When Evaluation Changes the Conclusion: Measuring the Reliability of Small RL Gains in Reasoning LLMs

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning (RL) post-training has become a widely adopted approach for improving the reasoning capabilities of large language models (LLMs). However, many reported improvements are relatively small, raising an important question: whether current evaluation protocols provide sufficient resolution to reliably measure such incremental gains.

We conduct a systematic empirical study of RL gain measurement reliability under different evaluation protocols. We evaluate Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Yang et al., 2024) with RL-trained variants including Static-KL GRPO, GRPO, and No-KL using greedy decoding, stochastic sampling, multi-seed evaluation, paired bootstrap confidence intervals, and evaluation-size scaling analysis.

Our results show that reproducibility across training seeds does not necessarily imply robustness across evaluation protocols. Under zero-shot greedy GSM8K evaluation, Static-KL achieves consistent improvements across three training seeds, whereas the same checkpoints yield near-zero gains under 4-shot evaluation. Under stochastic sampling with 16 generations, the measured RL–Base differences are only 0.2–0.5 percentage points, with confidence intervals including zero. Evaluation-size scaling further shows that larger evaluations provide more stable estimates of small effects.

1 Introduction

Reinforcement learning (RL) post-training has become an important approach for improving the reasoning capabilities of large language models (LLMs). By optimizing models with task-specific or verifiable rewards, RL methods have achieved notable progress in mathematical reasoning, code generation, and complex problem solving. However, as RL post-training methods improve model capabilities, the resulting gains are often relatively small, typically at the scale of a few percentage points. This raises a fundamental measurement question: **can current evaluation protocols reliably measure such small RL-induced improvements?**

Most RL post-training studies compare an RL-trained model with a baseline using benchmark accuracy. However, evaluation outcomes can vary with prompting, decoding, stochastic sampling, and evaluation-set size. When RL gains are small, such variation can be comparable to the measured improvement.

In this work, we systematically study the reliability of measuring small RL gains in reasoning language models. We evaluate RL variants under different decoding, prompting, stochastic sampling, and evaluation-size settings using paired bootstrap confidence intervals and uncertainty analysis.

Our experiments reveal four main findings.

First, reproducibility across training seeds does not imply robustness across evaluation protocols. Static-KL produces consistent gains across three training seeds under zero-shot greedy GSM8K evaluation, yet the same checkpoints yield near-zero gains under 4-shot evaluation.

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.

Second, evaluation variation can exceed incremental RL effects. Prompt changes produce accuracy shifts larger than the RL gains measured under the same benchmark.

Third, under stochastic decoding, the measured RL–Base differences are substantially smaller than under greedy evaluation, with confidence intervals including zero under the tested GSM8K sampling protocol.

Fourth, RL post-training changes observable generation behavior beyond final accuracy. These behavioral changes do not necessarily correspond to measurable benchmark improvements.

We instead study the reliability of the evaluation process itself and how the measured magnitude and empirical resolution of small RL gains depend on evaluation design and statistical uncertainty.

Specifically, we make the following contributions:

1. **Training reproducibility versus evaluation robustness.** We show that small RL gains can be reproducible across training seeds yet remain sensitive to evaluation protocols.
2. **A measurement perspective on RL evaluation.** We combine paired bootstrap confidence intervals and sample-size scaling to characterize small RL gains relative to evaluation variability.
3. **Evaluation variation versus RL effects.** We show that evaluation changes can produce accuracy differences comparable to or larger than measured RL gains.
4. **Generation behavior beyond accuracy.** We show that RL-induced generation changes do not necessarily correspond to reliably measurable accuracy improvements.

2 Related Work

RL post-training for reasoning. Reinforcement learning (RL) post-training has become an important paradigm for improving the reasoning capabilities of large language models (LLMs). Early reinforcement learning from human feedback (RLHF) demonstrated the effectiveness of policy optimization for aligning language models with human preferences (Schulman et al., 2017; Ouyang et al., 2022). More recently, reinforcement learning with verifiable rewards (RLVR) has enabled substantial progress on mathematical reasoning and other tasks with automatically checkable outcomes. DeepSeekMath introduced Group Relative Policy Optimization (GRPO), which estimates relative advantages among sampled responses without a separate critic model (Shao et al., 2024), while DeepSeek-R1 demonstrated that large-scale RL post-training can induce strong reasoning behavior (DeepSeek-AI et al., 2025). Related reasoning models and analyses have further investigated training stability, reward design, KL regularization, and the structural properties of RL post-training(Yang et al., 2024; Jaech et al., 2024; Liu et al., 2025; Samineni et al., 2025; Dai et al., 2025).

Recent work has questioned whether improvements observed after RL necessarily correspond to expanded reasoning capability. Yue et al. examine RL and base models across sampling budgets and show that improvements at small k need not imply an expanded reasoning capability boundary (Yue et al., 2025). Other studies identify alternative explanations or confounds behind apparent RLVR improvements, including spurious reward signals, contamination, and memorization-related effects (Shao et al., 2025; Wu et al., 2026; Tao et al., 2026; Chen et al., 2026a; Yan et al., 2026). These studies primarily investigate what RL training changes or why apparent improvements arise. Our work is complementary: rather than proposing a new RL algorithm or attributing gains to a particular training mechanism, we ask how reliably small RL–Base differences are measured under different evaluation protocols.

Evaluation sensitivity and reliability. Reliable LLM evaluation depends not only on the evaluated model, but also on benchmark construction, prompting, decoding, contamination control, and other evaluation choices. Broad evaluation frameworks such as HELM and BIG-bench have emphasized standardized and diverse evaluation (Liang et al., 2023; Srivastava et al., 2023), while LiveBench and Chatbot Arena address contamination-resistant and human-preference-based evaluation, respectively (White et al., 2025; Chiang et al., 2024). Benchmark design itself can substantially affect model comparisons and rankings (Alzahrani et al., 2024), while recent work has questioned

whether standard RL benchmarks can reliably distinguish genuine methodological progress (Chen et al., 2026b).

Prompting is another important source of evaluation sensitivity. Few-shot example selection and ordering can affect measured performance (Liu et al., 2022; Lu et al., 2022), and recent work has explicitly quantified prompt sensitivity in LLM evaluation (Chatterjee et al., 2024; Zhuo et al., 2024). More broadly, reproducibility studies have shown that conclusions about reasoning progress can depend on evaluation configuration (Hochlehnert et al., 2025).

Recent work is particularly relevant to RL reasoning evaluation. Wu et al. identify measurement gaps in RLVR evaluation arising from budget mismatch, attempt inflation and calibration drift, and benchmark contamination (Wu et al., 2025). Chilaka et al. study how prompt format, decoding, and statistical considerations affect pass@k comparisons in RL reasoning evaluation (Chilaka et al., 2026). These findings reinforce the view that reported RL improvements cannot always be interpreted independently of the evaluation protocol.

Our work asks a related but distinct question: *when an RL improvement is reproducible across training seeds, is it also robust across evaluation protocols?* We study small RL–Base effect estimates across zero- and few-shot prompting, prompt formulations, greedy and stochastic decoding, benchmarks, and model families. This distinction between training-seed reproducibility and evaluation-protocol robustness is central to our analysis.

Statistical uncertainty and benchmark resolution. Statistical uncertainty is central to interpreting small differences between models. Bootstrap resampling provides a general framework for estimating sampling uncertainty (Efron, 1979), and prior NLP work emphasizes the importance of appropriate statistical testing when comparing systems (Dror et al., 2018). In reinforcement learning, uncertainty across runs and the limitations of point-estimate comparisons have similarly motivated more statistically grounded evaluation practices (Agarwal et al., 2021).

Recent work has brought these concerns directly to language-model evaluation. Madaan et al. quantify variance in evaluation benchmarks (Madaan et al., 2024), while Miller develops a statistical perspective on uncertainty and error bars in language-model evaluations (Miller, 2024). Other work studies uncertainty estimation and evaluation robustness more broadly (Xie et al., 2025; Xia et al., 2025; Matta et al., 2026; Golikov et al., 2026). Bowyer et al. further highlight limitations of asymptotic uncertainty approximations in small-sample LLM evaluation (Bowyer et al., 2025).

Closely related to reasoning evaluation, Yang et al. develop statistical inference for pass@k crossovers between RLVR and base models, showing that visually apparent crossovers need not be statistically established (Yang et al., 2026). Their analysis focuses on inference over pass@k curves and crossover behavior. In contrast, we focus on the reliability of small RL–Base effect estimates across evaluation protocols.

Building on these foundations, we use paired bootstrap confidence intervals, bootstrap positive-difference fractions, and evaluation-size scaling to characterize the uncertainty and empirical resolution of small RL gains. We additionally introduce the Evaluation Detectability Ratio (DR) as a descriptive signal-to-uncertainty diagnostic relating the magnitude of an observed RL–Base difference to the empirical uncertainty scale of the corresponding evaluation protocol. DR is not a hypothesis test, statistical power calculation, or universal significance threshold; it is intended as a compact descriptive summary of effect magnitude relative to evaluation uncertainty.

3 Results

3.1 Small RL Gains Are Sensitive to Evaluation Protocol

We first investigate whether measured RL gains remain stable under different evaluation protocols. We train Static-KL GRPO models with three independent random seeds (42, 1337, and 3407) and evaluate them on GSM8K (Cobbe et al., 2021) using zero-shot and 4-shot greedy decoding. Figure 1 reports the measured Static-KL gains over the Base model with 95% paired-bootstrap confidence intervals.

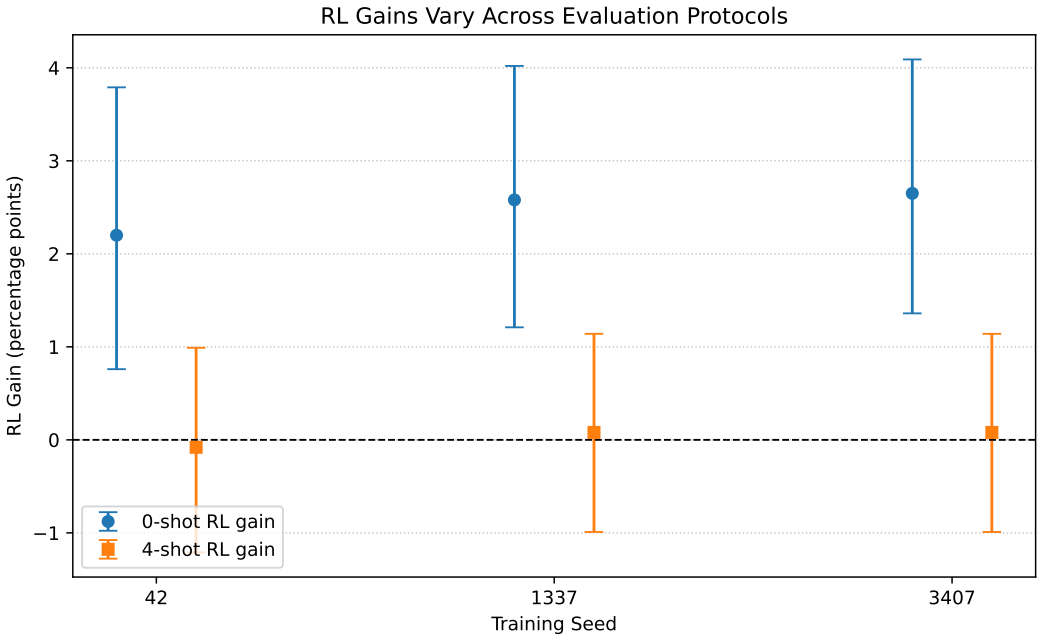


Figure 1: Measured Static-KL gains over Base on GSM8K across three training seeds under zero-shot and 4-shot greedy evaluation. Error bars denote 95% paired-bootstrap confidence intervals.

Under zero-shot evaluation, all Static-KL checkpoints obtain positive gains with confidence intervals excluding zero. The improvements are +2.20 percentage points (CI [0.76, 3.79]), +2.58 points (CI [1.21, 4.02]), and +2.65 points (CI [1.36, 4.09]) for seeds 42, 1337, and 3407, respectively. In contrast, under 4-shot evaluation, the gains become close to zero: -0.08 points (CI [-1.21, 0.99]), +0.08 points (CI [-0.99, 1.14]), and +0.08 points (CI [-0.99, 1.14]).

The consistency across training seeds shows that the zero-shot gain is reproducible across independent training runs. However, the same checkpoints yield near-zero gains under 4-shot evaluation, demonstrating that training-seed reproducibility does not imply robustness across evaluation protocols.

We further examine whether this sensitivity extends beyond GSM8K. Figure 2 summarizes Static-KL gains across four additional reasoning benchmarks—MATH-500 (Hendrycks et al., 2021), SVAMP (Patel et al., 2021), AMC23 (Art of Problem Solving, 2023), and StrategyQA (Geva et al., 2021)—under zero-shot and 4-shot evaluation.

Under zero-shot evaluation, GSM8K, MATH-500, and SVAMP show positive gains with confidence intervals excluding zero, whereas AMC23 and StrategyQA exhibit larger uncertainty. Under 4-shot evaluation, the measured differences become smaller and the confidence intervals generally include zero.

Table 1 summarizes the detectability ratio (DR), a descriptive signal-to-uncertainty diagnostic defined as

$$R_{95} = \frac{CI_u - CI_l}{2}, \qquad \mathrm{DR} = \frac{|\hat{\Delta}|}{R_{95}},$$

where $\hat{\Delta}$ is the measured RL–Base accuracy difference and R_{95} is the half-width of its 95% paired-bootstrap confidence interval.

For GSM8K, MATH-500, SVAMP, and AMC23, DR decreases substantially when moving from zero-shot to 4-shot evaluation, while StrategyQA remains low under both protocols. AMC23 further shows that a larger point estimate does not necessarily correspond to a larger DR when evaluation uncertainty is high. Descriptively, DR values near or below one correspond to measured differences

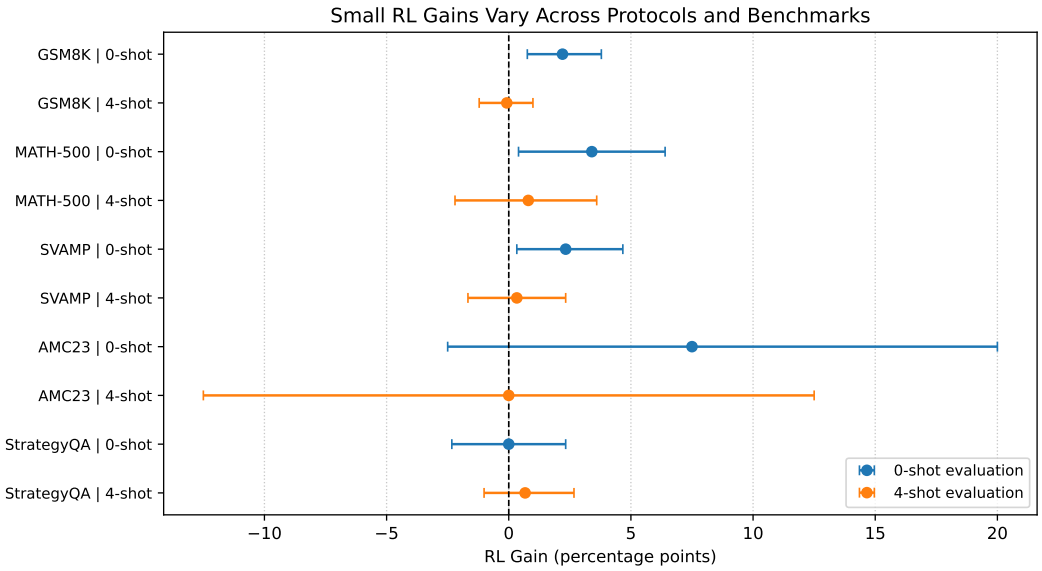


Figure 2: Measured Static-KL gains over Base across reasoning benchmarks under zero-shot and 4-shot evaluation. Error bars denote 95% paired-bootstrap confidence intervals.

Table 1: Evaluation detectability ratio (DR) for Static-KL relative to Base.

Benchmark	0-shot DR	4-shot DR
GSM8K	1.45	0.07
MATH-500	1.13	0.28
SVAMP	1.07	0.17
AMC23	0.67	0.00
StrategyQA	0.00	0.36

whose magnitude is comparable to or smaller than the empirical uncertainty scale of the corresponding protocol.

3.1.1 Prompt Variation Can Exceed RL Improvements

Evaluation configuration itself introduces substantial variation. On GSM8K, we compare two alternative zero-shot prompt formulations. Across Base, Static-KL, and No-KL models, changing the prompt produces accuracy shifts of 7.25, 6.94, and 6.42 percentage points, respectively.

These differences exceed the approximately 2–3 percentage-point RL gains observed under greedy evaluation; prompt formulation should therefore be explicitly controlled when measuring small improvements. This result does not imply that prompt variation is random noise; prompts may systematically alter generation behavior and response patterns.

3.2 Small RL–Base Differences Become Difficult to Resolve under Stochastic Evaluation

We next examine how measured RL–Base differences change when stochastic generation is introduced. For each GSM8K question, each model generates 16 responses using temperature sampling ($T = 0.9$). Comparisons are performed using paired bootstrap confidence intervals over the same 1,319 questions.

Under stochastic evaluation, all RL variants obtain positive point estimates, with measured RL–Base differences ranging from +0.2 to +0.5 percentage points, smaller than the corresponding differences

Table 2: Stochastic GSM8K evaluation with paired-bootstrap uncertainty. $P_{\text{boot}}(\Delta > 0)$ denotes the fraction of bootstrap resamples with positive RL–Base differences.

Model	Accuracy (%)	Δ (pp)	95% CI (pp)	$P_{\text{boot}}(\Delta > 0)$
Base	69.67	+0.00	–	–
Static-KL	69.98	+0.30	[-1.74,+2.50]	0.625
GRPO	70.20	+0.53	[-1.59,+2.65]	0.680
No-KL	69.90	+0.23	[-2.05,+2.43]	0.573

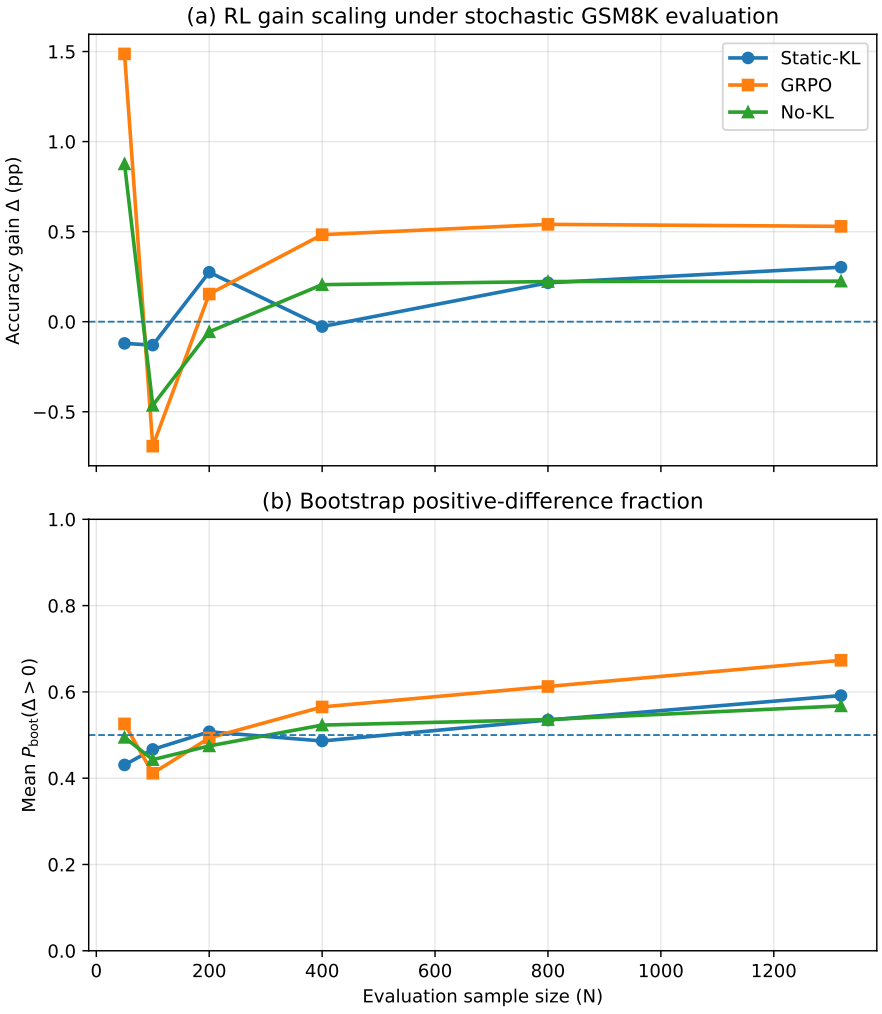


Figure 3: Evaluation-size scaling under stochastic GSM8K evaluation. (a) Mean RL–Base accuracy differences across repeated subset evaluations. (b) Mean bootstrap positive-difference fraction $P_{\text{boot}}(\Delta > 0)$ across repeated subset evaluations. The horizontal dashed line in (b) marks 0.5.

observed under greedy evaluation. All paired-bootstrap confidence intervals include zero, indicating that the measured differences are small relative to the uncertainty under this protocol.

To examine evaluation-size effects, we repeatedly sample GSM8K subsets with sizes $N \in \{50, 100, 200, 400, 800, 1319\}$.

For each subset size, we summarize the estimated RL–Base accuracy difference and the bootstrap positive-difference fraction $P_{\text{boot}}(\Delta > 0)$.

Figure 3(a) shows that estimated RL–Base differences vary substantially for small evaluation subsets. As the evaluation size increases, the estimates become more concentrated around the small full-set differences reported in Table 2. Figure 3(b) reports the mean bootstrap positive-difference fraction $P_{\mathrm{boot}}(\Delta > 0)$ across repeated subset evaluations. For Static-KL, this quantity remains near 0.5 across evaluation sizes, consistent with an effect that is small relative to the uncertainty of the stochastic evaluation protocol.

Importantly, $P_{\mathrm{boot}}(\Delta > 0)$ is descriptive rather than a frequentist p-value or CI-exclusion frequency. Increasing evaluation size reduces estimate variability, while values near 0.5 remain consistent with small RL–Base differences relative to uncertainty under the tested stochastic protocol.

These results show that evaluation protocol affects not only reported accuracy values but also the statistical resolution at which RL-induced improvements can be observed. Importantly, the impact of evaluation uncertainty is benchmark-dependent: while stochastic evaluation makes small gains difficult to resolve on GSM8K, larger improvements may remain distinguishable on other benchmarks such as MATH-500.

3.3 Cross-Model Validation Shows Similar Evaluation Sensitivity

To examine whether the observed evaluation sensitivity extends beyond Llama-3.1-8B-Instruct, we conduct an additional validation experiment on Qwen2.5-7B-Instruct (Yang et al., 2024) using the same Static-KL GRPO objective. Table 3 summarizes the measured gains.

Table 3: Cross-model validation of Static-KL RL gains on Qwen2.5-7B-Instruct.

Benchmark	Protocol	RL Gain (pp)	95% CI (pp)
GSM8K	0-shot	+2.88	[1.51,4.25]
GSM8K	4-shot	+0.99	[0.08,1.90]
MATH-500	0-shot	-0.60	[-3.00,1.60]
MATH-500	4-shot	+0.60	[-1.40,2.40]
SVAMP	0-shot	+2.33	[0.00,4.67]
SVAMP	4-shot	0.00	[-1.67,1.67]

The Qwen experiments show heterogeneous protocol sensitivity across benchmarks. On GSM8K, the measured RL–Base difference decreases from +2.88 percentage points under zero-shot evaluation to +0.99 points under 4-shot evaluation; the corresponding 4-shot 95% confidence interval is [0.08, 1.90]. On MATH-500 and SVAMP, several confidence intervals include zero, indicating that the measured differences are small relative to the uncertainty under these evaluation settings.

These results provide additional evidence that evaluation sensitivity is not restricted to a single model family.

3.4 RL Post-Training Changes Generation Behavior Beyond Accuracy

Finally, we analyze whether RL changes observable generation behavior even when accuracy differences remain difficult to resolve. Using stochastic GSM8K generations, we compare output length, correctness-conditioned generation length, Pass@1, and majority-vote accuracy (Wang et al., 2023).

Table 4: Generation behavior under stochastic GSM8K evaluation. Token counts describe output behavior and should not be interpreted as direct reasoning quality measures.

Model	Avg. Tokens	Correct Tokens	Wrong Tokens	Pass@1	Maj@16
Base	227.0	206.8	273.3	69.67	89.08
Static-KL	209.7	190.8	253.8	69.98	88.78
GRPO	221.2	201.5	267.7	70.20	89.84
No-KL	215.2	196.2	259.3	69.90	88.93

All RL variants modify generation-length distributions compared with Base. Static-KL produces shorter responses on average, whereas GRPO remains closer to the Base model.

However, these behavioral changes do not correspond to monotonic improvements in accuracy metrics. Static-KL reduces output length but does not improve majority-vote accuracy, while GRPO achieves the highest Maj@16 despite maintaining a generation length closer to Base.

These results indicate that RL post-training changes observable generation behavior beyond final accuracy differences. Nevertheless, behavioral changes alone should not be interpreted as direct evidence of improved reasoning capability.

4 Discussion

4.1 Evaluation Resolution as a Measurement Constraint

Our results suggest that evaluating small RL-induced improvements can be viewed as a measurement problem. An RL post-training procedure may modify model behavior, but whether such changes become observable as benchmark accuracy gains depends on the resolution of the evaluation protocol. Similar to measurement processes in empirical studies, an evaluation setting in which uncertainty is comparable to the measured difference provides limited resolution for distinguishing small model differences from measurement variability.

This distinction separates training-seed reproducibility from evaluation-protocol robustness. Our results show that the former does not necessarily imply the latter: reproducible RL gains can still depend on the measurement protocol under which they are evaluated. In this view, evaluation design choices act as part of the measurement instrument, not as incidental reporting details.

A key implication is that reproducibility and robustness characterize different dimensions of empirical reliability. Repeating RL training across random seeds tests whether an observed gain is reproducible under a fixed evaluation procedure, but does not establish that the gain persists when the measurement procedure changes. In our experiments, the zero-shot GSM8K gain is reproducible across three independently trained Static-KL checkpoints, yet largely disappears under 4-shot evaluation for the same checkpoints. Thus, multi-seed training evaluation and evaluation-protocol robustness should be treated as complementary rather than interchangeable checks.

4.2 Why Small RL Gains Are Difficult to Resolve

The experiments show that incremental RL improvements often operate where the measured difference is comparable to evaluation uncertainty: stochastic generation adds sampling variability, while alternative prompt formulations can substantially shift measured accuracy.

When the uncertainty associated with an evaluation protocol is comparable to the measured RL–Base difference, conclusions based on point estimates alone become less stable. RL post-training may introduce localized behavioral changes, but their measured benchmark-level impact depends on how evaluation procedures sample and aggregate model outputs.

Importantly, our findings do not indicate that RL post-training is ineffective. Instead, they demonstrate that the evidential strength of a reported improvement depends on the interaction between the training procedure and the measurement process used for evaluation. Consequently, conclusions about small gains require careful consideration of evaluation resolution and uncertainty. Reporting the evaluation protocol alongside the observed gain is therefore part of the empirical claim itself.

4.3 Practical Evaluation Checklist for Small RL Gains

Based on our findings, we recommend several practices for evaluating incremental improvements from RL post-training:

- **Report evaluation configurations explicitly.** Decoding strategy, temperature, sampling budget, and prompting protocol should be specified when reporting benchmark improvements.
- **Quantify uncertainty for small differences.** Confidence intervals or bootstrap-based uncertainty estimates should accompany small accuracy changes rather than relying only on point estimates.

- **Consider evaluation resolution.** Benchmark size and statistical uncertainty should be considered when interpreting improvements that are close to the measurement uncertainty scale.
- Separate behavioral changes from capability improvements and evaluate robustness across protocols when feasible.

5 LIMITATIONS

Our findings should be interpreted within the scope of the evaluated models, tasks, and protocols. First, our experiments focus on 7B–8B-parameter models and a limited set of reasoning benchmarks. The observed sensitivity may not generalize unchanged to larger models or other tasks, and benchmark uncertainty may depend on dataset size, difficulty, and response characteristics.

Second, greedy and stochastic decoding characterize different aspects of the model output distribution; we therefore do not interpret either protocol as intrinsically more reliable. Their differing estimates instead show that the measured magnitude and empirical resolution of small gains can depend on the evaluation functional.

Third, our empirical resolution analysis is descriptive rather than a formal statistical power analysis. The resolution scale R_{95} summarizes uncertainty under a particular evaluation configuration, while DR compares the magnitude of the measured RL–Base difference with this uncertainty scale.

Finally, our stochastic experiments use 16 generations per GSM8K question at a fixed temperature of $T = 0.9$. Different sampling temperatures, generation budgets, or aggregation rules may yield different measured effects. Generation length is treated only as an observable output statistic and not as a direct measure of reasoning quality or correctness.

Future work should examine whether similar evaluation sensitivity persists across larger models, broader reasoning domains, additional decoding configurations, sampling budgets, and evaluation metrics.

6 CONCLUSION

We investigated whether commonly used evaluation protocols can reliably measure small improvements introduced by RL post-training in reasoning language models.

Across multiple reasoning benchmarks and evaluation settings, we find that training-seed reproducibility does not imply evaluation-protocol robustness. For Llama-3.1-8B-Instruct, Static-KL produces consistent GSM8K gains across three training seeds under zero-shot greedy evaluation, while the same checkpoints yield near-zero gains under 4-shot evaluation. Controlled prompt variation can induce accuracy shifts larger than the RL gains being measured, and Qwen2.5-7B-Instruct experiments exhibit related protocol-dependent behavior in the tested settings.

Stochastic evaluation provides an additional perspective on this measurement problem. Under a 16-generation GSM8K sampling protocol, the evaluated RL variants show positive but small differences relative to Base, with 95% paired-bootstrap confidence intervals including zero. The evaluation-size scaling analysis further shows that small evaluation subsets can produce highly variable effect estimates, while larger evaluation sets provide more stable estimates under the tested stochastic protocol.

Beyond final accuracy, we show that RL post-training modifies observable generation behavior, including output-length distributions and correctness-conditioned response lengths. However, these behavioral changes do not necessarily correspond to reliably measurable accuracy improvements, highlighting the distinction between changing model behavior and improving benchmark performance. Overall, our results do not suggest that RL post-training is ineffective. Instead, they demonstrate that the evidential strength of a reported small RL improvement depends not only on the training algorithm and model, but also on the evaluation protocol and statistical resolution of the measurement setting. We recommend reporting evaluation configurations and uncertainty estimates when interpreting incremental RL improvements.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and presentation of this paper, including improving clarity, grammar, organization, and LaTeX formatting. The authors reviewed and verified all AI-assisted revisions to ensure that they accurately reflect the underlying experiments, analyses, and conclusions.

Generative AI tools were not used to generate experimental results, fabricate data, or replace the authors' verification of statistical analyses. All experimental results, numerical values, figures, tables, and quantitative claims reported in this paper were obtained from the authors' experiments and checked against the corresponding evaluation outputs and analysis scripts. The authors take full responsibility for the final content of this work and have independently verified all experimental results, analyses, and conclusions.

ETHICS STATEMENT

This work studies the reliability of evaluation methodologies for reinforcement-learning post-training of large language models. The experiments use established reasoning benchmarks and do not involve human subjects, private user data, or the collection of personally identifiable information.

The primary objective of this work is methodological: to improve the reliability and transparency of empirical claims about incremental improvements obtained through RL post-training. In particular, we examine how prompting protocols, decoding configurations, benchmark size, training randomness, and statistical uncertainty can affect conclusions about small model differences.

Because evaluation choices can influence empirical comparisons between models and training methods, we report confidence intervals and explicitly distinguish observed point estimates from differences that are resolved under the corresponding uncertainty scale. We also discuss the limitations of the proposed empirical measurement-resolution and detectability diagnostics and avoid interpreting them as universal statistical decision rules.

The stochastic and generation-behavior analyses are intended to characterize evaluation sensitivity and observable model behavior. They should not be interpreted as assessments of human-like reasoning processes or as evidence about internal cognitive mechanisms.

REPRODUCIBILITY STATEMENT

We provide the model configurations, RL post-training settings, benchmark details, prompting protocols, decoding settings, answer-extraction procedures, and statistical analysis methods required to reproduce the main experiments in Section 3 and the appendices.

The definitions and computation procedures for paired bootstrap confidence intervals, empirical measurement resolution R_{95}, the evaluation detectability ratio (DR), bootstrap-based ranking analysis, and sample-size scaling are described in the methodology section and Appendix C.

Appendix A reports the main RL post-training configurations, including LoRA parameters, KL coefficients, training data, random seeds, and hardware. Appendix B provides the zero-shot, few-shot, and prompt-variation templates used in the evaluation. Appendix C describes the paired-bootstrap procedures, empirical measurement-resolution metrics, detectability ratio, and resolution-scaling analyses. Appendix D documents the stochastic GSM8K evaluation and its within-benchmark scaling analysis. Appendix E provides the generation-behavior analysis and supplementary stochastic metrics. Appendix F summarizes the cross-model Qwen2.5-7B-Instruct evaluation.

All reported comparisons preserve question-level correspondence between models wherever paired statistical analysis is performed. Random seeds used for training, bootstrap resampling, and subset sampling are reported where applicable. Code for evaluation, paired-bootstrap analysis, evaluation-size scaling, and figure generation, together with prompt templates and experimental configurations, is provided in the supplementary material. These materials are intended to enable independent verification of the evaluation-sensitivity results reported in this paper.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.744.

Art of Problem Solving. AMC 2023 competition problems. `https://artofproblemsolving.com/wiki/index.php/AMC_12`, 2023.

Sam Bowyer, Laurence Aitchison, and Desi R. Ivanova. Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14550–14565. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.852.

Peter Chen, Xiaopeng Li, Ziniu Li, Wotao Yin, Xi Chen, and Tianyi Lin. Exploration vs exploitation: Rethinking RLVR through clipping, entropy, and spurious reward. In *International Conference on Learning Representations*, 2026a. arXiv:2512.16912.

Zihan Chen, Yiming Zhang, Hengguang Zhou, Zenghui Ding, Yining Sun, and Cho-Jui Hsieh. Rethinking RL evaluation: Can benchmarks truly reveal failures of RL methods? In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026b. doi: 10.18653/v1/2026.findings-acl.769.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024.

Saiakhil Chilaka, Neil Patel, Charles Xue, and Eric Shen. Pass@k is not a property of a model: Prompt format, decoding, and statistical power in evaluations of RL for reasoning. *Preprints.org*, 2026. doi: 10.20944/preprints202608.1495.v1. Preprint, not peer-reviewed.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Muzhi Dai, Shixuan Liu, and Qingyi Si. Stable reinforcement learning for efficient reasoning. *arXiv preprint arXiv:2505.18086*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

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 (Volume 1: Long Papers)*, pp. 1383–1392. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1128.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

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.

Pavel Golikov, Evgenii Opryshko, Gennady Pekhimenko, and Mark C. Jeffrey. Robust reasoning benchmark. *arXiv preprint arXiv:2604.08571*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. In *Conference on Language Modeling*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Carl Elman, et al. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.deelio-1.10.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

Krish Matta, Atharv Naphade, and Andy Zou. Rethinking uncertainty evaluation in large language models. *arXiv preprint arXiv:2607.19367*, 2026.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of NAACL-HLT*, pp. 2080–2094, 2021.

Soumya Rani Samineni, Durgesh Kalwar, Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. RL in name only? analyzing the structural assumptions in RL post-training for LLMs. *arXiv preprint arXiv:2505.13697*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. *arXiv preprint arXiv:2506.10947*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Yongding Tao, Tian Wang, Yihong Dong, Huanyu Liu, Kechi Zhang, Xiaolong Hu, and Ge Li. Detecting data contamination from reinforcement learning post-training for large language models. In *International Conference on Learning Representations*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations*, 2025.

Fang Wu, Aaron Tu, Weihao Xuan, Heli Qi, Xu Huang, Qingcheng Zeng, Shayan Talaei, Yijia Xiao, Peng Xia, Xiangru Tang, Yuchen Zhuang, Yinxi Li, Bing Hu, Hanqun Cao, Wenqi Shi, Rui Yang, Nan Liu, Huaxiu Yao, Ge Liu, Li Erran Li, Amin Saberi, Naoto Yokoya, Jure Leskovec, and Yejin Choi. Position: The hidden costs and measurement gaps of reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2509.21882*, 2025. doi: 10.48550/arXiv.2509.21882.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, Yanwei Fu, Qin Liu, Songyang Zhang, and Qi Zhang. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33944–33952, 2026. doi: 10.1609/aaai.v40i40.40687.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21381–21396, 2025. doi: 10.18653/v1/2025.findings-acl.1101.

Qiujie Xie, Qingqiu Li, Zhuohao Yu, Yuejie Zhang, Yue Zhang, and Linyi Yang. An empirical analysis of uncertainty in large language model evaluations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=J4xLuCt2kg`.

Lecheng Yan, Ruizhe Li, Guanhua Chen, Qing Li, Jiahui Geng, Wenxi Li, Longyue Wang, and Chenyang Lyu. Spurious rewards paradox: Mechanistically understanding how RLVR activates memorization shortcuts in LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Chen Yang, Xianyang Zhang, and Jun Chen. RLVR is a kernel, not a function: Statistical inference for pass@k crossovers. *arXiv preprint arXiv:2609.22547*, 2026.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, pp. 57654–57689. Curran Associates, Inc., 2025.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1950–1976. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.108.

A Training Details

The primary RL experiments are initialized from `Llama-3.1-8B-Instruct` and trained using the `trl` library with `GRPOTrainer`. Table 5 summarizes the main training configuration.

Table 5: Main training hyperparameters for RL post-training.

Hyperparameter	Value
Base model	Llama-3.1-8B-Instruct
Training framework	TRL GRPOTrainer
LoRA rank	16
LoRA alpha	32
LoRA target modules	q_proj, k_proj, v_proj, o_proj
LoRA dropout	0.05
Optimizer	AdamW
Learning rate	2×10^{-6}
Learning rate schedule	cosine decay
Batch size per device	2
Gradient accumulation steps	8
Number of generations per prompt	4
Maximum prompt length	512
Maximum completion length	512
Number of epochs	2
KL coefficient β (Static-KL)	0.02
KL coefficient β (No-KL)	0.0
Training data (Static-KL)	GSM8K train
Training data (No-KL HardMix)	GSM8K-Hard-Mix train
Random seeds (Static-KL)	42, 1337, 3407
Random seed (No-KL HardMix)	42
Precision	bfloat16
Gradient checkpointing	enabled
Hardware	$4\times$ A100 40GB

The three Static-KL checkpoints corresponding to random seeds 42, 1337, and 3407 are used for the multi-seed robustness analysis. The No-KL HardMix model differs from Static-KL in both KL regularization and training-data construction and is therefore treated as an additional RL-trained evaluation subject rather than as a controlled KL-only ablation.

The GRPO checkpoint used in the stochastic evaluation is included as an additional RL-trained variant for studying protocol sensitivity. We do not use comparisons among Static-KL, GRPO, and No-KL to claim superiority of one RL objective over another; the purpose of these variants is to provide multiple RL-induced model changes under a common evaluation framework.

For cross-model validation, a Static-KL GRPO variant is trained from `Qwen2.5-7B-Instruct` using the same LoRA rank, LoRA alpha, training procedure, and primary optimization configuration. The base checkpoint, tokenizer, and chat template are changed to the corresponding Qwen2.5 components.

B PROMPT TEMPLATES

B.1 ZERO-SHOT EVALUATION

For the primary zero-shot evaluation, the target question is formatted using the chat template associated with the corresponding model. The conceptual prompt structure is:

```
<|system|>
You are a helpful assistant.
<|user|>
{question}
<|assistant|>
```

The exact special tokens and message delimiters are generated automatically by the tokenizer and therefore differ between Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct.

B.2 FEW-SHOT EVALUATION

For few-shot evaluation, we prepend a fixed set of in-context demonstrations before the target question. The same demonstrations are used for all models within a benchmark comparison.

For a k-shot evaluation, the first k examples from the fixed demonstration list are used. The conceptual structure is:

```
<|system|>
You are a helpful assistant.

<|user|>
Q: {example_question_1}
<|assistant|>
A: {example_answer_1}

<|user|>
Q: {example_question_2}
<|assistant|>
A: {example_answer_2}

...

<|user|>
Q: {target_question}
<|assistant|>
```

The fixed set of eight GSM8K training examples used to construct the few-shot prompts is listed below. The primary comparison in the main text uses the first four examples for the 4-shot protocol.

1. **Q:** Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?
   A: Natalia sold 48/2 = 24 clips in May. Natalia sold 48+24 = 72 clips altogether in April and May. #### 72
2. **Q:** Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?
   A: Weng earns 12/60 = $0.2 per minute. Working 50 minutes, she earned 0.2 x 50 = $10. #### 10
3. **Q:** Betty is saving money for a new wallet which costs $100. Betty has only half of the money she needs. Her parents decided to give her $15 for that purpose, and her grandparents twice as much as her parents. How much more money does Betty need to buy the wallet?
   A: In the beginning, Betty has only 100 / 2 = $50. Betty's grandparents gave her 15 * 2 = $30. This means Betty needs 100 - 50 - 30 - 15 = $5 more. #### 5

4. **Q:** Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pages as yesterday. If she wants to read half of the remaining pages tomorrow, how many pages should she read?
   A: Julie read 12 x 2 = 24 pages today. She therefore read 12 + 24 = 36 pages in total. There are 120 - 36 = 84 pages remaining. Half of the remaining pages is 84/2 = 42. #### 42
5. **Q:** James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?
   A: He writes each friend 3*2=6 pages a week. So he writes 6*2=12 pages every week. That means he writes 12*52=624 pages a year. #### 624
6. **Q:** Mark has a garden with flowers. He planted plants of three different colors in it. Ten of them are yellow, and there are 80% more of those in purple. There are only 25% as many green flowers as there are yellow and purple flowers. How many flowers does Mark have in his garden?
   A: There are 80/100 * 10 = 8 more purple flowers than yellow flowers. Therefore, there are 10 + 8 = 18 purple flowers. Purple and yellow flowers total 10 + 18 = 28. There are 25/100 * 28 = 7 green flowers. The total is 28 + 7 = 35. #### 35
7. **Q:** Albert is wondering how much pizza he can eat in one day. He buys 2 large pizzas and 2 small pizzas. A large pizza has 16 slices and a small pizza has 8 slices. If he eats it all, how many pieces does he eat that day?
   A: He eats 32 slices from the large pizzas because 2 x 16 = 32. He eats 16 slices from the small pizzas because 2 x 8 = 16. Therefore, he eats 32 + 16 = 48 pieces. #### 48
8. **Q:** Ken created a care package to send to his brother, who was away at boarding school. Ken placed a box on a scale, and then he poured into the box enough jelly beans to bring the weight to 2 pounds. Then, he added enough brownies to cause the weight to double. Then, he added another 2 pounds of jelly beans. And finally, he added enough gummy worms to double the weight once again. What is the final weight of the box, in pounds?
   A: The initial weight is 2 pounds. Adding brownies doubles the weight to 4 pounds. Adding another 2 pounds gives 6 pounds. Doubling the weight again gives 12 pounds. #### 12

B.3 Controlled Prompt Variation

For the controlled GSM8K prompt-variation experiment, we use two zero-shot formulations:

- **Prompt A (step-by-step):**

```
Solve the problem step by step.
{question}
```

- **Prompt B (concise):**

```
Provide a concise answer without explanation.
{question}
```

Both prompts use the same underlying model chat template and greedy decoding configuration. Therefore, the comparison isolates the effect of this controlled prompt-formulation change while holding the model, benchmark, and decoding strategy fixed.

For completeness, Table 6 reproduces the corresponding full-set GSM8K results.

Table 6: Controlled prompt variation on the full GSM8K test set.

Model	Prompt A Acc.	Prompt B Acc.	Effect (pp)	95% CI (pp)
Base	75.21	68.01	7.25	[4.47, 10.01]
Static-KL	75.51	68.61	6.94	[4.25, 9.63]
No-KL HardMix	75.21	68.84	6.42	[3.64, 9.17]

C Bootstrap and Evaluation-Resolution Details

C.1 Paired Bootstrap Confidence Intervals

Let

$$a = (a_1, a_2, \ldots, a_n)$$

and

$$b = (b_1, b_2, \ldots, b_n)$$

denote binary correctness vectors for two models evaluated on the same n questions.

For the primary benchmark comparisons, we perform $B = 1000$ paired bootstrap resamples. At bootstrap iteration b, we sample

$$i_1, i_2, \ldots, i_n$$

uniformly with replacement from $\{1, \ldots, n\}$. The two models are evaluated on the same sampled indices, preserving question-level pairing.

The bootstrap accuracy difference is

$$d^{(b)} = \frac{1}{n}\sum_{j=1}^{n} a_{i_j} - \frac{1}{n}\sum_{j=1}^{n} b_{i_j}. \tag{1}$$

The 95% percentile-bootstrap confidence interval is defined by the 2.5th and 97.5th percentiles of

$$\left\{ d^{(1)}, d^{(2)}, \ldots, d^{(B)} \right\}.$$

All primary bootstrap analyses use a fixed random seed of 42 unless otherwise specified.

C.2 Empirical Measurement Resolution

For a particular model comparison and evaluation protocol, we define empirical measurement resolution as the half-width of its 95% paired-bootstrap confidence interval:

$$R_{95} = \frac{CI_{\text{upper}} - CI_{\text{lower}}}{2}. \tag{2}$$

This quantity characterizes the empirical uncertainty scale of the specific comparison. It depends on the benchmark, sample size, model pair, prediction structure, and evaluation protocol and should not be interpreted as an intrinsic property of the benchmark.

C.3 Evaluation Detectability Ratio

We define the evaluation detectability ratio as

$$DR = \frac{|G_{\text{RL}}|}{R_{95}}, \tag{3}$$

where G_{RL} is the observed accuracy difference between an RL-trained model and its Base model.

A larger DR indicates that the magnitude of the observed difference is large relative to the uncertainty scale of the corresponding evaluation setting.

DR is a descriptive diagnostic. It is not a hypothesis test, statistical power estimate, minimum detectable effect, or universal decision threshold.

C.4 Bootstrap-Based Ranking Analysis

For selected evaluation settings, we perform $B = 1000$ paired bootstrap resamples over test questions. Within each resample, we compute the accuracy of each evaluated model and identify the model with the highest accuracy.

When multiple models tie for the highest accuracy, the corresponding ranking credit is divided equally among the tied models. For model m, its bootstrap ranking frequency is therefore computed as the average fractional first-place credit across bootstrap resamples.

This statistic characterizes the stability of the observed ranking under resampling and is not interpreted as the probability that one model is intrinsically superior to another.

C.5 Cross-Benchmark Resolution Scaling

To study the relationship between evaluation-set size and empirical measurement resolution, we pool paired correctness outcomes from GSM8K, MATH-500, and SVAMP, yielding a total of 2,119 evaluation examples.

For each target size

$$s \in \{50, 200, 500, 1000, 2000\},$$

we perform the following procedure:

1. sample s questions without replacement from the pooled evaluation set;
2. apply paired bootstrap resampling to the sampled subset;
3. compute the 95% confidence interval of the Static-KL–Base accuracy difference;
4. record the confidence-interval half-width R_{95};
5. repeat the subset experiment 10 times and report the average half-width.

The resulting average half-widths are approximately 7.2, 3.8, 2.55, 1.75, and 1.25 percentage points for $N = 50, 200, 500, 1000$, and 2000, respectively.

Because the pooled benchmarks differ in difficulty, label structure, and model error patterns, this simulation is intended only as an empirical illustration of how uncertainty changes with evaluation size in our experimental setting. It should not be interpreted as a universal benchmark scaling law or formal sample-complexity result.

D Stochastic Sampling Evaluation

D.1 Sampling Protocol

We perform an additional stochastic evaluation on the full GSM8K test set containing 1,319 questions.

For each question and each evaluated model, we generate

$$n = 16$$

responses using temperature sampling with

$$T = 0.9.$$

The evaluated models are Base, Static-KL, GRPO, and No-KL. Predictions are aligned by question index so that all RL–Base comparisons preserve question-level pairing.

This experiment is designed as a complementary evaluation protocol rather than as a replacement for deterministic greedy decoding. Greedy and stochastic evaluation characterize different aspects of the model output distribution.

D.2 Stochastic Accuracy and Bootstrap Uncertainty

Table 7 summarizes the full-set stochastic comparison.

Table 7: Stochastic GSM8K evaluation and paired-bootstrap uncertainty.

Model	Accuracy (%)	Δ (pp)	95% CI (pp)	$P_{\mathrm{boot}}(\Delta > 0)$
Base	69.67	+0.00	–	–
Static-KL	69.98	+0.30	[-1.74, +2.50]	0.625
GRPO	70.20	+0.53	[-1.59, +2.65]	0.680
No-KL	69.90	+0.23	[-2.05, +2.43]	0.573

For these stochastic comparisons, we use $B = 10{,}000$ paired bootstrap resamples because the observed RL–Base differences are small.

We additionally report

$$P_{\mathrm{boot}}(\Delta > 0) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left[\Delta^{(b)} > 0\right]. \tag{4}$$

This quantity denotes the fraction of bootstrap resamples in which the estimated RL–Base difference is positive. It is a descriptive bootstrap statistic and is not interpreted as a frequentist p-value.

D.3 Within-Benchmark Sample-Size Scaling

For the stochastic GSM8K evaluation, we additionally study subset sizes

$$N \in \{50, 100, 200, 400, 800, 1319\}.$$

For each target size, repeated subsets are sampled from the full GSM8K evaluation set. For every subset, we compute the RL–Base accuracy difference and the bootstrap positive-difference fraction $P_{\mathrm{boot}}(\Delta > 0)$.

For a particular evaluation subset, $P_{\mathrm{boot}}(\Delta > 0)$ denotes the fraction of paired-bootstrap resamples in which the RL–Base accuracy difference is positive. We summarize this quantity across repeated subset evaluations by reporting its mean for each evaluation size.

Figure 3(a) reports the mean RL–Base accuracy difference across repeated subset evaluations, while Figure 3(b) reports the mean $P_{\mathrm{boot}}(\Delta > 0)$.

The quantity $P_{\mathrm{boot}}(\Delta > 0)$ is descriptive and should not be interpreted as a frequentist p-value, statistical power, or the frequency with which a 95% confidence interval excludes zero.

D.4 Stochastic Evaluation on MATH-500

To examine whether the observed evaluation sensitivity generalizes beyond GSM8K, we conduct additional stochastic evaluation on MATH-500. Each question is sampled with 16 generations under temperature sampling ($T = 0.9$). We compute question-level paired bootstrap confidence intervals over the 500 test questions.

Table 8: Stochastic MATH-500 evaluation with question-level paired bootstrap uncertainty.

Model	Accuracy (%)	Δ (pp)	95% CI (pp)	$P_{\mathrm{boot}}(\Delta > 0)$
Base	25.96	–	–	–
GRPO	28.31	+2.35	[1.39,+3.31]	1.000
Static-KL	24.66	-1.30	[-2.37,-0.24]	0.008

Unlike GSM8K, where stochastic evaluation reduces the detectability of small RL gains, MATH-500 exhibits a larger GRPO improvement that remains distinguishable under question-level bootstrap uncertainty. This result suggests that evaluation sensitivity depends jointly on benchmark characteristics and evaluation protocols.

E Generation Behavior Analysis

We use the stochastic GSM8K generations to characterize observable changes in generation behavior after RL post-training.

For each generated response, we record its generated-token length and whether the extracted final answer is correct. We then compute:

- mean generated-token length across all responses;
- mean generated-token length among correct responses;
- mean generated-token length among incorrect responses;
- Pass@1 accuracy;
- majority-vote accuracy using all 16 generations (Maj@16).

Table 9 reports the resulting statistics.

Table 9: Generation behavior under stochastic GSM8K evaluation.

Model	Avg. Tokens	Correct Tokens	Wrong Tokens	Pass@1	Maj@16
Base	227.0	206.8	273.3	69.67	89.08
Static-KL	209.7	190.8	253.8	69.98	88.78
GRPO	221.2	201.5	267.7	70.20	89.84
No-KL	215.2	196.2	259.3	69.90	88.93

All evaluated RL variants produce shorter generations on average than Base, although the magnitude of the difference varies across variants. Static-KL exhibits the largest reduction in average generated-token length.

Incorrect responses are longer than correct responses for every evaluated model. This pattern is descriptive and should not be interpreted causally. In particular, the results do not establish that longer reasoning causes incorrect answers or that shorter responses represent more efficient reasoning.

Generated-token length is therefore used only as an observable behavioral diagnostic. It may reflect changes in reasoning, verbosity, formatting, or termination behavior and should not be treated as a direct measurement of reasoning depth or reasoning quality.

F Cross-Model Validation

To examine whether the observed prompting-protocol sensitivity is specific to Llama-3.1-8B-Instruct, we additionally evaluate a Static-KL GRPO model based on Qwen2.5-7B-Instruct.

The Qwen Static-KL model is compared with its corresponding Base model under zero-shot and 4-shot greedy evaluation.

The Qwen experiments are intended as cross-model validation rather than as a comprehensive comparison between model families. On GSM8K, the measured Static-KL gain decreases from +2.88 percentage points under zero-shot evaluation to +0.99 percentage points under 4-shot evaluation. The 4-shot confidence interval remains slightly above zero and is therefore interpreted as a smaller but still positively resolved effect.

On MATH-500 and SVAMP, the exact pattern differs, but the measured differences remain sensitive to the prompting protocol. These results suggest that protocol sensitivity is not restricted to the Llama model family in the tested settings, while avoiding a claim of universal generalization.

Table 10: Cross-model validation with Qwen2.5-7B-Instruct.

Benchmark	Protocol	RL Gain (pp)	95% CI (pp)
GSM8K	0-shot	+2.88	[1.51, 4.25]
GSM8K	4-shot	+0.99	[0.08, 1.90]
MATH-500	0-shot	-0.60	[-3.00, 1.60]
MATH-500	4-shot	+0.60	[-1.40, 2.40]
SVAMP	0-shot	+2.33	[0.00, 4.67]
SVAMP	4-shot	0.00	[-1.67, 1.67]

G Additional Analysis and Interpretation Guidelines

The empirical analyses in this paper are intended to characterize the reliability of measuring small RL-induced differences rather than to establish a universal criterion for whether an RL method is effective.

In particular:

- a confidence interval that includes zero indicates that the observed difference is not resolved from zero at the corresponding empirical uncertainty scale; it does not establish that the true effect is exactly zero;
- greedy and stochastic evaluation characterize different functionals of the model output distribution and should not be interpreted as a hierarchy in which one protocol is inherently more valid than the other;
- prompt-induced accuracy variation is not treated as random statistical noise, because different prompts can systematically change model behavior;
- R_{95} and DR are descriptive evaluation diagnostics rather than formal statistical power measures;
- generation length is an observable output property and not a direct proxy for reasoning quality;
- results obtained on the tested 7B–8B-scale models and reasoning benchmarks should not be assumed to generalize universally to larger models, other RL objectives, or other evaluation domains.

These distinctions are important because the central claim of this work concerns *measurement reliability*: when expected improvements are small, the strength of an empirical conclusion depends jointly on the observed model difference and the protocol and statistical resolution under which that difference is measured.