PLANTED EFFECTS GIVE DETECTION FLOORS FOR REASONING-TRACE MEASUREMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Studies that intervene on a model's reasoning trace report how strongly its answer depends on the trace, but rarely the smallest effect their design could detect. Without that detection floor, a reader cannot tell a small reported change from item-sampling noise. We measure the floor for the causal importance of reasoning (CIR) by planting effects of known size in an open model's own reasoning. The answer must repeat a code word shown at the end of the reasoning, and switching that word on a chosen share of items plants the effect. The spread of the planted effects across items gives the floor. It shrinks with the square root of the item count, which also tells a new study how many items a planned effect needs. In a published reinforcement-learning study, and assuming item variance carries over from our calibration model, 34 of 36 task effects are resolvable. The exceptions are the study's smallest effects, including a near-zero change its design could not have resolved. Studies that report CIR effects should report this floor with them, and planting effects of known size in the studied model gives it.

1 INTRODUCTION

Interventions on the reasoning trace that a language model writes before answering (Wei et al., 2022; Kojima et al., 2022) are the standard evidence for how much the answer relies on that trace. Truncating, corrupting, resampling, or masking the trace and measuring the change in the answer gives a dependence score (Lanham et al., 2023; Turpin et al., 2023; Macar et al., 2026; Jia et al., 2026). Reinforcement learning with verifiable rewards (RLVR) scores only the final answer, which leaves open how training changes the role of the trace (Guo et al., 2025; Yu et al., 2025; Liu et al., 2025). Yu et al. (2026) use a dependence score, the causal importance of reasoning (CIR), to report this change for 40 procedurally generated tasks. On one task, Number format, the plotted values before and after training nearly coincide, and a reader could take the change as evidence that training left the answer's dependence on its trace unchanged.

Whether that reading holds depends on the design's *detection floor*. We define it as the smallest effect whose estimate would be distinguishable from zero at the design's item count. Standard advice for evaluation in natural language processing, reinforcement learning, and language-model benchmarks obtains such floors from significance tests and power analysis (Dror et al., 2018; Card et al., 2020; Henderson et al., 2018; Agarwal et al., 2021; Madaan et al., 2024; Miller, 2024). Both need the variance of the score, taken from a noise model such as the binomial variance of an accuracy or from reported replicates. CIR averages Jensen–Shannon (JS) divergences between answer probabilities, and its item variance depends on the model, the task, and the effect itself, with no formula to supply it. Of 11 trace-intervention studies we screened, the only one that reports task-level CIR effects gives a single training run and no sampling intervals (Appendix A). Existing controls target other confounds. Restarting generation from a trace prefix with the same remaining token budget separates problem difficulty from trace content (Bulut, 2026), and frozen copies of a model show how much apparent self-improvement arises when nothing has changed (Xu et al., 2026). These controls test whether an intervention isolates the intended counterfactual, and neither gives the smallest CIR effect that a published item count can resolve.

We measure the floor by planting effects of known size in the reasoning of an open calibration model, Qwen3-1.7B on Grade School Math 8K (GSM8K) (Yang et al., 2025; Cobbe et al., 2021). Like Yu et al. (2026), we score CIR on exact prefixes, literal initial segments of the trace with no

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.

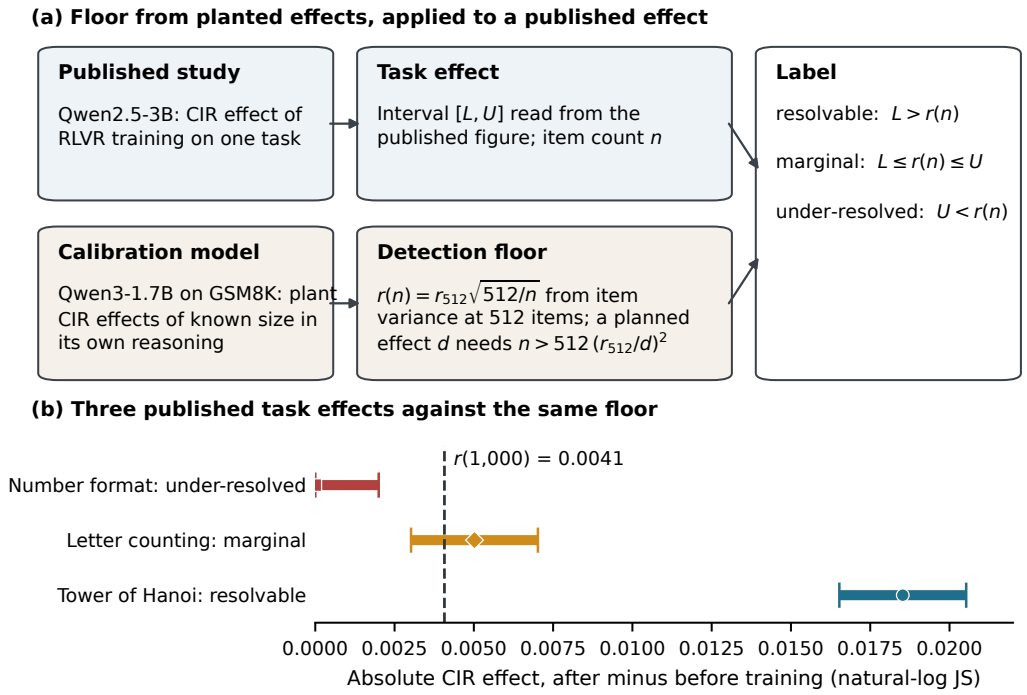


Figure 1: A published effect is resolvable only when its interval clears the floor that planted effects give at its item count. (a) CIR effects of known size planted in a calibration model's reasoning give the item variance, hence the floor $r(n)$ at n items, which the plot-reading interval $[L, U]$ of a published effect must clear. (b) Three published effects at 1,000 items receive different labels.

regenerated continuation. A nonce is an arbitrary code word. The prompt asks the model to copy the nonce shown at the end of its reasoning into its answer, and CIR scores the probability of that nonce. Showing the target nonce on some items and a different one on others changes how much the answer depends on the trace, and the share of items switched sets the planted CIR effect. The spread of the planted effects across items gives the half-width of a 95% interval for the mean effect. That half-width shrinks with the square root of the item count. One measurement therefore gives the floor at any item count, and inverting it gives the item count that a planned effect needs. Applying the floor to another model assumes that item variance carries over from the calibration model.

We audit the task effects that Yu et al. (2026) report for Qwen2.5-3B on ReasoningGym (Qwen Team, 2024; Stojanovski et al., 2025), and the floor resolves most effects but not the two smallest (Figure 1). Of the 36 effects whose size and item count we could recover, 34 lie wholly above their floors and are *resolvable*. Number format's near-zero change lies below its floor, and a change of that size was beyond what the study's item count could resolve. Letter counting's interval straddles its floor. Because the planted effects shape the item variance, we also recompute the floor from natural traces, unmodified continuations that reach correct and incorrect answers, and only these two labels move. Our contributions are:

- We measure a detection floor for exact-prefix CIR from planted effects of known size and invert it into an item-count rule (Section 3).
- We show that 34 of 36 classifiable task effects in a published RLVR study are resolvable, and its two smallest are not (Section 4).
- We find that natural-trace item variance on GSM8K and on MATH changes at most the labels of the two smallest effects (Section 4.3).

2 CIR AND THE AUDITED STUDY

CIR measures how much an answer's probability depends on the reasoning that precedes it (Yu et al., 2026). Let q be a question, t a sampled reasoning trace, and y the scored answer. Write $t_{1:k}$ for the first k units of t, which are whitespace-delimited words in the audited study and tokens in our calibration. For a set $\mathcal{K}$ of prefix positions,

$$\mathrm{CIR}(q,t,y) = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} \mathrm{JS}\big(\mathrm{Bern}\big(p(y \mid q, t_{1:k})\big) \,\big\|\, \mathrm{Bern}\big(p(y \mid q, t)\big)\big)\,, \tag{1}$$

where $p(y \mid \cdot)$ is the probability of the answer's full token sequence and $\mathrm{Bern}(p)$ is a Bernoulli distribution. With natural logarithms in JS, CIR lies in $[0, \ln 2]$, and the main text reports CIR values in these units (Appendix B uses CIR divided by $\ln 2$). A value near zero means that truncating the trace barely changes the answer probability.

The audited study reports CIR from a single sampled response per item and a single training run. It scores the joint probability of the model's own answer at prefix positions spaced evenly from the first prefix to the full trace (Table 1). Its Figure 2 plots task-mean CIR before and after RLVR training for each task. We define a task's effect as the after-minus-before difference in task-mean CIR.

3 A DETECTION FLOOR FROM PLANTED EFFECTS

A detection floor needs the item variance of CIR under a known effect. Because the audited study releases no item-level CIR values, we plant effects of known size in a calibration model's own reasoning and measure that variance there.

3.1 PLANTING EFFECTS OF KNOWN SIZE

Planting controls how much the scored answer depends on the last line of the model's own reasoning. The calibration model runs with its built-in thinking mode. It is a small open model with native reasoning traces and checkable answers, near the audited model's scale. Each calibration trace keeps the start of the model's own reasoning and ends with one extra line that shows one of four nonces. Table 1 gives the lengths, and an example reads

Trace	first 89 reasoning tokens, then `Verification nonce: RUBY.`
Scored answer	`Final: <answer>; Check: RUBY`

We score the probability of the target nonce after the correct final answer, which makes the target nonce the scored answer y in Equation 1. In the target condition the last line of the trace shows the target, and in the alternative condition it shows a different nonce. Only prefixes that reach the last line reveal the nonce. CIR therefore measures how much the probability of the target depends on that line.

We plant effects in two directions and pair each item with its own baseline. When we add the target, the baseline is the alternative condition, and when we remove it, the baseline is the target condition. Each item either switches to the other condition or keeps its baseline. Its paired effect h_j is its CIR in the condition it ends in minus its CIR at the baseline, and a kept item has $h_j = 0$.

The fraction of switched items sets the mean effect. A fixed hash of the item identifier and the design settings chooses the switched items (Appendix B), and we tune the fraction on separate items. We plant five sizes, each twice the previous one, in both directions. Each of three sampling seeds of the same checkpoint draws its evaluation questions from the same pool, and different seeds can draw the same question (Table 1). The planted effect is a synthetic switch on a subset of items, and Table 1 lists how it differs from the training-induced effects of the audited study.

3.2 FLOOR AT ANY ITEM COUNT

The variance of the planted item effects gives the floor at the calibration's item count. For seed s, let h_{sj} be the paired effect on item j, in natural-log JS, and let $\mathrm{Var}_j(h_{sj})$ be its variance across the

Property	Calibration	Audited study
Model	Qwen3-1.7B	Qwen2.5-3B
Task	GSM8K	ReasoningGym tasks
Sampling temperature	0.6	1.0
Trace	89 reasoning tokens and a nonce line	full trace
Prefix positions	all 96 depths	11 evenly spaced
Scored answer	target nonce	model's own answer
Evaluation items	512 per seed, three seeds	1,000 per task (one task 100)
Compared conditions	same reasoning tokens	traces from two checkpoints
Pairing by item	yes	not reported

Table 1: The calibration shares the CIR formula and units with the audited effects but differs in model, task, and trace construction.

items of seed s. The floor is

$$r_{512} = c_\alpha \sqrt{\frac{1}{3^2} \sum_{s=1}^{3} \frac{\mathrm{Var}_j(h_{sj})}{512}}, \qquad c_\alpha = t_2^{-1}(1-\alpha/2), \quad \alpha = 5\%. \tag{2}$$

Here t_2^{-1} is the quantile function of Student's t distribution with two degrees of freedom. The expression is the half-width of a 95% interval for the mean of three seed means. It treats the seeds as independent and does not estimate the covariance between seeds that share questions.

An estimate smaller than this half-width cannot be distinguished from zero, and r_{512} is therefore the detection floor at the calibration's item count. The t quantile widens the interval relative to a normal quantile, which suits the small number of seeds, and Table 2 also gives the count under a normal quantile.

A square-root law carries the floor to any item count. Because the variance of a mean falls as $1/n$, the floor for three seeds of n items each is

$$r(n) = r_{512}\sqrt{512/n}. \tag{3}$$

Inverting Equation 3 gives the item count at which a planned effect of magnitude d clears the floor, $n > 512\,(r_{512}/d)^2$. An effect exactly at the floor would be detected about half the time, and planning for 80% power raises the floor by the factor in Table 2.

The floor grows by more than threefold from the smallest planted size to the largest (Figure 2a). With a fraction f of items switched, each by a similar amount δ, the item variance is close to $f(1-f)\delta^2$, which indicates that the planting construction shapes much of the floor. As the planted effect approaches zero, fewer items switch and the floor approaches zero too, because unswitched items contribute exactly zero. Section 4.3 therefore compares it with item variance from natural traces.

We use the smallest planted floor, $r_{512} = 0.0057$, and write r for the floor $r(n_i)$ it gives at each task's item count. Other floors appear as multiples of r at equal item count. A larger floor can only reduce the number of resolvable effects. Among the planted floors, this choice is therefore the most favourable to the published effects. Most published effects, 23 of 36, exceed the largest planted effect (Appendix Table 7), and r may understate their uncertainty.

Variation between sampling seeds adds nothing measurable to the planted floor. The between-seed half-width, the t half-width from the spread of the three seed means, is at most 0.0001, far below r_{512}. The small value partly reflects the design, in which all seeds switch the same fraction of their items. Adding the excess between-seed variance, the variance of the seed means beyond what item variation explains, leaves the floor at r, because that excess is zero for planted CIR. For attention blocking and resampled continuation, two other dependence scores measured in the same calibration, between-seed variation can exceed item variation, and their floors would need a between-seed term (Appendix B).

Floor ($\times r$)	Variance source or definition	Resolvable	Marginal	Under-resolved
0.25	Fixed multiple	35	1	0
0.46	Normal quantile, three-seed mean	35	1	0
0.47	Natural traces, GSM8K	35	1	0
0.50	Fixed multiple	35	0	1
0.79	Normal half-width, one evaluation	34	1	1
1.00	Planted effects, item only	34	1	1
1.00	Planted effects, item plus seed	34	1	1
1.25	Detectable at 80% power	34	1	1
1.63	Natural traces, MATH	34	1	1
2.00	Fixed multiple	34	0	2
4.00	Fixed multiple	30	1	5
4.87	Attention-blocking seed spread	28	2	6

Table 2: Floors up to twice r leave 34 or 35 of the 36 task effects resolvable. Each row applies the labelling rule with the floor replaced by a multiple of r at equal item count. The item-plus-seed row adds the excess between-seed variance, which is zero here. The 80%-power row adds the 80% quantile of t with 2 degrees of freedom to c_α. The last row uses the largest between-seed half-width measured for attention blocking (Appendix B).

3.3 Applying the floor to published effects

The other screened studies measure different quantities, such as agreement with early answers, divergence under resampled continuations, attention masking, or accuracy under corruption (Appendix Table 4). A CIR floor does not transfer to them.

We read each effect from the drawing commands of the audited study's figure, which is a vector graphic. Each marker receives a reading tolerance of ± 0.0010, wider than the precision of the drawing coordinates, and Appendix A checks the extraction against the study's own counts. The after-minus-before difference therefore carries a tolerance of ± 0.0020. A signed interval $[a, b]$ becomes a magnitude interval $[L, U]$, equal to $[0, \max(|a|, |b|)]$ when it contains zero and to $[\min(|a|, |b|), \max(|a|, |b|)]$ otherwise. The interval covers plot-reading error only, and the label ignores the sign of an effect. We classify the 36 task results that have both markers and a positive item count (Appendix A).

We label each published effect by where its interval lies relative to its floor. Task i with n_i items has floor $r_i = r(n_i)$, and we call its effect resolvable when $L_i > r_i$, marginal when $L_i \le r_i \le U_i$, and under-resolved when $U_i < r_i$. Because r_i has the form of a 95% half-width, a resolvable label says that an estimate of the published size would be distinguishable from zero at the item variance of the calibration model. The floor r_i assumes three seeds of n_i items. The normal half-width of a single evaluation, the audited study's design, and the floor for detection at 80% power bracket r_i and give the same resolvable count (Table 2).

For example, Letter counting has $n_i = 1{,}000$ items, which gives

$$r_i = 0.0057\sqrt{512/1000} = 0.0041.$$

Its magnitude interval $[0.0030, 0.0070]$ contains that floor, and its label is marginal.

4 Results

4.1 Resolution of the published effects

If item variance carries over from the calibration model, 34 of the 36 classifiable task effects are resolvable at their item counts (Table 2). The other two, one marginal and one under-resolved, are the boundary cases of Section 4.2.

Most resolvable effects clear their floors by a wide margin. In Figure 2b we divide each interval by its floor, which places all task effects on one axis. In 25 of the 36, the lower end L_i exceeds ten times r_i. The resolvable effects include both increases and decreases in CIR (Appendix Table 7).

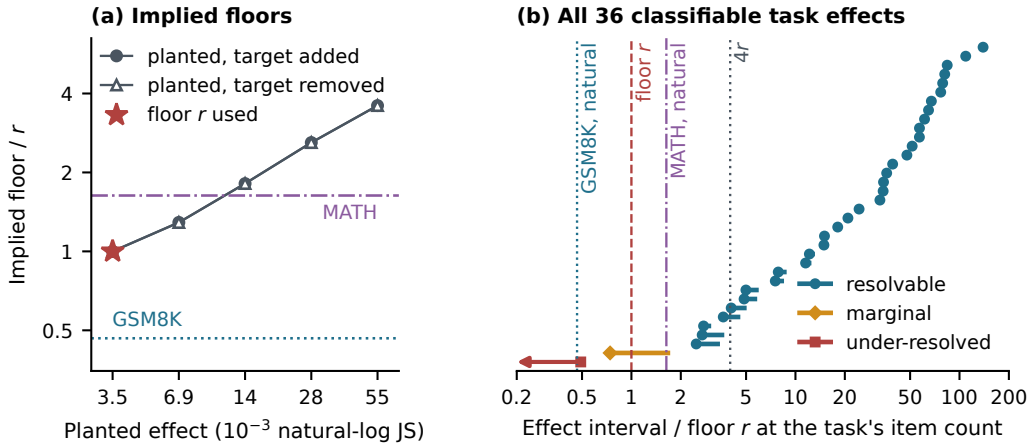


Figure 2: Natural-trace item variance changes only the labels of the two smallest published effects. (a) Floors implied by planted effects of increasing size and by natural traces on GSM8K and MATH (horizontal lines), as multiples of the floor r used in the audit, at equal item count. (b) Each bar is a task effect's magnitude interval divided by its floor, sorted by the lower end. An effect is resolvable under a floor m times r when its bar starts right of m. The under-resolved interval starts at zero, and its arrow runs off the log axis.

4.2 Boundary cases

Number format, the smallest published effect, lies below its floor (Table 3). Its magnitude is at most 0.0020, against a floor of 0.0041. The near-zero change of Section 1 is smaller than the design could resolve at 1,000 items. It cannot show that training left CIR unchanged. Letter counting's positive interval straddles its floor. Tower of Hanoi, shown for contrast, lies wholly above it.

Task	Items	Signed effect interval	Floor	Label	Natural-trace label
Number format	1,000	$[-0.0020, +0.0020]$	0.0041	under-resolved	marginal
Letter counting	1,000	$[+0.0030, +0.0070]$	0.0041	marginal	resolvable
Tower of Hanoi	1,000	$[+0.0165, +0.0205]$	0.0041	resolvable	resolvable

Table 3: Three effects at the same item count receive all three labels. The last column uses item variance from natural traces.

4.3 Sensitivity to the variance source

The effects whose lower end exceeds ten times their floors stay resolvable under all variance sources in Table 2, and the labels that change belong to effects closer to their floors (Figure 2).

Natural traces give a smaller floor. For each GSM8K question, we split the model's own continuations from a shared prefix by whether they reach the correct answer. The paired effect is the CIR of the correct answer on a correct-outcome trace minus its CIR on an incorrect-outcome trace, averaged over two such pairs. Without a planted effect, this measurement has no known effect to check it against. We obtain an implied floor of $0.47\,r$ at equal item count. With the labelling rule otherwise unchanged, Number format becomes marginal and Letter counting resolvable, and 35 task effects remain resolvable.

The natural-trace floor depends on the task. On the MATH benchmark with the same model (Hendrycks et al., 2021), scoring the 128 questions common to all seeds, the implied floor is $1.63\,r$ at equal item count, and the counts match those under r. Moving from GSM8K to MATH raised the natural-trace floor more than threefold. The audited setting changes both the model and the tasks, which suggests that a shift of this size is plausible there as well.

The resolvable count holds at 34 until the floor reaches about two and a half times r, where String synthesis, the resolvable effect with the smallest ratio of lower end to floor, becomes marginal (Appendix Table 7). Larger multiples, up to four times r and above the floor from any planted size, remove further effects in order of that ratio (Table 2).

5 Related work

Measuring trace dependence. Early answering, trace corruption, and paraphrase test whether answers depend on the stated reasoning (Lanham et al., 2023; Turpin et al., 2023). Later scores resample continuations (Macar et al., 2026), track information flow with attention masks (Jia et al., 2026), locate the point after which reasoning no longer changes the answer (Scalena et al., 2026), and measure recovery from injected perturbations (von Recum et al., 2026). Corruption format, classifier choice, and task accuracy can each change measured faithfulness (Garcia, 2026; Young, 2026; Bentham et al., 2024; Arcuschin et al., 2026; Paul et al., 2024). Faithful-by-construction reasoning, verifier benchmarks, trace monitorability, and legibility are related targets with their own measures (Lyu et al., 2023; Jacovi et al., 2024; Chen et al., 2025; Korbak et al., 2025; Kirchner et al., 2024). CIR, the score of the audited study, belongs to the dependence family (Yu et al., 2026).

Controls for trace-intervention claims. Matched-budget restarts show that apparent trace effects can reflect problem difficulty and continuation budget (Bulut, 2026). Comparing a model with frozen copies of itself measures how much apparent self-improvement arises when nothing has changed (Xu et al., 2026).

Statistical power in machine-learning evaluation. Card et al. (2020) find that many comparisons in natural language processing are underpowered, and Dror et al. (2018) set out significance testing for the field. In deep reinforcement learning, seed variation can reverse conclusions (Henderson et al., 2018), and Agarwal et al. (2021) recommend interval estimates over few runs. Seed and data-sampling variance has also been measured for machine-learning benchmarks (Bouthillier et al., 2021) and language-model evaluations (Madaan et al., 2024; Miller, 2024). In other sciences, low power inflates reported magnitudes and undermines replication (Button et al., 2013; Gelman & Carlin, 2014; Open Science Collaboration, 2015), and sample-size justification is part of study design (Lakens, 2022). These methods take the variance of a score from a noise model or from reported replicates. For CIR, which has neither, we obtain it from planted effects.

6 Limitations

We audit one published figure of exact-prefix CIR effects, and four of its rows lack a required input, which leaves the resolvable count for the whole figure between 34 and 38 of 40. No literature-wide rate follows from one figure. All variance measurements come from Qwen3-1.7B, a different model from the audited Qwen2.5-3B, and the calibration differs from the audited score in trace length, prefix grid, sampling temperature, and scored answer (Table 1). Without variance measured on the audited model, the resolvable count is neither an upper nor a lower bound, and Table 2 shows how it moves with the floor. The floor treats each published effect as a paired item-level difference. The audited study does not report its pairing, and an unpaired comparison would have a larger floor. The labels grade resolution alone. Whether the traces reflect the model's computation, whether the reward caused the changes, and whether the audited study's conclusions hold lie outside what they test.

7 Conclusion

A trace-intervention effect should be read against the smallest effect its design could resolve, and planting effects of known size in a model's own reasoning gives that floor at any item count. For the audited RLVR study, most task effects lie far above the floors we tested. The labels that change under other variance sources belong to effects close to their floors. The floor complements restart and frozen-copy controls, which test what an effect reflects. A trace-intervention study lets readers check its resolution when it reports

- the effect with its sampling interval, and whether the before and after estimates are paired by item,
- the scoring details, including the prefix positions and the scored answer,
- the item count, with item variance measured on the same model and task, and
- for claims about training, the variation between independently trained runs.

With that item variance, the floor for three seeds of n items is $r_{512}\sqrt{512/n}$, and a planned effect of magnitude d needs more than $512\,(r_{512}/d)^2$ items.

AI USE STATEMENT

AI assistants based on large language models were used for part of the execution of this work, under the authors' direction: implementing the calibration and analysis code, and extracting and reformatting the published figure data. All research decisions were made by the authors, who verified every AI-assisted output and every reported result. The authors take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

Section 3 and the appendix specify the inclusion rule, digitization, calibration design, floor calculation, classifier, model and task settings, and every classified value. The anonymized code supplement contains the audit inputs and the analysis that regenerates the classifications, tables, and figures.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, et al. Chain-of-thought reasoning in the wild is not always faithful, 2026.

Oliver Bentham, Nathan Stringham, and Ana Marasović. Chain-of-thought unfaithfulness as disguised accuracy, 2024.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, 2021.

Yigit Utku Bulut. It's the problem, not the path: Budget and difficulty confounds in LLM reasoning trajectories, 2026.

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013. doi: 10.1038/nrn3475.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9263–9274, 2020. doi: 10.18653/v1/2020.emnlp-main.745.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, et al. Reasoning models don't always say what they think, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems, 2021.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1383–1392, 2018. doi: 10.18653/v1/P18-1128.

Gabriel Garcia. The last word often wins: A format confound in chain-of-thought corruption studies, 2026.

Andrew Gelman and John Carlin. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014. doi: 10.1177/1745691614551642.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11694.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. Measuring mathematical problem solving with the MATH dataset, 2021.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, et al. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains, 2024.

Jinghan Jia, Joe Benton, and Eric Easley. Faithfulness as information flow: Evaluating and training faithful chain-of-thought reasoning, 2026.

Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-verifier games improve legibility of LLM outputs, 2024.

Takeshi Kojima, Shixiang Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, 2022.

Tomek Korbak, Mikita Balesni, Elizabeth A. Barnes, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety, 2025.

Daniël Lakens. Sample size justification. *Collabra Psychology*, 8(1), 2022. doi: 10.1525/collabra.33267.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, et al. Measuring faithfulness in chain-of-thought reasoning. In *International Conference on Machine Learning*, 2023.

Zichen Liu, Changyu Chen, Wenjun Li, et al. Understanding R1-Zero-like training: A critical perspective, 2025.

Qing Lyu, Shreya Havaldar, Adam Stein, et al. Faithful chain-of-thought reasoning, 2023.

Uzay Macar, Paul C. Bogdan, Senthooran Rajamanoharan, and Neel Nanda. Thought branches: Interpreting LLM reasoning requires resampling, 2026.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, et al. Quantifying variance in evaluation benchmarks, 2024.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349 (6251):aac4716, 2015. doi: 10.1126/science.aac4716.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning, 2024.

Qwen Team. Qwen2.5 technical report, 2024.

Daniel Scalena, Sara Candussio, Luca Bortolussi, Elisabetta Fersini, Malvina Nissim, and Gabriele Sarti. Beyond the commitment boundary: Probing epiphenomenal chain-of-thought in large reasoning models, 2026.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, et al. Reasoning Gym: Reasoning environments for reinforcement learning with verifiable rewards, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Sam Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

Alexander von Recum, Leander Girrbach, and Zeynep Akata. Are reasoning LLMs robust to interventions on their chain-of-thought? In *International Conference on Learning Representations*, 2026.

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

Cheng Xu, Nan Yan, Liming Chen, and M-Tahar Kechadi. Phantom gains: Auditing self-improvement against a measured null, 2026.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report, 2025.

Richard J. Young. Measuring faithfulness depends on how you measure: Classifier sensitivity in LLM chain-of-thought evaluation, 2026.

Qinan Yu, Alexa R. Tartaglini, Peter Hase, Carlos Guestrin, and Christopher Potts. Outcome rewards do not guarantee verifiable or causally important reasoning. In *International Conference on Machine Learning*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, et al. DAPO: An open-source LLM reinforcement learning system at scale, 2025.

A AUDIT PROTOCOL

A candidate paper was eligible when its task-level effects used the exact-prefix CIR scoring rule of Equation 1 and reported an effect estimate with a recoverable positive evaluation count. Inclusion also required an open autoregressive model with at most 4B parameters, tasks with mechanically verifiable outcomes, and recoverable item, resample, and training-trajectory counts. Candidates came from a fixed conceptual literature search plus a starting set of known trace-intervention papers. The targeted screen contains 11 candidate trace-intervention papers, including the one compatible study; it is not a random sample of the literature. Early-answer agreement, resampled-outcome Kullback–Leibler (KL) divergence, attention-masked KL, and sentence-level confidence are different quantities, so a CIR floor does not transfer to them. The exclusions follow this scoring rule and ignore the direction of any reported effect. Table 4 lists the excluded candidates.

Candidate paper	Primary exclusion
Lanham et al. (2023)	Early-answer agreement, not CIR
Macar et al. (2026)	Resampled-outcome KL, not CIR
Bulut (2026)	Restart contrast, not a CIR effect
Bentham et al. (2024)	Answer-letter agreement
Young (2026)	Hint-acknowledgment classifier
Scalena et al. (2026)	Sentence-level confidence
Garcia (2026)	Accuracy under corruption
von Recum et al. (2026)	Intervention recovery accuracy
Paul et al. (2024)	Mediation accuracy contrast
Jia et al. (2026)	Entropy and attention-masked KL

Table 4: Candidate papers excluded by the scoring rule.

The published figure's vector paths recover 40 after-training markers and 38 before-training markers. Among the 38 signed effects, which include two rows without a positive item count that Table 7 omits, the extraction reproduces the study's reported count of 19 tasks whose CIR decreased. The *cryptarithm* and *modulo-grid* rows lack before-training markers. The *binary-matrix* and *manipulate-matrix* rows have signed effects, but the audited study's item-generation configuration does not support a defensible positive evaluation count. Thus 38 rows have signed effects, two of those lack a defensible item count, and 36 can be classified.

The extraction assigns each vector marker a tolerance of ± 0.0010 natural-log JS. The signed after-minus-before difference then has tolerance ± 0.0020. The classifier acts on the magnitude interval $[L, U]$ defined in Section 3. Equality falls in the marginal class:

$$\text{label}([L, U], r) = \begin{cases} \text{resolvable}, & L > r, \\ \text{under-resolved}, & U < r, \\ \text{marginal}, & \text{otherwise}. \end{cases} \tag{4}$$

The audited study evaluates Qwen2.5-3B with one response per item at temperature 1.0 and evaluates CIR at 11 prefix positions along each trace. It reports one training trajectory, without a recoverable seed identifier or sampling intervals.

B FLOOR CALIBRATION

The calibration code works in normalized CIR, CIR divided by $\ln 2$, and this appendix reports its values in those units. We write $u_{512} = r_{512}/\ln 2$ for the normalized floor at 512 items; multiplying any value here by $\ln 2$ gives natural-log JS.

Calibration model and prompt. The calibration uses Qwen3-1.7B with its built-in thinking mode enabled, on held-out GSM8K test questions. Three sampling seeds, 17, 29, and 43, each supply 512 held-out items. Each seed samples 96 tokens of native reasoning at temperature 0.6 and top-p 0.95; each calibration trace keeps the first 89 of them and appends the nonce line. The calibration's CIR averages Equation 1 over every prefix depth from 1 to 96 tokens, where the audited study uses 11 positions along full traces. The system prompt, reproduced verbatim, is:

```
Solve the problem carefully using your native thinking mode.
A verification nonce may be shown at the end of your
thinking.  After thinking, answer exactly as 'Final:
<numeric answer>; Check:  <nonce>'.  Copy the shown nonce
exactly.
```

The user message is `Question:` followed by the GSM8K question. Each calibration trace keeps the first 89 native reasoning tokens and ends with the line `Verification nonce:` {`nonce`}`.`, so every trace has 96 tokens. The nonce is one of `MESA`, `RUBY`, `TANGO`, or `NOVA`. The target and alternative nonces rotate deterministically across items. The evaluator appends `Final:` {`gold`}`; Check:` after the reasoning, where {`gold`} is the correct numeric answer, and scores the probability of the target nonce tokens.

Planted effects. When the target is added, baseline traces show the alternative nonce and assigned traces show the target; when it is removed, baseline traces show the target and assigned traces show the alternative. Assignment is a deterministic hash of the protocol, score, trace condition, direction, sampling seed, planted size, and item identifier. Unassigned items keep their baseline, so their paired effect is zero. The planted sizes are mean effects of 0.005, 0.010, 0.020, 0.040, and 0.080 normalized CIR in each direction. Table 5 gives the item-only floor at each planted effect. Both directions agree to the fourth decimal, and the floor rises monotonically with the planted effect, from 0.0082 to 0.0295. The audit uses the smallest entry, the choice most favourable to the published effects. Larger entries correspond to multiples of r, and the four-times row of Table 2 exceeds every entry.

	Item-only floor u_{512}		Between-seed half-width	
Planted effect	Target added	Target removed	Target added	Target removed
0.005	0.0082	0.0082	0.0000	0.0000
0.010	0.0106	0.0106	0.0000	0.0000
0.020	0.0149	0.0149	0.0000	0.0000
0.040	0.0213	0.0213	0.0000	0.0000
0.080	0.0295	0.0295	0.0000	0.0001

Table 5: The planted-CIR floor grows with the planted effect and is nearly identical in both directions. Values are in normalized CIR at 512 items, from Equation 2; between-seed half-widths use the spread of the three seed means.

Variance components. The item-plus-seed variant adds the smallest nonnegative estimate of excess between-seed variance across the calibration groups, defined as the variance of the three seed means minus the item variance of a seed mean. For planted CIR the excess estimate is zero in every group. With three seeds, a zero estimate does not show that between-seed variation is absent. Table 6 compares planted CIR with two other dependence scores measured in the same calibration. Attention blocking masks the trace from the answer's attention, and resampled continuation regenerates the trace from each prefix before scoring the answer. Both were measured under the nonce line and under an outcome switch, which moves assigned items between continuations that reach the correct

answer and continuations that reach an incorrect one. The between-seed half-width exceeded the item-only floor in 10 of the 50 groups of score, condition, planted effect, and direction, all in the outcome-switch condition of these two scores. A floor for those scores therefore needs a between-seed term. The largest between-seed half-width, 0.0400 at 512 items under attention blocking, sets the attention-blocking row of Table 2, which leaves 28 task effects resolvable. For planted CIR the largest between-seed half-width is 0.0001.

Score	Trace	Largest between-seed half-width	Seed-dominated groups
Exact-prefix CIR	nonce line	0.0001	0 of 10
Attention blocking	nonce line	0.0125	0 of 10
Attention blocking	outcome switch	0.0400	4 of 10
Resampled continuation	nonce line	0.0054	0 of 10
Resampled continuation	outcome switch	0.0180	6 of 10

Table 6: Between-seed variation is negligible for planted CIR but not for every score. Values are in normalized CIR at 512 items; each score and trace condition has ten groups of planted effect and direction. A group is seed-dominated when its between-seed half-width exceeds its item-only floor. Nonce-line rows plant effects with the verification line. Outcome-switch rows plant effects by switching assigned items between correct-outcome and incorrect-outcome continuations, unlike the unplanted natural-trace contrast of Section 4.3.

Natural-trace substitution. The natural-trace sensitivity uses the same three sampling seeds, each contributing 512 GSM8K questions whose continuations reach both outcomes. For each question, the model samples continuations from a shared prefix of its own reasoning that ends at 24, 48, or 72 tokens, chosen by hash, and two continuations that reach the correct answer are paired with two that reach an incorrect one. The item effect is the CIR of the correct answer on a correct-outcome trace minus its CIR on an incorrect-outcome trace, averaged over the two pairs. No effect is planted, and the mean contrast was close to zero, so the measurement gives item variation without a known effect to recover. We divide each item effect by $\ln 2$ and apply Equation 2 without changing the published intervals, item counts, or classification rule. The resulting floor is 0.0038 normalized CIR at 512 items, and the excess between-seed component is zero. Number format changes from under-resolved to marginal, and Letter counting from marginal to resolvable, under both the item-only and the item-plus-seed variants.

Task shift to MATH. The task-shift measurement keeps Qwen3-1.7B and the natural-trace contrast while moving to MATH. It scores the 128 questions that all three seeds sampled, with an item-only floor of 0.0268 and a separate between-seed half-width of 0.0130, both in normalized CIR at 128 items. The variance of the three seed means is about one quarter of the item variance of a seed mean, so the excess between-seed variance is truncated at zero. To confirm that the MATH setup reproduces the model's published accuracy, a separate check runs the same model with thinking disabled, the setting for which a published MATH-500 accuracy exists; on those 500 questions its accuracy is 0.726 against the published 0.730 (Yang et al., 2025). This measurement shows how much item variance can move with the task, for the calibration model only.

C Complete task-level classification

Table 7 lists every classified task with its signed effect interval, item count, floor, and labels under the planted-effect and natural-trace variance sources. Only Number format and Letter counting change between the two label columns. The code supplement contains each digitized effect with its interval, and the extraction code that recovers the marker coordinates from the published figure.

Table 7: All 36 classifiable task results. Signed intervals are after-minus-before CIR effects read from the published figure, with ± 0.0020 reading tolerance; r is the floor at each item count n. Labels: R resolvable, M marginal, UR under-resolved; the last column substitutes natural-trace item variance.

Task	n	Signed effect interval	Floor r	Label	Natural-trace label
Advanced geometry	1,000	$[+0.5672, +0.5712]$	0.0041	R	R
Base conversion	1,000	$[+0.3216, +0.3256]$	0.0041	R	R
Binary alternation	1,000	$[-0.1435, -0.1395]$	0.0041	R	R
Bitwise arithmetic	1,000	$[-0.1035, -0.0995]$	0.0041	R	R
Caesar cipher	1,000	$[-0.0238, -0.0198]$	0.0041	R	R
Count bits	1,000	$[-0.0188, -0.0148]$	0.0041	R	R
Countdown	1,000	$[+0.1333, +0.1373]$	0.0041	R	R
Family relationships	1,000	$[-0.1506, -0.1466]$	0.0041	R	R
Fraction simplification	1,000	$[+0.0470, +0.0510]$	0.0041	R	R
Futoshiki	1,000	$[-0.0360, -0.0320]$	0.0041	R	R
Largest island	1,000	$[+0.0612, +0.0652]$	0.0041	R	R
Leg counting	1,000	$[+0.4444, +0.4484]$	0.0041	R	R
Letter counting	1,000	$[+0.0030, +0.0070]$	0.0041	M	R
Letter jumble	1,000	$[+0.1946, +0.1986]$	0.0041	R	R
Mini Sudoku	100	$[+0.0355, +0.0395]$	0.0129	R	R
Number format	1,000	$[-0.0020, +0.0020]$	0.0041	UR	M
Number sorting	1,000	$[-0.1442, -0.1402]$	0.0041	R	R
Palindrome generation	1,000	$[+0.1594, +0.1634]$	0.0041	R	R
Polynomial multiplication	1,000	$[+0.0851, +0.0891]$	0.0041	R	R
Pool matrix	1,000	$[-0.0243, -0.0203]$	0.0041	R	R
Power function	1,000	$[+0.3428, +0.3468]$	0.0041	R	R
Quantum lock	1,000	$[-0.0647, -0.0607]$	0.0041	R	R
Rotate matrix	1,000	$[-0.0150, -0.0110]$	0.0041	R	R
Sentence reordering	1,000	$[-0.0777, -0.0737]$	0.0041	R	R
Shortest path	1,000	$[-0.0347, -0.0307]$	0.0041	R	R
Simple equations	1,000	$[+0.2637, +0.2677]$	0.0041	R	R
Simple geometry	1,000	$[+0.2494, +0.2534]$	0.0041	R	R
Simple integration	1,000	$[+0.3130, +0.3170]$	0.0041	R	R
Spiral matrix	1,000	$[-0.0535, -0.0495]$	0.0041	R	R
String manipulation	1,000	$[-0.3350, -0.3310]$	0.0041	R	R
String splitting	1,000	$[+0.2322, +0.2362]$	0.0041	R	R
String synthesis	1,000	$[-0.0141, -0.0101]$	0.0041	R	R
Tower of Hanoi	1,000	$[+0.0165, +0.0205]$	0.0041	R	R
Tsumego	1,000	$[-0.2784, -0.2744]$	0.0041	R	R
Word sequence reversal	1,000	$[-0.2136, -0.2096]$	0.0041	R	R
Word sorting	1,000	$[-0.2360, -0.2320]$	0.0041	R	R