requirements described in Appendix C. Cross-model analyses use the intersection of the corresponding eligible item sets.

We set the matched model pairs, endpoints, estimator, and bootstrap procedure before examining the signed-versus-magnitude results; Appendix C gives further details.

OpenBookQA is the main comparison; BoolQ is treated as supporting evidence because its comparison is more sensitive to immediate output termination (continuations that end without producing an answer; Appendix F).

Primary confidence intervals use 5,000 item-bootstrap resamples, with matched items resampled jointly across the two models. Full resampling details and the two-stage bootstrap sensitivity analysis are reported in Appendix D.

4.2 Perturbation protocol

For the primary BoolQ and OpenBookQA experiments, we apply masking, deletion, and typed substitution to each of the four predefined spans, yielding 12 perturbation records per complete item.

Masking replaces the selected word with a mask token, deletion removes it, and typed substitution replaces it with a precomputed same-type distractor word. Appendix C provides a worked example of all three operators, together with full details of span selection, perturbation implementation, continuation sampling, decoding, parsing, and the GSM8K-specific construction.

In the main experiments, continuations are sampled at temperature 0.7. Appendix D.3 shows that the model-ordering reversal also persists at other temperatures.

4.3 Broader analysis and external reanalysis

For the broader descriptive analysis, a post-hoc parser-validity check excluded Qwen3.6 BoolQ because fewer than 5% of its clean outputs could be parsed as valid answers. After this exclusion, the analysis contains 42 model–dataset populations, 78,853 perturbation records, and 47 matched comparisons. Full quality-control criteria and threshold checks are reported in Appendix G.

As an external check, we reanalyze the independently released reasoning-step effects of Du et al. (2026). We use the same disagreement rule as in our span analysis: signed and magnitude-only scoring disagree only when their top-step sets have no overlap. The result is reported in §5.2, with source, eligibility, and sampling details in Appendix H.

5 RESULTS

5.1 SIGNED AND MAGNITUDE-ONLY SCORING CAN REVERSE MODEL ORDERING

We first ask whether changing only how the perturbation effect is scored can alter a cross-model conclusion when the underlying perturbation records and matched items are held fixed.

Gemma-4 and Qwen3.6 share 132 OpenBookQA items with usable perturbation records for both models. Their model-specific eligible populations contain 144 and 150 items, respectively. On this matched set, the Gemma-4-minus-Qwen3.6 difference in average perturbation score is $D_S = -0.178$ under signed scoring and $D_M = +0.143$ under magnitude-only scoring. The change in sign means that the two scoring rules reverse the models' ordering by average perturbation score (Table 2).

Table 2: **Main OpenBookQA matched comparison.** Differences are Gemma-4 minus Qwen3.6 over the 132 shared items. The opposite signs show that signed and magnitude-only scoring order the models differently by average perturbation score. Intervals are primary 95% item-bootstrap confidence intervals.

Scoring rule	Difference	95% CI	Score ordering
Signed	$D_S = -0.178$	$[-0.238, -0.117]$	Gemma-4 $<$ Qwen3.6
Magnitude-only	$D_M = +0.143$	$[0.091, 0.196]$	Gemma-4 $>$ Qwen3.6

The reversal is stable across several checks. Removing each matched item in turn preserves the reversal in all 132 leave-one-item-out analyses, and the two-stage bootstrap preserves the reversal in all 5,000 replicates (Appendix D). As another check, removing the fourth perturbation site, the site selected closest to the end of the trace, produces nearly identical estimates ($D_S = -0.179$, $D_M = +0.144$). The reversal is therefore not driven by a single item or by the latest-position perturbation site.

The supporting BoolQ comparison also reverses. Gemma-2 and Gemma-4 share 142 BoolQ items with usable perturbation records for both models; each model-specific eligible population contains 200 items. On this matched set, the Gemma-2-minus-Gemma-4 difference is $D_S = -0.344$ under signed scoring and $D_M = +0.290$ under magnitude-only scoring, again reversing the models' ordering by average perturbation score. We treat BoolQ as supporting evidence because its comparison is more sensitive to output termination, as described in §4.1.

The reversal also persists when we change the continuation-sampling temperature. For both matched comparisons, $D_S < 0$ and $D_M > 0$ at $T \in \{0.3, 0.7, 1.0\}$; the OpenBookQA reversal also persists at $T = 1.3$. The effect sizes become smaller as temperature increases, but the two scoring rules continue to rank the models in opposite orders (Appendix D.3).

5.2 SIGNED AND MAGNITUDE-ONLY SCORES CAN PRIORITIZE DIFFERENT REASONING SPANS

We next ask whether the scoring choice changes which part of a reasoning trace is ranked first. Using the top-span disagreement measure from §3.2, the two scores disagree when their top-span sets have no common span.

Each population–operator result uses the eligible items for that population: 200 for Gemma-2 BoolQ, 200 for Gemma-4 BoolQ, 144 for Gemma-4 OpenBookQA, and 150 for Qwen3.6 OpenBookQA. Disagreement is substantial in three of the four primary populations, reaching approximately 40–56% across perturbation operators. Gemma-4 on BoolQ shows lower disagreement, approximately 13.5–16.5% across operators. Exact operator-level estimates and confidence intervals are shown in Fig. 2 and reported in Appendix E.

Appendix E shows that disagreement is especially common when all four perturbations raise reference-answer support ($X < 0$): rates range from 54–96% on such items, compared with 6–22% on items containing at least one harmful effect ($X > 0$). This pattern follows from the two score definitions. When all four effects are negative, signed scoring selects the least negative effect, whereas magnitude-only scoring selects the largest absolute change. The observed disagreement therefore reflects a

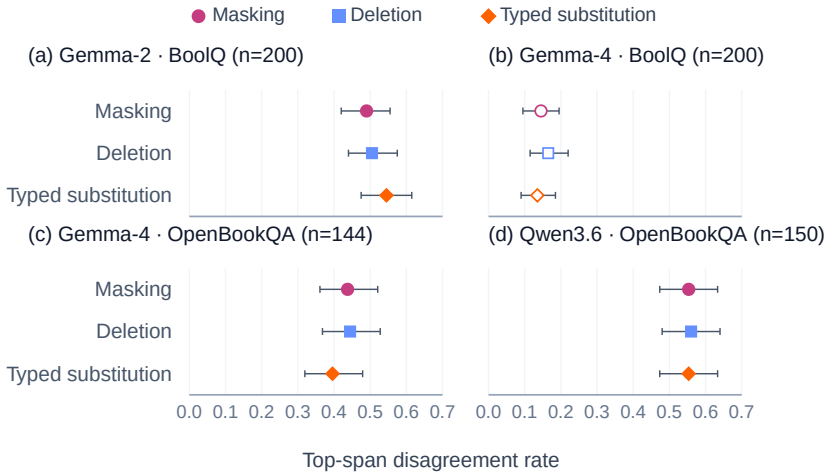


Figure 2: **Signed and magnitude-only scoring can select different top reasoning spans.** Top-span disagreement rates are shown across the four primary populations and three perturbation operators, with 95% item-cluster bootstrap intervals ($B = 5{,}000$). A disagreement is counted only when the two top-span sets have no common span; any shared top span counts as agreement. Exact values are reported in Appendix E.

difference in what the two scores prioritize, not evidence that one provides a universally better measure of span importance.

Evidence from an independently released step-level analysis. We also reanalyze the successive-prefix correctness effects released by Du et al. (2026). Top-step disagreement occurs in 2,247 of 6,127 eligible trajectories (36.7%). In this dataset, the effect measures the change from adding the next reasoning step rather than perturbing part of a completed trace. We therefore treat this as external evidence for the step-selection result rather than as a replication of the model-ordering result. Confidence intervals and robustness checks are reported in Appendix H.

5.3 Helpful and harmful contributions clarify model-ordering reversals

We next ask why the same matched records can produce opposite model orderings. We separate each perturbation effect into a harmful contribution and a helpful contribution: $H = \max(X, 0)$ and $P = \max(-X, 0)$. H is positive when perturbation lowers reference-answer support, whereas P is positive when perturbation raises it. Signed scoring combines them as $X = H - P$, while magnitude-only scoring uses $|X| = H + P$.

For each model, $E[H]$ and $E[P]$ are the average harmful and helpful contributions across perturbation records. For a model pair, Δ_H and Δ_P denote the corresponding between-model differences. The algebra relating these quantities to model-ordering reversals is given in Appendix A.

For the main OpenBookQA comparison, $\Delta_H = -0.017$ and $\Delta_P = +0.161$. The models therefore differ much more in how strongly perturbations raise reference-answer support than in how strongly they lower it (Fig. 3, panel b).

This difference is not apparent from how often helpful effects occur. In the two full OpenBookQA populations, $X < 0$ occurs in 62.5% of Gemma-4 records and 59.1% of Qwen3.6 records. Helpful effects therefore occur at similar rates, while their average contribution differs substantially (Table 3).

A breakdown by output type shows that about 70% of the helpful-contribution difference is associated with valid-but-wrong outputs, while the remainder is associated with invalid outputs. Restricting the analysis to valid outputs still preserves the reversal on the subset where the comparison is defined. Because many bootstrap replicates are undefined under this restriction, we treat this only as supporting sensitivity evidence. Full results are reported in Appendix F.

Table 3: **Effect-sign frequencies and average harmful and helpful contributions across key populations.** $H = \max(X, 0)$ and $P = \max(-X, 0)$, with $E[H]$ and $E[P]$ denoting their average contributions across perturbation records.

Population	$P(X > 0)$	$P(X < 0)$	$P(X = 0)$	$E[H]$	$E[P]$
Gemma-2 BoolQ	.122	.626	.253	.023	.383
Gemma-4 BoolQ	.169	.214	.617	.051	.069
Gemma-4 OpenBookQA	.083	.625	.292	.013	.261
Qwen3.6 OpenBookQA	.227	.591	.182	.032	.105
Gemma-4 GSM8K	.060	.066	.874	.039	.046

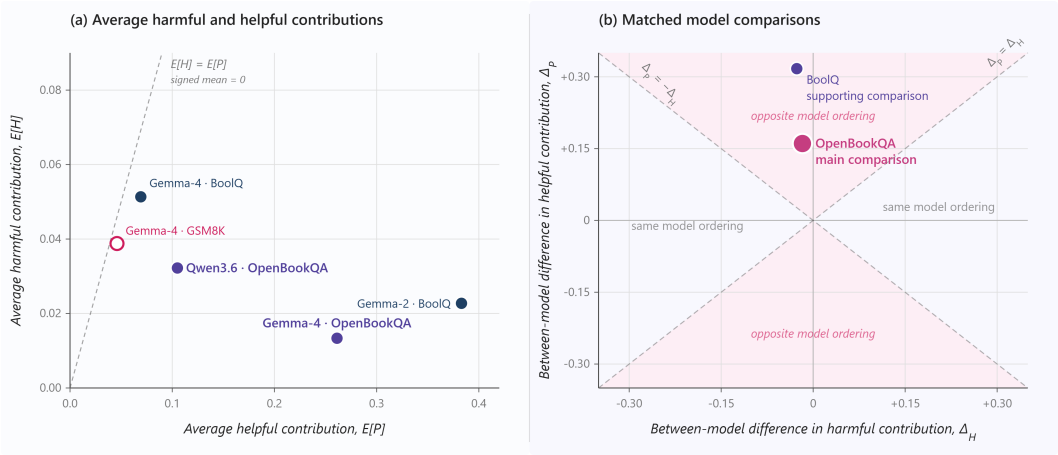


Figure 3: **Helpful and harmful contributions characterize signed–magnitude differences. (a)** Average harmful contribution $E[H]$ and helpful contribution $E[P]$ for the four primary populations and Gemma-4 GSM8K. **(b)** Between-model differences in average harmful and helpful contributions for the matched comparisons.

5.4 Divergence is widespread but not universal

We finally ask whether the same pattern appears beyond the two primary comparisons. The broader analysis contains 42 model–dataset populations and 78,853 perturbation records. Among 47 matched model comparisons, 36 (77%) change ordering between signed and magnitude-only scoring. In 23 of these comparisons, the 95% intervals for the two model differences exclude zero in opposite directions (Appendix G).

These 47 comparisons are not independent. They reuse some of the same models, datasets, items, and experimental settings, so we treat this analysis as descriptive evidence rather than as 47 separate replications.

The two scores do not always disagree. Gemma-4 on GSM8K is the clearest example (Fig. 3, panel a). Of its 1,763 perturbation records, 1,541 (87.4%) have exactly zero effect. Its average harmful and helpful contributions are also small, $E[H] \approx 0.039$ and $E[P] \approx 0.046$, and the retained GSM8K matched comparisons do not reverse model ordering.

One reason is that clean reference-answer support is already very high on this population, so there is little room for a perturbation to increase it. This shows that signed and magnitude-only scoring do not necessarily lead to different conclusions in every setting.

Overall, the difference between the two scores depends on how often perturbations lower, raise, or leave unchanged reference-answer support, and on how large those changes are.

6 DISCUSSION

Reporting chain-of-thought evaluation. Signed and magnitude-only scores answer different questions. A magnitude-only model ordering therefore reflects perturbation sensitivity, not how strongly perturbation harms reference-answer support. Neither score is inherently inappropriate: magnitude-based scoring is appropriate when the target is sensitivity regardless of direction, while signed scoring is appropriate when direction itself matters. When both questions are relevant, we recommend reporting both scores, together with $P(X > 0)$, $P(X < 0)$, $P(X = 0)$, $E[H]$, and $E[P]$; Appendix I provides a reporting checklist.

Implications for drawing the right conclusions from CoT evaluation. The span-selection results expose the same distinction at the item level. The two criteria diverge particularly when all candidate perturbations raise reference-answer support: magnitude-only scoring selects the span producing the largest such change, whereas signed scoring selects the largest signed effect, which in our definition is the least helpful perturbation. Magnitude-based methods such as TTS (Zhao et al., 2025) and the step-prioritization use of rollout-effect magnitude in Du et al. (2026) therefore answer a different selection question from a direction-preserving criterion; on all-helpful items, an absolute-effect ranking selects the most helpful perturbation. By contrast, NLDD (Ye et al., 2026) retains the sign of corruption-induced effects and therefore does not collapse helpful and harmful changes. Our results extend the signed-versus-absolute dependence Wang et al. (2026) found for attribution to behavior-derived effects and model ordering, and add sign retention to normalization (Edin et al., 2025) and classifier choice (Young, 2026) as axes that change rankings.

Open questions. An important open question is why perturbations increase reference-answer support so often. Because eligible items already have a correct answer in the recorded clean trace, one possibility is that a helpful perturbation shifts the sampled continuation distribution back toward an answer the model already produces in the clean condition. Output termination provides a concrete example in BoolQ, while the OpenBookQA output-state decomposition shows that invalid-output contribution accounts for only about 30% of the between-model helpful-effect difference (Appendix F); we therefore do not attribute helpful effects to a single mechanism.

Limitations. The main evidence comes from two matched model comparisons, with OpenBookQA as the main comparison and BoolQ as supporting evidence. The broader analysis reuses models, datasets, items, and design choices, so its comparisons should not be treated as independent replications. The study is also limited to items whose recorded clean-trace answer matches the reference answer, four predefined perturbation sites per trace (with a different line-block construction for GSM8K), a 60-token continuation-sampling budget, fixed top-p and top-k settings, English QA benchmarks, and open-weight models. The matched model-ordering reversal persists across the tested continuation temperatures (Appendix D.3), but generalization beyond these decoding choices remains open. The effect is also not universal: some populations show little difference between signed and magnitude-only scoring. Further limitations related to finite sampling, parsing, output termination, the post-hoc parser-validity check, and the sample-count tail in the external release are discussed in Appendix I.

7 CONCLUSION

Signed and magnitude-only scoring can lead to different conclusions even when applied to exactly the same realized CoT perturbation records. In our experiments, discarding effect direction can reverse matched model orderings and change which reasoning spans are prioritized, particularly when perturbations increase reference-answer support. These differences are not universal, but they show that signed and magnitude-only scoring should not be treated as interchangeable. Perturbation-based evaluations should therefore make the treatment of direction explicit and report both scores when both the direction and the size of the change are relevant.

AI USE STATEMENT

We used generative AI tools to aid and polish writing, retrieve and synthesize related literature, support research ideation and execution, and draft portions of the manuscript. As part of research ideation and execution, the tools assisted with refining hypotheses, providing feedback on analysis and experimental methodology, implementing analysis code, and interpreting results. We additionally used generative AI for figure layout and plotting code, reference formatting, and manuscript organization. We did not use generative AI tools to generate synthetic experimental data or to formulate or prove mathematical claims.

The authors reviewed all AI-assisted work. AI-assisted analysis code and reported numerical results were checked against the stored experimental records and result tables, while literature references and factual claims were checked against the cited sources. The authors made the final scientific decisions and take responsibility for the final content of the paper, including its text, claims, analyses, figures, and results.

REPRODUCIBILITY STATEMENT

Reproducibility materials are organized across the appendix and supplementary package. Appendix C documents the generation, perturbation, parsing, matching, and model-revision protocol. Appendix D reports the matched-model results and robustness checks, including sensitivity to continuation temperature. The following appendices report the span-selection and output-state analyses. Appendix G documents parser-validity quality control, threshold sensitivity, and the retained descriptive grid; and Appendix H documents the external reanalysis of Du et al. (2026), including the release version, effect field, clustering unit, and parsing correction.

The supplementary package contains analysis code, prompts, parser definitions, seeds, sampling and bootstrap configurations, population information, result tables, and supporting metadata. Signed and magnitude-only analyses use the same within-model perturbation records, with cross-model analyses restricted to shared usable items. Re-running the primary bootstrap with its recorded seed reproduces the reported primary confidence intervals.

An anonymous version of the code and reproducibility materials is available at `https://anonymous.4open.science/r/cot-evaluation-0853/`.

REFERENCES

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which LLM reasoning steps matter? In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=VnSlfeRCaU.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv*, abs/1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

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*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Kevin Du, Alexander M. Hoyle, Laura Ruis, and Acyr Locatelli. Legibility is not interpretability: Evaluating judged importance versus actual importance in chain-of-thought reasoning steps. In

Third Conference on Language Modeling, 2026. URL https://openreview.net/forum?id=Qhex5RJmzs.

Joakim Edin, Andreas Geert Motzfeldt, Casper L. Christensen, Tuukka Ruotsalo, Lars Maaløe, and Maria Maistro. Normalized AOPC: Fixing misleading faithfulness metrics for feature attributions explainability. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1730, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.86. URL https://aclanthology.org/2025.acl-long.86/.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21/.

Yoav Gur-Arieh, Ana Marasović, and Mor Geva. Faithfulness metrics don't measure faithfulness: A meta-evaluation with ground truth. *arXiv*, abs/2605.25052, 2026. URL https://arxiv.org/abs/2605.25052.

Yunseok Han, Yejoon Lee, and Jaeyoung Do. RFEval: Benchmarking reasoning faithfulness under counterfactual reasoning intervention in large reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2Gc8aj0afg.

Anna Hedström, Philine Bommer, Kristoffer K. Wickstrøm, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. The meta-evaluation problem in explainable AI: Identifying reliable estimators with MetaQuantus. *Transactions on Machine Learning Research*, 2023a. URL https://openreview.net/forum?id=j3FK00HyfU.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023b. URL https://jmlr.org/papers/v24/22-0142.html.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv*, abs/2307.13702, 2023. URL https://arxiv.org/abs/2307.13702.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015/.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38489–38517. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25t.html.

Uzay Macar, Paul C. Bogdan, Senthooran Rajamanoharan, and Neel Nanda. Thought branches: Interpreting LLM reasoning requires resampling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bVsAuIOvJ5.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. FaithCoT-Bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lN3yKqqzF1.

Jingyi Sun, Qianli Wang, Pepa Atanasova, Nils Feldhus, and Isabelle Augenstein. Investigating the interplay between contextual and parametric chain-of-thought faithfulness under optimization. *arXiv*, abs/2605.24960, 2026. URL https://arxiv.org/abs/2605.24960.

Siyuan Wang, Yanchen Liu, and Xiang Ren. Segment-level attribution for selective learning of long reasoning traces. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=C1UD4FLFPL. arXiv:2602.00425.

Xiaoyu Xu, Yulan Pan, Xiaosong Yuan, Zhihong Shen, Minghao Su, Yuanhao Su, and Xiaofeng Zhang. Reasoning fails where step flow breaks. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26337–26353, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1212. URL https://aclanthology.org/2026.acl-long.1212/.

Donald Ye, Max Loffgren, Om Kotadia, Linus Wong, and Jonas Rohweder. Mechanistic evidence for faithfulness decay in chain-of-thought reasoning. In *ICLR 2026 Workshop on Latent & Implicit Thinking: Going Beyond CoT Reasoning*, 2026. URL https://openreview.net/forum?id=wVj7dB7waI. arXiv:2602.11201.

Richard J. Young. Measuring faithfulness depends on how you measure: Classifier sensitivity in LLM chain-of-thought evaluation. *arXiv*, abs/2603.20172, 2026. URL https://arxiv.org/abs/2603.20172.

Jiachen Zhao, Yiyou Sun, Weiyan Shi, and Dawn Song. Can aha moments be fake? identifying true and decorative thinking steps in chain-of-thought, 2025. URL https://arxiv.org/abs/2510.24941. arXiv:2510.24941; listed by the authors as Findings of EMNLP 2026.

A Notation and estimands

The primitive behavioral perturbation effect is

$$X = p_{\text{ref}}(\text{clean}) - p_{\text{ref}}(\text{perturbed}).$$

The two scores applied to each perturbation record are the signed score X and the magnitude-only score $|X|$.

Aggregation and interpretation. The magnitude-only transformation is applied to each perturbation record before item- or model-level aggregation. We therefore average realized values of $|X|$ rather than taking the absolute value of an averaged signed effect. Under signed scoring, larger X means a larger decrease in reference-answer support. If all candidate spans have $X < 0$, the signed criterion selects the least support-increasing perturbation.

We separate each effect into harmful and helpful contributions,

$$H = \max(X, 0), \qquad P = \max(-X, 0),$$

so that

$$X = H - P, \qquad |X| = H + P.$$

For a matched model comparison,

$$D_S = \Delta_H - \Delta_P, \qquad D_M = \Delta_H + \Delta_P.$$

Therefore,

$$D_S D_M = \Delta_H^2 - \Delta_P^2,$$

and a nonzero ordering reversal satisfies

$$D_S D_M < 0 \iff |\Delta_P| > |\Delta_H|.$$

This identity applies to the linearly aggregated matched model-ordering endpoint. It does not automatically extend to medians, ranks, correlations, or span-selection endpoints.

For reasoning spans indexed by j,

$$T_S = \arg\max_j X_j, \qquad T_M = \arg\max_j |X_j|.$$

We call

$$\mathbb{1}[T_S \cap T_M = \varnothing]$$

top-span disagreement. All tied top spans are retained. A disagreement is counted only when the two top-span sets have no common member, so any overlap counts as agreement.

All of these quantities describe behavioral responses to perturbations of recorded reasoning traces. They are not direct measures of semantic faithfulness, mechanistic reliance, or pre-answer causal necessity.

B Extended literature positioning

The compact comparison in Table 1 focuses on the dimensions most directly related to our empirical question. Here we provide a more detailed view of the surrounding literature by distinguishing the primitive quantity being measured, how effect direction is treated, whether signed and magnitude-only versions are compared on the same underlying quantity, and what downstream conclusion is studied.

Several prior methods use closely related ingredients. TTS begins from signed behavioral treatment effects and uses their absolute magnitude for reasoning-step importance (Zhao et al., 2025); NLDD retains signed corruption-induced effects (Ye et al., 2026); Du et al. (2026) define signed rollout-based step advantage and use its absolute magnitude to identify consequential steps; and Wang et al. (2026) directly compare original and absolute integrated-gradient scores for reasoning-token prioritization. Related attribution work also discards direction when it is not part of the target construct: Step-Saliency uses absolute attention–gradient influence during step-level aggregation (Xu et al., 2026).

Outside this specific setting, MIB contrasts high-value and high-magnitude component selection (Mueller et al., 2025), while normalized AOPC and alternative CoT-faithfulness operationalizations show that measurement choices can alter model rankings (Edin et al., 2025; Young, 2026). Quantus and MetaQuantus provide broader precedent for treating the reliability and specification of interpretability evaluation as objects of study (Hedström et al., 2023b;a).

We therefore do not claim mathematical novelty for X versus $|X|$, for signed versus magnitude-only importance in general, or for the broader observation that measurement choices can change scientific conclusions. To our knowledge, the narrower combination studied here has not previously been isolated: holding each model's realized behavior-derived CoT perturbation records fixed, changing only whether effect direction is retained, and testing the consequences for matched cross-model ordering and reasoning-span selection. Tables 4, 5, and 6 expand the main-text comparison along these dimensions.

Table 4: **Detailed positioning relative to reasoning and CoT precedents (part 1).**

Work	Effect / primitive	Treatment of direction	Paired?	Reported consequence	Relation to our setting
Lanham et al. (2023)	Answer behavior after CoT truncation, corruption, paraphrasing, or replacement	Primarily answer-change or answer-consistency measures rather than a signed scalar effect	–	Behavioral CoT faithfulness under intervention	Perturbation-based CoT evaluation, but without a comparison of signed and magnitude-only scoring for the same primitive effect.
Bogdan et al. (2025)	Sentence-level counterfactual importance from resampled reasoning continuations	Uses divergence between downstream answer distributions, so the importance score is direction-discarding	–	Identification of influential reasoning sentences	Uses fine-grained behavioral interventions, but a nondirectional distributional importance measure rather than a signed reference-answer effect.
Macar et al. (2026)	Reasoning influence estimated through on-policy resampling and counterfactual continuations	Includes resampling-based causal effects; signed and magnitude-only scoring are not directly contrasted	–	Causal influence and resilience of reasoning steps	Uses a closely related intervention setting, but does not isolate sign retention on the same realized perturbation effects.
TTS (Zhao et al., 2025)	Context-conditioned treatment effect on answer confidence	Defines signed treatment effects and uses their absolute magnitude for direction-agnostic step importance	–	Reasoning-step importance	Uses a related signed behavioral primitive, but does not compare conclusions under signed and magnitude-only scoring.
NLDD (Ye et al., 2026)	Corruption-induced change in answer-supporting logit margin	Retains sign; negative values can indicate that corruption increases answer support	–	Step-level faithfulness decay and reasoning horizon	Retains signed behavioral perturbation effects, but does not report a corresponding magnitude-only analysis of the same effects.

Note. This table expands Table 1 by separating the underlying effect primitive, treatment of direction, whether signed and magnitude-only scoring are directly compared, the reported consequence, and the distinction from our setting. "Paired" denotes applying both scorings to the same underlying realized quantity and comparing their consequences.

Definitions for Table 1. **Behavior-derived** denotes a quantity estimated from a change or contrast in model outputs under intervention, rather than from an attribution score or a classifier applied to text. **Both scores, same values** means that signed and magnitude-only scoring are applied to the same underlying realized quantities and their consequences are compared. **Model order** denotes a demonstrated change in ordering between models. **Reasoning-unit selection** denotes a demonstrated change in reasoning-span, step, or token prioritization, or in a downstream conclusion based on that prioritization.

Notes for Table 1. [a] Wang et al. (2026) report downstream selective-fine-tuning performance rather than a direct token-set disagreement rate. [b] The paired step-selection comparison for Du et al. (2026) is our reanalysis of their released effects and is not an analysis reported in the original work. [c] Perturbation records are held fixed within each model; cross-model comparisons use shared usable items rather than identical reasoning traces or sampled continuations.

Table 5: **Detailed positioning relative to reasoning and CoT precedents (part 2).** Continuation of Table 4.

Work	Effect / primitive	Treatment of direction	Paired?	Reported consequence	Relation to our setting
Du et al. (2026)	Rollout-based signed step advantage	Defines signed advantage and uses its absolute magnitude to identify consequential steps	–[a]	Judged versus behaviorally estimated step importance	Uses both direction and magnitude in the original method, but does not compare signed versus magnitude-only prioritization. That paired comparison is performed in our reanalysis.
Wang et al. (2026)	Token-level integrated-gradient attribution	Compares original signed IG scores with their absolute values	✓	Token prioritization and downstream selective-fine-tuning performance	Directly compares signed and absolute reasoning-token attribution scores, but the primitive is attribution rather than a behavior-derived perturbation effect, and no matched model-order comparison is studied.
Xu et al. (2026)	Attention–gradient step influence	Uses absolute influence during step aggregation to avoid cancellation of positive and negative contributions	–	Step-saliency analysis	Uses direction-discarding aggregation, but is attribution-based rather than behavioral perturbation-based.
This paper	**Reference-answer effect** $X = p_{\mathrm{ref}}(\mathrm{clean}) - p_{\mathrm{ref}}(\mathrm{perturbed})$	**Directly compares X with $\lvert X\rvert$ on the same realized within-model perturbation records**	✓	**Matched cross-model ordering and reasoning-span selection**	**Isolates sign retention while holding each model's realized behavioral CoT perturbation records fixed and matching models on shared usable items.**

[a] Du et al. (2026) define signed step advantage and use its absolute magnitude in their consequential-step criterion. They do not compare the reasoning steps selected by signed versus magnitude-only scoring; the paired step-selection analysis reported in this paper is our reanalysis of their released effects.

Table 6: **Detailed positioning relative to broader measurement-dependence precedents.** These works establish that choices in interpretability measurement and evaluation can change reported conclusions, but they study different primitives or measurement axes from the fixed behavioral CoT perturbation effects considered here.

Work	Effect / primitive	Measurement choice	Paired?	Reported consequence	Relation to our setting
MIB (Mueller et al., 2025)	Causal component scores for circuit localization	Contrasts high-value and high-magnitude component selection	–	Circuit construction and mechanistic-interpretability evaluation	Contrasts signed-value and magnitude-based selection outside CoT reasoning.
Edin et al. (2025)	AOPC perturbation-curve faithfulness score	Changes normalization rather than sign treatment	–	Cross-model faithfulness rankings	Shows measurement-dependent model ordering under a change in normalization rather than direction retention.
Quantus (Hedström et al., 2023b)	Families of explanation-quality estimators	Studies evaluator design and parameterization rather than signed versus magnitude scoring	–	Evaluation and ranking of explanation methods	Shows that interpretability conclusions can depend on evaluator specification.
Meta-Quantus (Hedström et al., 2023a)	Meta-evaluation of explanation-quality estimators	Studies estimator reliability rather than manipulating sign treatment	–	Reliability of explanation-quality estimators	Addresses evaluator reliability rather than CoT perturbations or sign retention.
Young (2026)	Classifier judgments applied to the same behaviorally influenced CoT traces	Compares alternative measurement operationalizations rather than signed and magnitude-only transformations	–	Faithfulness rates and model rankings	Related CoT evidence for measurement-dependent model ordering, but the measurement change is classifier definition rather than sign retention.

C Experimental protocol

C.1 Populations, eligibility, and matching

The primary analyses use BoolQ, a passage-based yes/no question-answering benchmark (Clark et al., 2019), and OpenBookQA, a multiple-choice science question-answering benchmark (Mihaylov et al., 2018). The four primary model–dataset populations are Gemma-2 BoolQ, Gemma-4 BoolQ, Gemma-4 OpenBookQA, and Qwen3.6 OpenBookQA. Their final eligible population sizes are 200, 200, 144, and 150 items, respectively.

An item is eligible for a model if the answer parsed from its recorded clean trace matches the dataset's reference answer, the recorded trace has at least six words, and four perturbation sites can be constructed. Populations are the first 150 eligible items in candidate order, except BoolQ, for which the target is 200. Because eligibility is model-specific, the eligible sets can differ across models. The matched BoolQ comparison contains 142 shared items, and the matched OpenBookQA comparison contains 132 shared items. No additional post-match exclusions are applied to the OpenBookQA pair.

For Gemma-4, a generation-integrity criterion excludes traces not produced under the specified generation configuration. Six otherwise eligible OpenBookQA items fail this criterion and are excluded, leaving 144 eligible items. The same criterion excludes seven Gemma-4 ARC-Challenge items and three Gemma-4 GSM8K items in the broader analysis. It was applied before the present signed-versus-magnitude analysis. Qwen3.6 OpenBookQA required no analogous exclusion.

Outputs were not regenerated based on correctness, parser validity, or effect values. Technical retries were limited to execution failures or resuming interrupted runs, and generation-integrity exclusions were not replaced.

Analysis plan. The matched model pairs, endpoints, estimator, and bootstrap procedure were recorded before examination of the signed-versus-magnitude results. The analysis plan is included with the supplementary materials. The analysis was not externally preregistered, and the shortlist of comparisons was informed by earlier analyses using different endpoints.

Additional model–dataset populations provide descriptive breadth and boundary cases. Their retained scope is reported in Appendix G.

C.2 Recorded traces, perturbation sites, and operators

For non-GSM8K populations, the recorded trace is split on whitespace. Four target relative positions, 0.125, 0.375, 0.625, and 0.875 of the word sequence, define candidate locations. At each target, the nearest substitution-eligible word is selected. A substitution-eligible word is a non-stopword of at least three letters for which a same-type distractor is available. If no such word is available, the first non-empty word in that quarter is used. The stored normalized position can therefore differ from the target position.

GSM8K uses a different construction. Its traces are line-structured, and the four spans are contiguous, near-equal blocks of non-empty reasoning lines.

The recorded clean traces use dataset-specific decoding settings. For BoolQ, StrategyQA (Geva et al., 2021), ARC-Challenge (Clark et al., 2018), and OpenBookQA, the clean trace is generated with seeded sampling at temperature 0.7. GSM8K and AQuA-RAT (Ling et al., 2017) use greedy trace generation.

For non-GSM8K populations, masking replaces the selected word with `[MASK]`, deletion removes it, and typed substitution replaces it with one precomputed distractor of the same type. Type is a surface-form class—yes/no, true/false, number, single letter, capitalized word, or plain word—determined from the string rather than from part of speech. A suffix heuristic is used only to choose a fallback pool, and capitalization is preserved. For GSM8K, masking replaces the selected block with `[MASKED_REASONING]`, deletion removes the block, and typed substitution edits a number or arithmetic operator inside the block.

Worked operator example. For the OpenBookQA item used in Fig. 1 (`openbookqa_116`, Qwen3.6), the second perturbation site selects the word "classic" in the recorded trace: "…Stones weathering down to pebbles is a *classic* example of wind erosion/abrasion." The three operators modify this same selected word:

Operator	Perturbed rendering	X
Masking	"…is a `[MASK]` example of wind erosion/abrasion."	-0.20
Deletion	"…is a example of wind erosion/abrasion."	-0.30
Typed substitution	"…is a *specific* example of wind erosion/abrasion."	-0.28

Typed substitution replaces "classic" with the same-type distractor "specific", while deletion removes the selected word without repairing the surrounding text. All three records use the item's shared clean pool, for which $p_{\mathrm{ref}} = 35/50 = 0.70$. The displayed effects follow the paper's convention $X = p_{\mathrm{ref}}(\mathrm{clean}) - p_{\mathrm{ref}}(\mathrm{perturbed})$.

Applying the three operators to the four sites gives 12 perturbation records for a complete item.

For each item, one clean pool of 50 continuations is sampled once and reused across all 12 perturbation records. Each item–operator–site condition has its own pool of 50 perturbed continuations. The clean pool is therefore shared within item, whereas perturbed pools are record-specific.

Because each pool contains 50 continuations, p_{ref} and the resulting primitive effects are quantized in increments of $1/50$.

The same continuation pools are used for signed and magnitude-only scoring; no separate records are generated for the two scoring rules. Continuations are sampled at temperature 0.7, with top-$p = 0.9$, top-$k = 50$, and a 60-token continuation budget.

The supplementary materials provide the prompts, site-construction rules, operator implementations, seeds, and population information used in the experiments.

C.3 Task parsers

For BoolQ, the parser reads a yes/no answer from the task-specific answer position and otherwise assigns an invalid state. For OpenBookQA, the parser accepts the explicit answer pattern or a bare A–D label and uses the last-standalone-letter fallback before assigning an invalid state. Parsed answers are compared with the dataset's reference answer, not with the model's own clean-context answer.

Invalid parses remain in the denominator when estimating p_{ref}, as defined in Section 3.1. Parsing is therefore part of the measurement rather than a preprocessing step that removes observations. Appendix F reports parser, output-state, and valid-output sensitivity analyses.

C.4 Model revisions and reproducibility

Gemma-4-12B, Qwen3.6-35B-A3B, and OLMo-3-32B-Think were revision-pinned before generation and checked against the loaded checkpoint at run time. The Gemma-2 BoolQ run log did not record a revision string; its revision was recovered from the available cached model snapshot and matched to the same model revision used elsewhere in the study.

Table 7: **Revision status for models used in the primary analyses and broader checks.**

Model	Revision status
Gemma-2-9B-IT	recovered from cache
Gemma-4-12B	revision-pinned
Qwen3.6-35B-A3B	revision-pinned
OLMo-3-32B-Think	revision-pinned

The supplementary materials provide the prompts, parser definitions, generation and sampling settings, model-revision information, population information, analysis code, and result tables needed to reproduce or verify the reported analyses.

Using the reported seed and bootstrap configuration reproduces the primary confidence intervals.

D MATCHED MODEL-ORDERING RESULTS AND ROBUSTNESS

Table 8: **Population diagnostics and matched comparisons.** Block A reports effect-sign frequencies, average harmful and helpful contributions, and top-span disagreement for the four primary populations and Gemma-4 GSM8K. Block B reports the two matched model comparisons. OpenBookQA is the main comparison; BoolQ is supporting and termination-sensitive.

A. Population diagnostics

Population	$P(X>0)$	$P(X<0)$	$P(X=0)$	$E[H]$	$E[P]$	Top-span disagreement
Gemma-2 · BoolQ	0.122	0.626	0.253	0.023	0.383	0.490–0.545
Gemma-4 · BoolQ	0.169	0.214	0.617	0.051	0.069	0.135–0.165
Gemma-4 · OpenBookQA	0.083	0.625	0.292	0.013	0.261	0.396–0.444
Qwen3.6 · OpenBookQA	0.227	0.591	0.182	0.032	0.105	0.553–0.560
Gemma-4 · GSM8K	0.060	0.066	0.874	0.039	0.046	0.030[†]

B. Matched comparisons

Comparison	n	D_S	D_M	Δ_H	Δ_P	Status
Gemma-4 · Qwen3.6 · OpenBookQA	132	-0.178	$+0.143$	-0.0175	$+0.1607$	Main comparison; opposite ordering
Gemma-2 · Gemma-4 · BoolQ	142	-0.344	$+0.290$	-0.0267	$+0.3171$	Supporting; termination-sensitive

Top-span disagreement counts a disagreement only when $T_S \cap T_M = \varnothing$; all tied top spans are retained and any overlap counts as agreement. For the four primary populations, Block A reports the min–max range across masking, deletion, and typed substitution. Gemma-4 BoolQ shows lower disagreement than the other three primary populations. [†]For Gemma-4 GSM8K, 0.030 is the cross-operator mean of the three rates (0.029556), not an operator range. Here $H = \max(X, 0)$ and $P = \max(-X, 0)$.

D.1 OPENBOOKQA MAIN COMPARISON

For Gemma-4 minus Qwen3.6 on 132 matched OpenBookQA items,

$$D_S = -0.1781439394, \qquad D_M = +0.1431691919,$$

with

$$\Delta_H = -0.0174873737, \qquad \Delta_P = +0.1606565657.$$

The identities

$$D_S = \Delta_H - \Delta_P, \qquad D_M = \Delta_H + \Delta_P$$

hold to numerical precision.

Primary item bootstrap. The matched-item analysis resamples the 132 shared OpenBookQA items as complete statistical units. Using $B = 5{,}000$ replicates and seed 20260911, the same resampled item indices are used for both models and both scoring rules. The resulting estimates and 95% intervals are

$$D_S = -0.1781\ [-0.2383, -0.1169], \qquad D_M = +0.1432\ [0.0909, 0.1962].$$

The two-stage item-plus-continuation bootstrap gives

$$D_S:\ [-0.2412, -0.1151], \qquad D_M:\ [0.0877, 0.1947],$$

with all 5,000 replicates preserving the reversal. All 132 leave-one-item-out analyses also preserve the reversal.

Two-stage bootstrap implementation. The continuation-resampling analysis preserves the dependence structure of the experiment. Matched items are resampled jointly across the two models using one outer item-draw matrix. Within each replicate, each selected item's shared 50-sample clean

continuation pool is resampled once and the resulting clean count is reused across all 12 perturbation records for that item. Each record-specific 50-sample perturbed pool is resampled separately.

If the outer bootstrap selects the same item more than once, the implementation reuses that item's single within-replicate continuation resample with multiplicity weighting rather than generating a new inner resample for each duplicate occurrence. This preserves the shared-clean dependence structure.

Latest-position-site exclusion. Excluding the fourth perturbation site, selected closest to the end of the trace, and recomputing from the remaining three sites leaves the result essentially unchanged:

$$D_S = -0.17875\ [-0.2399, -0.1180], \qquad D_M = +0.14397\ [0.0924, 0.1961].$$

After exclusion,

$$\Delta_H = -0.0174, \qquad \Delta_P = +0.1614,$$

so the between-model difference in helpful contribution remains much larger than the difference in harmful contribution. This is a positional sensitivity check; it does not establish that all answer-bearing content has been removed.

D.2 BoolQ supporting comparison

For Gemma-2 minus Gemma-4 on 142 matched BoolQ items,

$$D_S = -0.3437558685, \qquad D_M = +0.2903521127,$$

with

$$\Delta_H = -0.0267018779, \qquad \Delta_P = +0.3170539906.$$

Latest-position-site exclusion. Excluding the fourth, latest-position perturbation site gives

$$D_S = -0.3482 \quad [-0.4332, -0.2670],$$

and

$$D_M = +0.2947 \quad [0.2266, 0.3645].$$

We treat BoolQ as supporting, termination-sensitive evidence. Valid-output and non-empty conditioning attenuate the result more strongly than in OpenBookQA, and the Gemma-2 helpful effects are largely associated with output termination (Appendix F).

D.3 Sensitivity to continuation temperature

We also vary the continuation-sampling temperature while keeping the matched items, reasoning traces, perturbations, parser, and other sampling settings fixed. Table 9 shows the matched model comparisons. For both datasets, signed and magnitude-only scoring continue to rank the models in opposite orders at every tested temperature.

Table 9: **Sensitivity of matched model ordering to continuation temperature.** The same matched item sets are used at every temperature. Intervals are 95% matched-item bootstrap intervals with $B = 5{,}000$ and seed 20260925.

Dataset	T	D_S	95% CI	D_M	95% CI
OpenBookQA	0.3	-.208	[-.285,-.133]	+.185	[.120,.253]
	0.7	-.178	[-.238,-.119]	+.143	[.092,.194]
	1.0	-.168	[-.219,-.117]	+.132	[.090,.177]
	1.3	-.138	[-.180,-.098]	+.107	[.073,.141]
BoolQ	0.3	-.425	[-.520,-.327]	+.356	[.272,.439]
	0.7	-.344	[-.421,-.264]	+.290	[.225,.355]
	1.0	-.272	[-.334,-.207]	+.219	[.169,.271]

All bootstrap intervals remain on opposite sides of zero, and all 5,000 bootstrap replicates preserve the reversal in every condition. The effect sizes decrease as temperature increases, so their magnitude is temperature-dependent even though the model-ordering conclusion remains the same.

E FULL SPAN-SELECTION RESULTS

Top-span disagreement retains all tied top spans and counts any overlap between T_S and T_M as agreement. Table 10 reports all 12 primary population-by-operator results and their 95% item-cluster bootstrap intervals. Each cell uses $B = 5{,}000$, and all 5,000 replicates are defined.

Table 10: **Full top-span disagreement results for the four primary populations.** Intervals are 95% item-cluster bootstrap intervals with $B = 5{,}000$; all 5,000 replicates are defined in every cell.

Population	Operator	Rate	CI low	CI high
Gemma-2 BoolQ	Masking	0.490	0.420	0.555
Gemma-2 BoolQ	Deletion	0.505	0.440	0.575
Gemma-2 BoolQ	Typed substitution	0.545	0.475	0.615
Gemma-4 BoolQ	Masking	0.145	0.095	0.195
Gemma-4 BoolQ	Deletion	0.165	0.115	0.220
Gemma-4 BoolQ	Typed substitution	0.135	0.090	0.185
Gemma-4 OpenBookQA	Masking	0.4375	0.3611	0.5208
Gemma-4 OpenBookQA	Deletion	0.4444	0.3681	0.5278
Gemma-4 OpenBookQA	Typed substitution	0.3958	0.3194	0.4792
Qwen3.6 OpenBookQA	Masking	0.5533	0.4733	0.6333
Qwen3.6 OpenBookQA	Deletion	0.5600	0.4800	0.6400
Qwen3.6 OpenBookQA	Typed substitution	0.5533	0.4733	0.6333

Top-span disagreement remains substantial under masking, deletion, and typed substitution in Gemma-2 BoolQ, Gemma-4 OpenBookQA, and Qwen3.6 OpenBookQA. Gemma-4 BoolQ shows lower disagreement.

As a positional sensitivity check, we remove the fourth, latest-position perturbation site. Top-span disagreement remains approximately 40–54% in Gemma-2 BoolQ, Gemma-4 OpenBookQA, and Qwen3.6 OpenBookQA. Gemma-4 BoolQ remains lower, at approximately 12.5–15.0%.

Additional diagnostics include unique-top disagreement, alternative tie summaries, rank-based comparisons, and output resampling. The primary span-selection endpoint remains top-span disagreement.

E.1 DISAGREEMENT BY EFFECT-SIGN PATTERN

Table 11 separates items for which all four span effects are helpful from items containing at least one harmful effect.

Table 11: **Top-span disagreement by effect-sign pattern.** "All helpful" means all four span effects satisfy $X < 0$. "At least one harmful" means at least one span satisfies $X > 0$.

Population	Operator	n	All helpful	At least one harmful	Overall disagreements	Disagreements: harmful stratum	Disagreements: all-helpful stratum
Gemma-2 BoolQ	Masking	200	118	46	98	10 (.22)	87 (.74)
Gemma-2 BoolQ	Deletion	200	108	46	101	8 (.17)	85 (.79)
Gemma-2 BoolQ	Typed substitution	200	113	52	109	11 (.21)	94 (.83)
Gemma-4 BoolQ	Masking	200	34	51	29	6 (.12)	20 (.59)
Gemma-4 BoolQ	Deletion	200	35	45	33	10 (.22)	20 (.57)
Gemma-4 BoolQ	Typed substitution	200	37	45	27	7 (.16)	20 (.54)
Gemma-4 OpenBookQA	Masking	144	86	17	63	3 (.18)	57 (.66)
Gemma-4 OpenBookQA	Deletion	144	86	16	64	3 (.19)	58 (.67)
Gemma-4 OpenBookQA	Typed substitution	144	85	17	57	1 (.06)	52 (.61)
Qwen3.6 OpenBookQA	Masking	150	83	42	83	6 (.14)	71 (.86)
Qwen3.6 OpenBookQA	Deletion	150	71	52	84	5 (.10)	68 (.96)
Qwen3.6 OpenBookQA	Typed substitution	150	73	51	83	4 (.08)	69 (.95)

Items with no harmful effect and at least one zero effect fall in neither sign-pattern stratum. The overall disagreement counts correspond to the reported population-by-operator values. This stratification describes the same endpoint rather than a separate one.

Across these cells, top-span disagreement is 54–96% on all-helpful items and 6–22% on items containing at least one harmful effect, as reported in §5.2.

F OUTPUT-STATE AND PARSER DIAGNOSTICS

For the analyses below, each sampled continuation is first assigned to one of three states: it gives the dataset's reference answer, it gives another valid but wrong answer, or it is invalid under the task parser.

F.1 BREAKDOWN OF HELPFUL-EFFECT CONTRIBUTIONS

On the matched OpenBookQA items, Gemma-4 has an average helpful contribution of 0.260997. Of this, approximately 0.137311 is associated with valid-but-wrong outputs and 0.123687 with invalid outputs. Qwen3.6 has an average helpful contribution of 0.100341, split into approximately 0.024444 and 0.075897, respectively.

The between-model difference is

$$\Delta_P \approx 0.160657,$$

with approximately 0.112866 associated with valid-but-wrong outputs and 0.047790 with invalid outputs.

Thus, 70.2% of the between-model helpful-contribution difference is associated with valid-but-wrong outputs and 29.8% with invalid outputs. Because the clean and perturbed continuation pools are unpaired, this is an aggregate breakdown rather than a transition analysis of individual continuations. This matched-item breakdown differs slightly from the full-population means reported in Table 3.

F.2 VALID-OUTPUT-CONDITIONED SENSITIVITY

As a secondary sensitivity estimand,

$$p_{\mathrm{ref|valid}} = \frac{n_{\mathrm{ref}}}{n_{\mathrm{ref}} + n_{\mathrm{wrong-valid}}}.$$

This gives

$$D_{S,\mathrm{valid}} = -0.150272, \qquad D_{M,\mathrm{valid}} = +0.109812.$$

Among defined two-stage replicates, the intervals are

$$[-0.21980, -0.08916]$$

and

$$[0.05628, 0.16606],$$

respectively, and all 616 of 616 defined replicates preserve the reversal.

Only 616 of 5,000 requested replicates are defined under the zero-denominator rule; 4,384 are undefined. The conditional estimate is defined for both models on 1,364 perturbation conditions spanning 115 of the 132 matched items. Because the estimand is undefined for many bootstrap replicates, we treat this as a supporting sensitivity analysis rather than primary inference.

F.3 OUTPUT TERMINATION

For this diagnostic, every stored generation in the primary records is classified more finely as reference answer, other valid answer, empty termination, or other invalid output. Empty termination means empty or whitespace-only generated text and is treated as an immediate end-of-turn.

For Gemma-2 BoolQ, empty terminations account for approximately 83–91% of the helpful gain in reference-answer support across operators, and 87% of helpful records include some empty-termination contribution. The helpful effects therefore largely reflect perturbations that make the model emit an answer instead of terminating.

For Gemma-4 BoolQ and Gemma-4 OpenBookQA, empty terminations are absent. They are essentially absent for Qwen3.6 OpenBookQA, appearing in only 2 of 1,800 perturbation records. Invalid outputs in these populations are otherwise non-empty unparsable text. This difference is why the BoolQ model comparison is treated as termination-sensitive.

G BOUNDARIES, QUALITY CONTROL, AND DESCRIPTIVE BREADTH

G.1 LOW-DIVERGENCE BOUNDARIES

Gemma-4 GSM8K is the clearest low-divergence case. Exactly 1,541 of 1,763 perturbation records have zero effect, or 87.4%, with

$$E[H] \approx 0.039, \qquad E[P] \approx 0.046.$$

Its top-span disagreement rates are approximately 0.0340, 0.0342, and 0.0204 across the three primary operators, giving a cross-operator mean of

$$0.029556 \approx 0.030.$$

The larger value near 0.074 belongs to a different unique-top-conditional endpoint and is not the top-span disagreement rate used in the paper. GSM8K also differs by design: it uses line-block spans and a block-level mask token rather than the single-word construction used for the other populations. Clean reference-answer support is already high in the relevant GSM8K populations, which leaves less room for perturbations to increase it. Its low divergence therefore should not be read as evidence that signed and magnitude-only scoring are generally equivalent.

StrategyQA provides a second, less extreme boundary. Most retained matched orderings are stable in the StrategyQA populations retained without a parser-validity flag, although some populations still show divergence. The Qwen3.6 StrategyQA population is flagged for low parser validity in the quality-control analysis below. These cases show that discarding direction does not matter equally in every population.

G.2 POST-HOC PARSER-VALIDITY QUALITY CONTROL

A post-hoc, rule-based parser-validity check examined the descriptive grid. Qwen3.6 BoolQ had very low parser validity: only 4.84% of its clean outputs and 9.56% of its perturbed outputs were parsed as valid task answers under the parser.

This was the only population removed by the quality-control rule. The exclusion weakens rather than strengthens the descriptive evidence for signed-versus-magnitude divergence because Qwen3.6 BoolQ had shown substantial disagreement between the two scoring rules and was involved in three ordering reversals whose confidence intervals excluded zero in opposite directions.

The quality-control rule was applied post hoc from parser-validity diagnostics. Because the threshold was not chosen independently of those diagnostics, we report threshold sensitivity below. No population was regenerated or reparsed as part of this quality-control analysis.

Applying either a 5% or 10% clean-validity cutoff excludes the same single population. The next-lowest clean-validity rates are approximately 12.1% for Qwen3.6 AQuA-RAT, 16.5% for Gemma-4 AQuA-RAT, and 19.7% for Qwen3.6 StrategyQA. These populations remain in the descriptive grid but are flagged for their parser-validity limitations.

The complete quality-control diagnostics and retained-scope inventory are included in the supplementary materials.

G.3 POST-QC DESCRIPTIVE BREADTH

The broader analysis spans six datasets: StrategyQA (Geva et al., 2021), BoolQ (Clark et al., 2019), ARC-Challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), GSM8K (Cobbe et al., 2021), and AQuA-RAT (Ling et al., 2017).

StrategyQA contains yes/no questions requiring implicit reasoning, ARC-Challenge is multiple-choice science QA, GSM8K contains grade-school math word problems, and AQuA-RAT contains multiple-choice algebraic word problems.

After the Qwen3.6 BoolQ exclusion, the retained descriptive grid contains 42 model–dataset populations and 78,853 perturbation records across nine model identifiers, five model families, and six datasets.

Across the 47 retained matched comparisons,

$$36/47 = 76.6\%$$

change ordering between signed and magnitude-only scoring. In 23 comparisons, the 95% intervals for the two model differences exclude zero in opposite directions.

These comparisons share models, items, datasets, and experimental design choices. They are therefore descriptive, dependence-structured evidence rather than independent replications.

The complete 42-population retained inventory, together with the excluded Qwen3.6 BoolQ row, is included in the supplementary scope table.

H External reanalysis of Du et al.

H.1 Source and native estimand

We reanalyzed the public `kducohere/MC-Math-Rollouts` data released by Du et al. (2026). The exact release version used in the analysis is listed in the supplementary materials.

The analysis uses the released `reasoning_prefixes_final_answer_dist_analysis.csv` files and the signed field `is_correct_proportion_diff`. This field corresponds to the successive-prefix correctness effect

$$A_t = \widehat{P}(\text{correct} \mid s_t) - \widehat{P}(\text{correct} \mid s_{t-1}),$$

equivalently the rollout-based Monte-Carlo step advantage of Du et al. (2026) under a correctness reward.

We keep the released sign orientation and compare A_t with $|A_t|$ without rescaling, renormalization, or sign reversal. The original method uses rollout-based advantages for reasoning-step importance, with self-advantage as its default formulation. The signed-versus-magnitude top-step comparison below is our reanalysis and is not an analysis reported by Du et al.

H.2 Eligibility, endpoint, and statistical unit

The eligibility rules cover 7,200 candidate trajectory files. Of these, 6,127 are present in the official release. The remaining 1,073 files are absent from the release and were not excluded by our analysis.

All 6,127 available trajectories are usable under the trajectory-level rules, comprising 531,035 eligible step records. We impose no additional eligibility threshold on the released `sample_n` values; the finite-rollout limitation is discussed in Section H.5.

For each trajectory,

$$T_S = \arg\max_t A_t, \qquad T_M = \arg\max_t |A_t|.$$

We count *top-step disagreement* when

$$T_S \cap T_M = \varnothing.$$

A disagreement is counted only when the two top-step sets have no common member; all tied top steps are retained, so ties favor agreement.

Bootstrap resampling is clustered by benchmark–prompt problem, giving 180 clusters.

H.3 Main result and robustness

Top-step disagreement occurs in

$$\frac{2{,}247}{6{,}127} = 0.366737 \approx 36.7\%$$

of eligible trajectories.

A 5,000-replicate cluster bootstrap gives

$$95\% CI = [0.333494, 0.401179],$$

reported as 33.3–40.1%. All 5,000 replicates are defined.

Leave-one-benchmark-out estimates range from 35.3% to 40.3%, while leave-one-configuration-out estimates range from 27.1% to 40.5%. A secondary original-answer endpoint gives

$$1,123/6,127 = 18.3\%$$

top-step disagreement.

No external matched model-ordering analysis was performed. This reanalysis tests only the step-selection consequence of retaining versus discarding sign under a different released effect construction.

H.4 Floating-point parser correction

The initial implementation used a default floating-point CSV parser that merged a small number of distinct released decimal values and thereby created unintended exact ties.

Re-parsing under the exact-equality rule changed six of 6,127 trajectory classifications, all from agreement to top-step disagreement and none in the opposite direction. The corrected estimate increased from 36.58% to 36.67%. All results reported in this paper use the corrected value.

Because tied top steps count as agreement, the original floating-point merging enlarged top sets and worked against finding top-step disagreement.

H.5 Finite-rollout sampling limitation

The released Du et al. files provide Monte-Carlo correctness proportions and a per-prefix `sample_n`, but not the underlying Bernoulli outcomes for individual rollouts. We therefore treat the released proportions as fixed quantities in the external reanalysis and apply no minimum `sample_n` filter.

Most prefix estimates are based on 50 rollouts: 95.7% of the eligible step records have `sample_n`=50. A small low-sample tail remains: 4.3% use fewer than 50 rollouts, including 2.5% with fewer than 49 and some values as low as one. We apply no minimum-`sample_n` filter.

The reported cluster-bootstrap interval therefore captures variation across the 180 benchmark–prompt problem clusters but does not account for Monte-Carlo variation inside the released per-prefix correctness proportions. This limitation concerns the external reanalysis only; our completed-trace experiments use stored continuation pools and separate continuation-resampling analyses.

I Detailed limitations and reporting checklist

For perturbation-based CoT evaluation, we recommend reporting at least the following:

1. the primitive behavioral effect and its sign convention;
2. whether direction is retained, discarded, or otherwise transformed before aggregation;
3. the intervention operator and reasoning-site construction;
4. the clean and perturbed continuation-sampling configuration;
5. the parser, reference-answer definition, and invalid-output handling;
6. the statistical unit, aggregation order, and matched-item rule;
7. how ties are handled for span-level selection;
8. the resampling procedure and whether it accounts for variation from the sampled continuations;
9. the evidence status of each analysis: primary comparison, supporting comparison, follow-up robustness check, supporting sensitivity, descriptive breadth, or external evidence under a different estimand;
10. construct boundaries: behavioral perturbation response should not be equated with semantic necessity or mechanistic reliance without additional evidence.

No universally preferred score. The paper does not identify a universally preferred scoring rule. Signed scoring is appropriate when whether a perturbation helps or harms reference-answer support is substantively meaningful. Magnitude-only scoring is appropriate when the question concerns change size regardless of direction. When both interpretations matter, both should be reported rather than treated as interchangeable.

Dependence and generalization. The 42 retained model–dataset populations and 47 matched comparisons are not independent replications. They share models, datasets, items, generation procedures, and analysis design. The broader results therefore show descriptive breadth within this study and should not be interpreted as an estimate of how common reversals are across reasoning systems generally.

Model-specific eligibility. Eligibility is model-specific, so analyzed populations for different models need not contain the same items. The two matched cross-model analyses address this by restricting comparison to shared usable items. The reported model-ordering differences should not be interpreted as causal effects of model scale, architecture, or training recipe.

Parser and output-state dependence. The primitive effect depends on the parser and on how invalid outputs are handled. Invalid parses remain in the denominator of p_{ref}. The output-state breakdown shows that both valid-but-wrong and invalid outputs contribute to the helpful-contribution difference in the main OpenBookQA comparison. The valid-output-conditioned analysis preserves the reversal on its defined support but has many undefined bootstrap replicates, so it remains supporting rather than primary evidence.

Sampling, termination, and quality control. The 50-sample clean and perturbed continuation pools quantize the primitive effect in steps of $1/50$. The two-stage bootstrap preserves the OpenBookQA reversal, but the valid-output-conditioned analysis is defined on a restricted subset of items. BoolQ is sensitive to immediate output termination, as detailed in Appendix F. The broader descriptive analysis also includes a post-hoc parser-validity exclusion. That exclusion removed evidence favoring the broader divergence pattern, while three low-validity populations were retained and flagged (Appendix G). In the Du et al. release, a small low-sample tail remains and no minimum sample-count filter was applied (Appendix H).

Boundary populations and headroom. Low divergence can arise when effects are sparse or when clean reference-answer support leaves little room for movement. Gemma-4 GSM8K is the clearest example: most perturbation records have zero effect and clean support is already high. GSM8K also differs in intervention design because it uses line-block spans and a block-level mask token. These features limit direct generalization from GSM8K to the single-word perturbation populations.

External reanalysis. The Du et al. analysis provides external evidence for the step-selection consequence of sign treatment under a different reasoning-effect construction. It does not replicate the completed-trace perturbation experiment or the matched model-ordering result. Its bootstrap treats the released Monte-Carlo proportions as fixed and therefore does not account for finite-rollout variation in those released estimates.

Behavioral rather than mechanistic interpretation. The study measures how sampled answer behavior changes after perturbing a recorded reasoning trace. The results do not by themselves establish that a selected span was internally necessary for the original answer, that the recorded reasoning is mechanistically causal, or that one model is intrinsically more faithful than another.