Good Reasoning Makes Good Demonstrations: Implicit Reasoning Quality Supervision via In-Context Reinforcement Learning

Tiehua Mei[1*], **Minxuan Lv**[2*], **Leiyu Pan**[3], **Zhenpeng Su**[2†], **Hongru Hou**[1], **Hengrui Chen**[1], **Ao Xu**[1], **Deqing Yang**[1†]

[1]School of Data Science, Fudan University
[2]University of Chinese Academy of Sciences
[3]College of Intelligence and Computing, Tianjin University

☒ thmei24@m.fudan.edu.cn, suzhenpeng13@163.com

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) improves reasoning in large language models but treats all correct solutions equally, potentially reinforcing flawed traces that arrive at correct answers by chance. We observe that *better reasoning makes better demonstrations*: high-quality solutions serve as more effective in-context examples than low-quality ones. We term this teaching ability **Demonstration Utility**, and show that the policy model's own in-context learning ability provides an efficient way to measure it, yielding a quality signal termed **Evidence Gain**. To leverage this signal during training, we introduce **In-Context RLVR**, which prepends demonstrations before each rollout. Theoretically, we prove that this simple input modification implicitly reweights rewards by a factor approximately proportional to Evidence Gain, assigning higher weights to high-quality traces without requiring costly computation. Experiments on mathematical reasoning benchmarks demonstrate consistent improvements in both accuracy and reasoning quality over standard RLVR baselines. Our codes and datasets are available at https://github.com/Mithas-114/IC-DAPO.

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a powerful paradigm for improving LLM reasoning (Shao et al., 2024; Guo et al., 2025a), especially in domains such as mathematics where correctness can be checked by rules (Su et al., 2025c). By using outcome-level supervision, RLVR avoids costly process annotations and scales well (Mroueh, 2025). However, this simplicity comes with a limitation: all correct solutions receive equal reward, regardless of the reasoning used to obtain them (Do et al., 2025). This is problematic since models can produce flawed reasoning traces that coincidentally get correct answers, particularly when final answers are simple values that can be guessed (Guo et al., 2025b). Consequently, reinforcing such traces may corrupt internal reasoning strategies, degrading performance on other problems (MacDiarmid et al., 2025).

A natural solution is to use process reward models (PRMs) (Zhang et al., 2025b; Ye et al., 2025) that score intermediate steps. However, PRMs typically require extensive human annotation or auxiliary trained evaluators (Lightman et al., 2024). This raises a key question: *Can we encourage high-quality reasoning within RLVR without requiring step-level supervision or trained reward models?*

Demonstration Utility as Global Quality Signal. Our key insight is that *high-quality reasoning traces are better teachers than low-quality ones* (Min et al., 2022). Consider two solutions that both arrive at the correct answer: one is coherent and complete; the other contains redundant or unclear steps. When used as in-context demonstrations, the former provides transferable problem-solving patterns that help the model generate better solutions, while the latter provides less reference value (Li et al., 2025). We term this teaching ability **Demonstration Utility**. Crucially, the policy model's own in-context learning (ICL) ability provides a natural way to measure Demonstration Utility. Specifically, we construct a held-out validation set composed of questions and high-quality reference reasoning traces. We propose computing the average increase in the model's log-likelihood of generating these references after a candidate reasoning trace is prepended as a demonstration. We call this measure **Evidence Gain** (§2). Unlike PRMs that require external evaluators, Evidence Gain leverages the intrinsic ICL capability of the policy model itself. Experiments in Section 2 confirm that this intrinsic signal effectively distinguishes good reasoning from bad.

[*]Equal contribution.
[†]Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 30452–30469

Implicit Reward Reweighting via In-Context RLVR. While Evidence Gain provides a reasoning quality signal, computing it as rewards would introduce substantial overhead. Fortunately, we show that this explicit computation is unnecessary. Our key idea is to reverse the process: instead of computing Evidence Gain *after* generation as rewards, we use the same validation set to guide training *before* generation. Specifically, before each rollout, we sample a demonstration from the validation set and prepend it to the current question, then perform standard RL updates in this demonstration-conditioned setting, a procedure we term **In-Context RLVR**. Theoretically, we prove that this training objective is equivalent to standard zero-shot RLVR but with *rewards implicitly reweighted by a factor approximately proportional to* $\exp(\Delta)$, where Δ denotes Evidence Gain (Theorem E.3, Theorem E.4). Consequently, high-quality traces with greater teaching utility receive amplified gradient signals, while low-quality traces receive relatively lower weights through this implicit reweighting mechanism.

Contributions. **(1)** We introduce **Evidence Gain**, a quality signal that measures reasoning quality by leveraging the policy model's intrinsic ICL ability, requiring no external evaluators or step-level supervision. **(2)** We show that this signal can be seamlessly integrated into training via **In-Context RLVR**, which prepends demonstrations during training. We prove (Theorem E.3) that this simple input modification implicitly reweights rewards, and further show (Theorem E.4) that the weight factor is approximately proportional to $\exp(\Delta)$. **(3)** Experiments across mathematical benchmarks validate that our method improves both accuracy and reasoning quality over competitive baselines, while introducing less than 5% training overhead.

2 Evidence Gain as Quality Measure

This section formally defines **Evidence Gain**, a quality signal that measures reasoning quality by leveraging the policy model's intrinsic in-context learning ability, and validates it empirically.

Our basic idea is that, when used as demonstrations, high-quality reasoning traces provide more valuable problem-solving patterns, while low-quality reasoning (even with correct answers) provides less reference value due to flaws such as inconsistent logic (Li et al., 2025). This motivates us to quantify the quality of a solution by its teaching ability as a demonstration, formalized as follows.

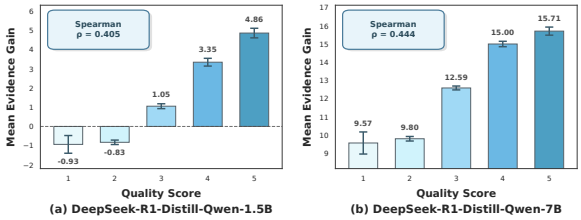

Figure 1: Mean Evidence Gain by quality score with 95% confidence intervals on two models.

2.1 Definition

Let π_θ denote policy model. Given a question q, a model-generated reasoning trace r, and a held-out validation set $\mathcal{E} = \{(e_q, e_r)\}$ composed of questions e_q and high-quality reference reasoning traces e_r, we define **Evidence Gain** as:

$$\Delta(q, r) = \mathbb{E}_{e \sim \mathcal{E}} \left[\log \pi_\theta(e_r | q, r, e_q) - \log \pi_\theta(e_r | e_q) \right] \quad (1)$$

Intuitively, Δ measures how much prepending the pair (q, r) improves the model's ability to generate reference solutions. Averaging over $\mathcal{E}$ ensures that high Δ reflects transferable reasoning patterns rather than spurious matches to any single sample.

2.2 Correlation with Reasoning Quality

We validate Evidence Gain on DeepSeek-R1-Distill-Qwen at 1.5B and 7B scales using KlearReasoner-MathSub-30K dataset (Su et al., 2025a). First, we sample 3,000 questions, generate 8 responses per question, and retain traces with correct final answers (12,251 for 1.5B and 16,910 for 7B). We then employ DeepSeek-V3.2 (DeepSeek-AI et al., 2025), a strong LLM-based evaluator, to assess the reasoning quality. The evaluator scores each solution across multiple dimensions including logical coherence and redundancy, and assigns an overall quality score on a 1–5 scale. Next, we sample 100 new questions to construct a held-out validation set $\mathcal{E}$. For each $e_q \in \mathcal{E}$, we generate a correct solution using DeepSeek-R1-0528 (Guo et al., 2025a). We leverage a feature of DeepSeek-R1: it first produces a draft chain-of-thought inside `<think> </think>`, and then outputs a more polished reasoning solution afterward. We treat the content following `</think>` as the high-quality reference trace e_r.

Figure 1 shows an interesting pattern: model ability determines the absolute baseline of Δ, while reasoning quality differentiates its relative magnitude. The 7B model, with stronger ICL ability, ex-

	Length	LogProb	MajorVote	Δ (Ours)
1.5B	−0.147	0.129	0.079	**0.405**
7B	−0.161	0.178	0.109	**0.444**

Table 1: Spearman correlation (ρ) between proxy signals and reasoning quality. Evidence Gain achieves stronger correlation than alternatives.

tracts useful information from any reasoning trace, resulting in uniformly positive Δ values greatly higher than those of 1.5B. However, the relative ordering remains consistent within each scale: high-quality traces yield higher Δ than low-quality ones. This relative difference is what matters for RL training. These results confirm that Evidence Gain effectively distinguishes reasoning quality. Human evaluation in Appendix C.3 supports these findings.

2.3 Comparison with Other Proxy Signals

To contextualize Evidence Gain, we compare it against three representative proxy signals of reasoning quality from prior work: response length (Xin et al., 2026), log-probability (Kadavath et al., 2022), and majority voting (Wang et al., 2023). Detailed descriptions of these signals are provided in Appendix C.2. Table 1 reports Spearman correlations with reasoning quality scores.

Evidence Gain achieves stronger correlation with reasoning quality than all three proxies. This is because Evidence Gain captures transferable problem-solving patterns through the model's ICL mechanism, encoding richer quality information than any single surface-level feature (e.g., reasoning length). This is further validated by our per-dimension analysis in Appendix C.1, which shows that Evidence Gain reflects multiple aspects of reasoning quality rather than a single dimension.

3 Implicit Reward Reweighting

The correlation between Evidence Gain (Δ) and reasoning quality (§2) suggests that upweighting high-Δ traces during training could improve reasoning. However, explicitly computing Δ for each rollout is prohibitively expensive.[1] In this section, we show that explicit computation is unnecessary. Evidence Gain can be seamlessly integrated into training through **In-Context RLVR**.

In-Context RLVR modifies standard RLVR by a simple input-side change. The core idea is to reverse the process: instead of computing Evidence Gain *after* generation to reweight rewards, we utilize the demonstration set to guide training *before* generation. Specifically, before each rollout, we sample a demonstration $e = (e_q, e_r)$ uniformly at random from the demonstration set $\mathcal{E}$, prepend it to the current question q, and generate response r from the demonstration-conditioned policy $\pi_\theta(r|e, q)$. Standard RL updates are then performed using correctness reward $R(q, r) \in \{0, 1\}$.

3.1 Theoretical Foundation

We establish that In-Context RLVR implicitly performs reward reweighting, where the reweighting factor is approximately proportional to $\exp(\Delta)$. We first define the In-Context RLVR objective:

$$J_{\text{IC}}(\theta) = \mathbb{E}_q \mathbb{E}_{e\sim\mathcal{E}, r\sim\pi_\theta(\cdot|e,q)}[R(q, r)]. \quad (2)$$

Our theoretical result proceeds in two steps (full derivation in Appendix E).

Step 1: Bayesian Identity. Since demonstrations are sampled independently from training questions, the demonstration question e_q alone carries no information relevant to solving q.[2] Under this assumption, the following Bayesian identity holds:

$$\pi_\theta(r|e, q) = \pi_\theta(r|q) \cdot \frac{\pi_\theta(e_r|q, r, e_q)}{\pi_\theta(e_r|e_q)}.$$

This identity shows that the conditioned policy equals the base policy multiplied by a likelihood ratio. Using this, we prove (Theorem E.3) that the In-Context RLVR objective can be rewritten as:

$$J(\theta) = \mathbb{E}_q \mathbb{E}_{r\sim\pi_\theta(\cdot|q)} \left[R(q, r) \cdot w(q, r)\right], \quad (3)$$

where $w(q, r) = \mathbb{E}_e[\exp(\Delta_e)]$ with $\Delta_e = \log \pi_\theta(e_r|q, r, e_q) - \log \pi_\theta(e_r|e_q)$.

Step 2: Log-Linear Relationship. The weight $w = \mathbb{E}_e[\exp(\Delta_e)]$ differs from $\exp(\Delta)$ where $\Delta = \mathbb{E}_e[\Delta_e]$ is the Evidence Gain. However, we further show (Theorem E.4) that

$$\log w(q, r) \approx \Delta(q, r) + c,$$

where c is a model-specific constant. This log-linear approximation is empirically tight: Pearson correlations between $\log w$ and Δ exceed 0.8 at both 1.5B and 7B scales (Figure 7), confirming that $w(q, r) \propto \exp(\Delta(q, r))$ in practice.

Combining these two steps, In-Context RLVR implicitly reweights rewards by a factor w approximately proportional to $\exp(\Delta)$, assigning higher weights to traces with higher Evidence Gain.

[1] Computing Δ for ∼12K samples over 100 demonstrations requires approximately 80 hours on H800.

[2] This assumption is empirically verified in Appendix E.2.

Method	AIME24	AIME25	HMMT25	MATH500	AMC23	Olympiad	Average	Time/Step (s)
DS-R1-Distill-Qwen-1.5B	29.2	24.1	13.1	86.0	73.7	51.8	46.3	–
+ GRPO (Shao et al., 2024)	33.4	28.1	16.6	88.3	79.3	56.2	50.3	457.4
+ IC-GRPO (Ours)	38.3 ↑	30.6 ↑	17.7 ↑	89.5 ↑	82.5 ↑	56.9 ↑	52.6 ↑	461.8
+ DAPO (Yu et al., 2025a)	40.0	28.4	19.2	90.0	84.4	61.6	53.9	459.6
+ CISPO (MiniMax et al., 2025)	32.9	25.1	13.2	85.8	80.9	54.9	48.8	466.3
+ GSPO (Zheng et al., 2025)	42.5	33.6	19.0	90.3	85.9	62.6	55.7	437.3
+ CE-GPPO (Su et al., 2025b)	42.8	32.5	20.5	91.0	85.8	61.8	55.7	464.0
+ IC-DAPO (Ours)	**45.6** ↑	**34.2** ↑	**19.7** ↑	**90.6** ↑	**86.2** ↑	**62.1** ↑	**56.4** ↑	477.2
DS-R1-Distill-Qwen-7B	54.5	39.1	26.2	93.6	90.6	67.0	61.8	–
+ GRPO (Shao et al., 2024)	55.3	40.3	24.5	93.7	88.8	65.6	61.4	305.6
+ DAPO (Yu et al., 2025a)	62.0	45.9	27.4	94.1	92.3	69.9	65.3	303.1
+ CE-GPPO (Su et al., 2025b)	64.2	**50.3**	28.9	95.3	93.3	71.6	67.3	292.5
+ IC-DAPO (Ours)	**66.5** ↑	49.8 ↑	**29.4** ↑	**95.6** ↑	**93.7** ↑	**71.7** ↑	**67.8** ↑	315.6

Table 2: Performance comparison across mathematical reasoning benchmarks. **Bold** and underline indicate the best and second-best results respectively. ↑ denotes improvement over the corresponding baseline (GRPO or DAPO). Notably, training times are incomparable across scales due to different GPU configurations (32 vs. 128 GPUs). Given this high GPU requirements at 7B scales, GSPO and CISPO are evaluated only at 1.5B.

3.2 Interpretation of Implicit Reweighting

The reweighted reward $R \cdot w$ in Eq. 3 implies a two-stage selection mechanism. First, the binary reward R filters out traces with incorrect answers. Second, among correct traces, the weight $w \propto \exp(\Delta)$ differentiates reasoning quality, assigning higher weights to high-quality traces and lower weights to low-quality ones. While training explicitly samples from $\pi_\theta(r|e, q)$, In-Context RLVR implicitly optimizes the base policy $\pi_\theta(\cdot|q)$ with rewards reweighted by Evidence Gain.

Eq. 3 shows that our method employs the model's own ICL ability to guide optimization, with the policy serving as both the learner and the implicit quality evaluator. A natural concern is whether Evidence Gain remains a valid quality signal as the policy evolves, since our validation in Section 2 uses a fixed model. We address this in Section 4, showing that the correlation between Evidence Gain and reasoning quality remains stable throughout training.

Notably, while J_{IC} and J share the same expectation, they differ in variance. The explicit reweighting term $w \propto \exp(\Delta)$ in J would introduce prohibitive reward variance. J_{IC} avoids this instability by shifting the sampling distribution directly.

4 Experiments

To validate our framework, we combine In-Context RLVR with DAPO (Yu et al., 2025a), yielding **IC-DAPO**. We choose DAPO as backbone because it is a widely adopted RLVR method whose key techniques (e.g., clip-higher) have been incorporated into many subsequent methods (Yue et al., 2025; Su et al., 2025b), making it a representative baseline for evaluating input-side modifications. All details of this section are provided in Appendix B.

4.1 Setup

Dataset. Our training data is derived from KlearReasoner-MathSub-30K (Su et al., 2025a), which contains 30K mathematical reasoning problems. We partition training data into three disjoint subsets: (1) a **training set** for policy optimization, (2) a **demonstration set** $\mathcal{E}$ containing 1,082 question-reasoning pairs used for demonstration during IC-DAPO training, and (3) a **held-out set** $\mathcal{E}_0$ of 100 additional examples reserved for the correlation analysis in §4.3. Both $\mathcal{E}$ and $\mathcal{E}_0$ are constructed following the procedure described in §2.

Baselines. We compare against several popular RLVR methods, including both standard outcome-based algorithms and more advanced objective-modifying variants. This selection directly tests whether our input-side modification can match algorithmic innovations in policy optimization. We exclude PRM-based methods as they require costly overhead that our method aims to avoid.

Training and Evaluation. We train DeepSeek-R1-Distill-Qwen at 1.5B and 7B scales. We conduct evaluations across various authoritative mathematical reasoning benchmarks, including AIME24, AIME25, HMMT25, MATH500 (Lightman et al., 2024), AMC23 and OlympiadBench (He et al., 2024). Following Su et al. (2025b), we report avg@4 scores on MATH500 and OlympiadBench, and avg@32 scores on all other benchmarks. Crucially, *all evaluation is conducted in zero-shot mode*, ensuring fair comparison with baselines.

4.2 Main Results

Table 2 presents benchmark performance. IC-DAPO outperforms DAPO by +2.5 average points at both scales, with gains particularly pronounced on competition benchmarks: +5.6 on AIME24 and +5.8 on AIME25 for the 1.5B model. This supports our hypothesis that implicit quality reweighting helps more on challenging problems where correct-but-low-quality traces are most harmful, which is further validated in Section 4.3. To verify that this improvement generalizes beyond DAPO, we also apply In-Context RLVR to GRPO at 1.5B scale. IC-GRPO achieves +2.3 average improvement over GRPO across all benchmarks at this scale, confirming that the implicit reweighting mechanism transfers across different RL optimizers.

Beyond improvements over the corresponding baselines, IC-DAPO also matches or exceeds methods that modify the RL objective (e.g., GSPO, CISPO), achieving the highest average score at both scales while only altering the *input distribution*. This suggests that input-side modification constitutes an improvement axis orthogonal to policy optimization algorithms. We further compare wall-clock training time per training step and find that IC-DAPO incurs slight overhead ($<5\%$), confirming its practicality.

4.3 Analysis

Our theory (§3) predicts that In-Context RLVR implicitly upweights high-Δ traces. To verify this, we track training dynamics by computing Δ on the held-out set $\mathcal{E}_0$ and assessing reasoning quality via Deepseek-V3.2, following procedures in §2.

Q1: Does implicit reweighting occur? Figure 2 (left) shows that mean Evidence Gain increases steadily under IC-DAPO throughout training, while DAPO exhibits smaller and slower growth. This confirms that the conditioned objective steers the policy toward traces with higher demonstration utility, exactly as predicted.

Q2: Does this improve reasoning quality? Figure 2 (middle) shows higher Δ corresponds to improved quality scores. Note that $\mathcal{E}_0$ used for evaluation is disjoint from $\mathcal{E}$ used for training, ensuring unbiased evaluation. This shows that by upweighting traces with high teaching utility, we encourage better reasoning rather than merely correct answers.

Q3: Is Evidence Gain valid throughout training? Finally, we address the concern from §3.

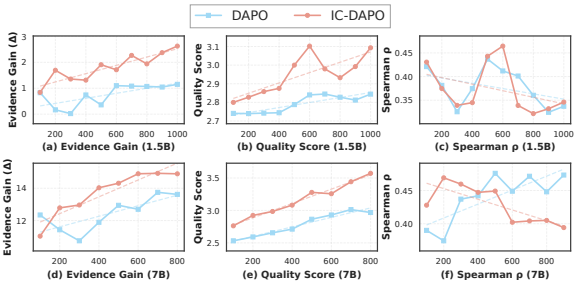

Figure 2: Training dynamics of Evidence Gain, quality score, and their correlation (ρ) across training steps.

Method	Scale	Easy	Medium	Hard
DAPO	1.5B	98.3	90.1	23.1
IC-DAPO	1.5B	98.8 (+0.5%)	93.5 (+3.8%)	26.0 (**+12.6%**)
DAPO	7B	98.6	97.8	39.2
IC-DAPO	7B	99.3 (+0.7%)	98.2 (+0.4%)	43.2 (**+10.2%**)

Table 3: Difficulty-stratified analysis. Parenthetical values denote relative improvement over DAPO.

Figure 2 (right) shows that the Spearman correlation between Δ and quality remains stable (around $\rho \approx 0.4$) across training steps, confirming that the policy model's intrinsic ICL signal remains a robust quality indicator as training progresses.

Q4: Where do the gains come from? To examine whether quality-based reweighting benefits problems that demand higher reasoning quality, we stratify evaluation by difficulty. We rank all problems from 6 benchmarks in Table 2 by backbone model accuracy and divide them into equal thirds. Table 3 shows that gains concentrate on hard problems: +12.6% relative improvement at 1.5B and +10.2% at 7B. On easier problems where baseline accuracy is near-perfect, there is limited room for quality-based reweighting to contribute. This pattern confirms that implicit quality reweighting provides the greatest advantage precisely where reasoning quality matters most.

5 Conclusions

We introduce Evidence Gain, a quality signal that measures reasoning quality based on the policy model's intrinsic ICL ability. To leverage this signal, we propose In-Context RLVR, which prepends demonstrations during training and implicitly reweights rewards by Evidence Gain to encourage high-quality traces. Experiments confirm improvements in both accuracy and reasoning quality over standard RLVR, providing a practical approach to improve reasoning quality in RLVR.

Limitations

This work has two main limitations. First, although In-Context RLVR demonstrates consistent improvements across mathematical reasoning benchmarks, its generalization to other reasoning-intensive domains such as STEM problem-solving remains an open question due to computational constraints. Second, constructing the demonstration set requires access to a strong model (e.g., DeepSeek-R1) for generating high-quality reference traces. Alternative construction strategies that reduce this dependency should be further developed.

References

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *Preprint*, arXiv:2505.16400.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *Preprint*, arXiv:2505.22617.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Heejin Do, Jaehui Hwang, Dongyoon Han, Seong Joon Oh, and Sangdoo Yun. 2025. What defines good reasoning in llms? dissecting reasoning steps with multi-aspect evaluation. *Preprint*, arXiv:2510.20603.

Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. *Preprint*, arXiv:2210.10760.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *ICLR 2023*. OpenReview.net.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025a. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nat.*, 645(8081):633–638.

Jiaxing Guo, Wenjie Yang, Shengzhong Zhang, Tongshan Xu, Lun Du, Da Zheng, and Zengfeng Huang. 2025b. Right is not enough: The pitfalls of outcome supervision in training llms for math reasoning. *Preprint*, arXiv:2506.06877.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. 2025. Skywork open reasoner 1 technical report. *Preprint*, arXiv:2505.22312.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. *Preprint*, arXiv:2310.01798.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, and Oyvind Tafjord. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *Preprint*, arXiv:2502.07374.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. Numinamath. `https://huggingface.co/AI-MO/NuminaMath-CoT`. Technical report available at `https://github.com/`

project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. In-context learning with retrieved demonstrations for language models: A survey. *Preprint*, arXiv:2401.11624.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. Notion Blog.

Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, and Carson Denison. 2025. Natural emergent misalignment from reward hacking in production rl. *Preprint*, arXiv:2511.18397.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

MiniMax, :, Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, Chengjun Xiao, Chengyu Du, Chi Zhang, Chu Qiao, Chunhao Zhang, Chunhui Du, Congchao Guo, and 109 others. 2025. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *Preprint*, arXiv:2506.13585.

Youssef Mroueh. 2025. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *Preprint*, arXiv:2503.06639.

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

Zhenpeng Su, Leiyu Pan, Xue Bai, Dening Liu, Guanting Dong, Jiaming Huang, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. 2025a. Klear-reasoner: Advancing reasoning capability via gradient-preserving clipping policy optimization. *Preprint*, arXiv:2508.07629.

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. 2025b. Ce-gppo: Coordinating entropy via gradient-preserving clipping policy optimization in reinforcement learning. *Preprint*, arXiv:2509.20712.

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Tiehua Mei, Zijia Lin, Yuntao Li, Wenping Hu, Ruiming Tang, Kun Gai, and Guorui Zhou. 2025c. Entropy ratio clipping as a soft global constraint for stable reinforcement learning. *Preprint*, arXiv:2512.05591.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *Preprint*, arXiv:2312.08935.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 27723–27730. AAAI Press.

Rihui Xin, Han Liu, Zecheng Wang, Yupeng Zhang, Dianbo Sui, Xiaolin Hu, and Bingning Wang. 2026. Surrogate signals from format and length: Reinforcement learning for solving mathematical problems without ground truth answers. *Preprint*, arXiv:2505.19439.

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang. 2025. Self-rewarding correction for mathematical reasoning. *Preprint*, arXiv:2502.19613.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Chenlu Ye, Zhou Yu, Ziji Zhang, Hao Chen, Narayanan Sadagopan, Jing Huang, Tong Zhang, and Anurag Beniwal. 2025. Beyond correctness: Harmonizing process and outcome rewards through rl training. *Preprint*, arXiv:2509.03403.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. Flask: Fine-grained language model evaluation based on alignment skill sets. *Preprint*, arXiv:2307.10928.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, and Tiantian Fan. 2025a. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. 2025b. Rlpr: Extrapolating rlvr to general domains without verifiers. *Preprint*, arXiv:2506.18254.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, and 8 others. 2025. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *Preprint*, arXiv:2504.05118.

Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. 2025a. Consistent paths lead to truth: Self-rewarding reinforcement learning for llm reasoning. In *Advances in Neural Information Processing Systems*.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025b. The lessons of developing process reward models in mathematical reasoning. *Preprint*, arXiv:2501.07301.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

A Related Work

Reinforcement Learning with Verifiable Rewards. RLVR has become a dominant paradigm for improving LLM reasoning (Lambert et al., 2025; Guo et al., 2025a). By using rule-based correctness signals, RLVR avoids reward hacking and scales well. GRPO (Shao et al., 2024) optimizes policies via group-based reward normalization without requiring a separate critic model. Subsequent work addresses entropy collapse (Yu et al., 2025a), sequence-level optimization (Zheng et al., 2025), and training stability (MiniMax et al., 2025). Despite the advances, a fundamental limitation persists. Binary correctness rewards assign equal reward to correct solutions regardless of reasoning quality, potentially reinforcing spurious traces that get correct answers through flawed logic (Zhang et al., 2025b; MacDiarmid et al., 2025).

Process Reward Models. PRMs address this limitation by providing step-level feedback. Lightman et al. (2024) demonstrate substantial gains from process supervision over outcome supervision, though their approach required approximately 800,000 human-annotated step labels. Automated alternatives such as Math-Shepherd (Wang et al., 2024) construct step-level labels through Monte Carlo estimation, trading annotation cost for computational overhead from repeated rollouts. These methods share a common constraint. Obtaining reliable process signals demands either substantial human effort or significant compute resources.

Quality-Aware Reward Reweighting. Standard RLVR assigns equal reward to all correct traces regardless of reasoning quality. Several approaches address this limitation through reward reweighting, which we categorize into three families.

External reward models (e.g., PRM (Zhang et al., 2025b)) train a separate model to score reasoning quality at each step. However, this requires enormous annotation effort, adds per-step evaluation cost throughout training, and remains susceptible to reward hacking (Gao et al., 2022).

Self-evaluation methods (e.g., Self-Rewarding (Yuan et al., 2024), RLPR (Yu et al., 2025b)) let the model judge the quality of its own output. This avoids an external model but introduces different costs: each quality assessment requires additional forward passes during training. Moreover, without reliable external guidance, the model may reinforce its own errors through self-consistent illusions (Huang et al., 2024).

Proxy signals (e.g., length (Xin et al., 2026), log-probability (Kadavath et al., 2022), majority voting (Wang et al., 2023)) are lightweight but capture only a single surface dimension of quality. Majority voting, for instance, measures only final-answer consistency and cannot distinguish rigorous reasoning from lucky guessing.

Our method, Evidence Gain, belongs to the proxy signal family. The key difference from existing proxy signals is that Evidence Gain integrates multiple quality dimensions into a single scalar through the model's ICL mechanism, rather than relying on one surface feature. We quantitatively compare Evidence Gain against representative proxy signals in Section 2.

Self-Derived Reasoning Quality Signals. An emerging direction leverages signals derived from the model itself rather than external supervision.

Zhang et al. (2025a) construct intrinsic rewards from trajectory consistency and volatility, which requires computing distances between each intermediate state and all distinct final answers across sampled rollouts, incurring $O(NTK)$ additional forward passes per prompt beyond standard algorithm. Xiong et al. (2025) train models to perform iterative self-correction loops, where the model sequentially detects errors, revises outputs, and decides when to terminate. In contrast, our method measures reasoning quality through demonstration utility, motivated by findings that high-quality reasoning traces can serve as effective in-context demonstrations (Min et al., 2022; Li et al., 2025). Crucially, this signal can be integrated implicitly into the training objective via In-Context RLVR, adding less than 5% overhead without any explicit quality computation. Importantly, our approach differs fundamentally from demonstration selection methods (Luo et al., 2024), which develop retrieval or optimization strategies to identify the best demonstrations for each input query. In contrast, we leverage demonstrations to assess the quality of candidate reasoning traces; since Evidence Gain (Eq. 1) is defined as an average over the validation set, all demonstrations are equally important for every query in our method.

B Experimental Details

B.1 Datasets and Preprocessing

Our training data is derived from KlearReasoner-MathSub-30K (Su et al., 2025a), which contains approximately 30K high-quality mathematical reasoning problems collected from several curated sources, including Skywork-OR1 (He et al., 2025), Acereason (Chen et al., 2025), NuminaMath (LI et al., 2024), and DeepScaleR (Luo et al., 2025). To mitigate potential data contamination, the dataset has been processed with 9-gram deduplication against the evaluation benchmarks.

We first randomly partition 28k samples from the full dataset for policy optimization. For the remaining 2k samples, we generate reasoning traces using DeepSeek-R1-0528 (Guo et al., 2025a) and filter the outputs using rule-based validators to retain only those with correct final answers, yielding approximately 1,200 valid examples. From this filtered set, we randomly select 100 examples to form the held-out set $\mathcal{E}_0$ for the correlation analysis in Section 4.3, with the remaining 1,082 examples forming the demonstration set $\mathcal{E}$ used for prepending demonstrations during training. For each example in $\mathcal{E}$ and $\mathcal{E}_0$, we extract the content following `</think>` as the reference reasoning trace following the procedure described in Section 2. We manually verify the quality of these solutions.

Demonstration Pipeline Cost. The demonstration construction pipeline is lightweight and reproducible. The entire process involves 2,000 API calls to DeepSeek-R1, correctness verification via math-verify, and extraction of the content following `</think>`. This takes approximately one hour at a total API cost of approximately $50. This is a one-time, pre-training cost with zero expense during training or inference. For comparison, training a PRM requires many step-level annotations. Scaling to new domains requires only replacing the source problems, with no changes to the pipeline itself.

B.2 Details for Main Experiments

We train DeepSeek-R1-Distill-Qwen-1.5B[3] and DeepSeek-R1-Distill-Qwen-7B[4]. We evaluate on six mathematical reasoning benchmarks: AIME24, AIME25, HMMT25, MATH500, AMC23, and OlympiadBench. For evaluation metrics, we report avg@4 scores on MATH500 (Lightman et al., 2024) and OlympiadBench (He et al., 2024), and avg@32 scores on all other benchmarks, following prior work (Su et al., 2025b). At inference, we set the maximum generation length to 32k tokens for AIME24 and AIME25, and 16k tokens for the other datasets. For answer extraction, we follow the standard practice adopted in Yang et al. (2024): parsing the contents enclosed within the `\boxed{}` structure in model outputs to identify the final answer. Answer correctness is judged by math-verify[5], which performs symbolic comparison to handle equivalent mathematical expressions.

All evaluation is conducted in zero-shot mode without any demonstrations, ensuring fair comparison with baseline methods and validating that our approach requires no modification to deployment. This zero-shot evaluation also empirically confirms our theoretical claim in Section 3: while demonstrations are used during training to enable implicit quality reweighting, In-Context RLVR implicitly optimizes the base policy which can operate without demonstrations.

[3] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

[4] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

[5] https://github.com/huggingface/Math-Verify

B.3 Details for Training Dynamics Analysis

Theoretical analysis in Section 3 shows that In-Context RLVR implicitly reweights rewards by Evidence Gain, with the reweighted reward $R \cdot w$ implying a two-stage selection mechanism: first, the binary reward R filters out traces with incorrect answers ($R = 0$); second, among correct traces ($R = 1$), the weight $w \propto \exp(\Delta)$ differentiates reasoning quality. Therefore, Evidence Gain is designed to distinguish reasoning quality *among correct solutions*. To validate this implicit reweighting mechanism empirically (Section 4.3), we track both Evidence Gain and reasoning quality scores exclusively on traces with correct final answers.

Specifically, we use checkpoints from both DAPO and IC-DAPO at 1.5B and 7B scales. Every 100 training steps, we randomly sample 100 queries from the training set. Importantly, at each chosen step, DAPO and IC-DAPO share the same set of sampled queries to ensure fair comparison. For each query, we generate 8 rollouts and retain only those with correct final answers. We then compute Evidence Gain on the held-out set $\mathcal{E}_0$, and assess reasoning quality using DeepSeek-V3.2 as described in Appendix B.4. Finally, we compute the Spearman correlation ρ between Evidence Gain and quality scores to verify that Evidence Gain remains a valid quality signal throughout training.

B.4 LLM-based Quality Evaluation

To automatically assess reasoning quality, we employ DeepSeek-V3.2 (DeepSeek-AI et al., 2025) as an LLM-based evaluator. Our evaluation rubric is informed by prior work on solution quality assessment (Ye et al., 2024; Xia et al., 2025; Golovneva et al., 2023). To ensure comprehensive coverage, we define eight complementary dimensions as shown in Table 4. For each reasoning trace, the evaluator assigns a score from 1 to 5 on each dimension along with explicit textual explanations for justification, and finally provides an overall quality score from 1 to 5. The complete prompt template is provided in Appendix D.

B.5 Implementation of IC-DAPO

Following standard DAPO, we sample 8 rollouts per training question. To integrate In-Context RLVR, for m of these rollouts, we independently sample a random demonstration from $\mathcal{E}$ and prepend it to the question before generation; the remaining $n = 8 - m$ rollouts are generated from the original question. All 8 rollouts are then scored and updated using DAPO objective. In our experiments, we set $m = 6$ and $n = 2$. This mixed sampling strategy serves two purposes: (1) it preserves sufficient reward variance within each group for DAPO's group-based normalization to compute meaningful advantages, and (2) it increases the diversity of input contexts within each training batch. Our experiments validate the effectiveness of this configuration. Practitioners may explore alternative values of m to adapt to different tasks.

Dimension	Definition
Repetition	Same steps or ideas repeated
Redundancy	Unnecessary or verbose content
Logical Consistency	Contradictions
Relevance	Off-topic exploration
CoT-Ans Alignment	Answer derived from reasoning
Reasoning Rigor	Claims justified without leaps
Clarity	Easy to follow, well-structured
Completeness	All necessary steps present

Table 4: Quality evaluation dimensions.

B.6 Implementation of Baselines

GRPO optimizes policies via group-based reward normalization without requiring a separate critic model. Following Shao et al. (2024), we adopt symmetric clipping bounds with $\epsilon = 0.2$.

DAPO extends GRPO by introducing asymmetric clipping bounds and dynamic sample filtering to mitigate entropy collapse. Following Yu et al. (2025a), we set the lower and upper clipping thresholds to $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$, respectively.

CISPO applies clipping directly to the importance sampling weights rather than to the final policy update. Following Cui et al. (2025), we set symmetric clipping bounds with $\epsilon = 0.2$.

GSPO employs a sequence-level importance ratio to enhance training stability and scalability. Following Zheng et al. (2025), we set the lower and upper clipping thresholds to $\epsilon_{\text{low}} = 0.0003$ and $\epsilon_{\text{high}} = 0.0004$, respectively.

CE-GPPO reintroduces gradient signals from tokens outside the clipping interval in a bounded manner through a stop-gradient operation, enabling fine-grained control over policy entropy dynamics. We directly report results from the original paper (Su et al., 2025b). As this work provides two sets of evaluation results under different configurations ($\beta_1 = 0.5, \beta_2 = 1$ and $\beta_1 = 0.75, \beta_2 = 1$),

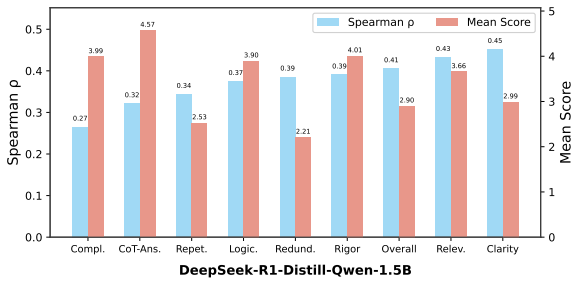

Figure 3: Spearman correlation between Evidence Gain and individual quality dimensions, alongside mean scores for each dimension. Evaluated on DeepSeek-R1-Distill-Qwen-1.5B with the same setup as Section 2.

we report their average scores in Table 2.

C Additional Experiments

C.1 What Does Evidence Gain Capture?

To investigate which aspects of reasoning quality Evidence Gain most effectively captures, we conduct a fine-grained correlation analysis across the eight dimensions defined in our evaluation rubric (Appendix B.4). Figure 3 presents results on DeepSeek-R1-Distill-Qwen-1.5B. We observe that all correlations are positive, ranging from $\rho = 0.27$ to $\rho = 0.45$, suggesting that Evidence Gain reflects multiple aspects of reasoning quality rather than a single dimension.

Notably, Clarity ($\rho = 0.45$) and Relevance ($\rho = 0.43$) show stronger correlations than Completeness ($\rho = 0.27$) and CoT-Answer Alignment ($\rho = 0.32$). This pattern can be explained by our setup. Since we only evaluate traces that arrive at correct answers (Appendix B), these traces are already filtered for answer validity. As shown by the mean scores in Figure 3, Completeness (mean 3.99) and CoT-Answer Alignment (mean 4.57) exhibit high scores with limited variance among correct traces, because reaching the right answer typically requires including necessary steps and properly deriving the conclusion. In contrast, Clarity (mean 2.99) and Relevance (mean 3.66) show lower mean scores even among correct solutions, since a trace can reach the right answer while still being poorly organized. This difference in variance explains why Evidence Gain shows stronger correlations with dimensions that have greater room to discriminate among correct traces.

C.2 Comparison with Proxy Signals

This section provides detailed descriptions of the three proxy signals compared with Evidence Gain in Section 2.

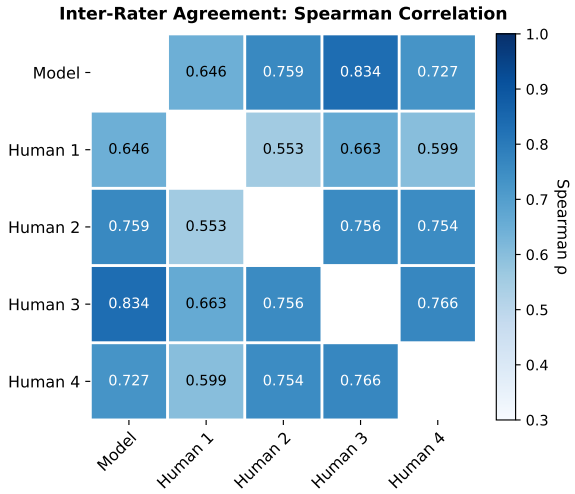

Figure 4: Spearman correlation matrix between DeepSeek-V3.2 quality scores and four human expert ratings on 100 sampled reasoning traces, together with inter rater correlations among experts.

- **Length**: token count of the reasoning trace.
- **LogProb**: log-probability of rollouts, reflecting the model's confidence in its own output.
- **MajorVote**: a binary indicator of whether the answer matches the majority answer across rollouts, measuring answer-level consistency.

C.3 Human Evaluation

We further validate whether the automatic quality scores produced by DeepSeek-V3.2 are consistent with human judgement. We first randomly sample 100 question-reasoning pairs generated by Deepseek-R1-Distill-Qwen-1.5B whose final answers are verified correct by rule. We then ask **four** human experts to separately assign an overall quality score between 1 and 5 for each reasoning trace. Annotators follow the same rubric for Deepseek-V3.2 (Appendix B.4). Finally, we compute Spearman correlation coefficients between the DeepSeek-V3.2 scores and each human expert, as well as the correlations among the human experts.

Figure 4 shows strong agreement between DeepSeek-V3.2 and human experts. The correlations between DeepSeek-V3.2 and individual experts fall between 0.65 and 0.83, with an average of 0.74, which is comparable to human expert agreement. Overall, DeepSeek-V3.2 performs within the variability of human judgement on this task,

Method	AIME24	AIME25	HMMT25	MATH500	AMC23	Olympiad	Average
DS-R1-Distill-Qwen-1.5B	29.2	24.1	13.1	86.0	73.7	51.8	46.3
+ DAPO (Yu et al., 2025a)	40.0	28.4	19.2	90.0	84.4	61.6	53.9
+ IC-DAPO (V3.1)	44.5 ↑	32.3 ↑	19.5 ↑	90.3 ↑	85.8 ↑	61.7 ↑	55.7 ↑
+ IC-DAPO (R1)	**45.6** ↑	**34.2** ↑	**19.7** ↑	**90.6** ↑	**86.2** ↑	**62.1** ↑	**56.4** ↑
DS-R1-Distill-Qwen-7B	54.5	39.1	26.2	93.6	90.6	67.0	61.8
+ DAPO (Yu et al., 2025a)	62.0	45.9	27.4	94.1	92.3	69.9	65.3
+ IC-DAPO (V3.1)	63.3 ↑	47.5 ↑	29.2 ↑	95.5 ↑	92.6 ↑	70.8 ↑	66.4 ↑
+ IC-DAPO (R1)	**66.5** ↑	**49.8** ↑	**29.4** ↑	**95.6** ↑	**93.7** ↑	**71.7** ↑	**67.8** ↑

Table 5: Ablation on demonstration quality. IC-DAPO (R1) uses refined reasoning traces from DeepSeek-R1, while IC-DAPO (V3.1) uses solutions from DeepSeek-V3.1, a non-reasoning model. Both variants outperform DAPO, with R1 demonstrations yielding stronger results. **Bold** and underline indicate best and second-best results. ↑ denotes improvement over DAPO.

Method	Scale	HumanEval	LCB	IFBench	IFEval	MMLU
DAPO	1.5B	73.6	30.4	9.1	38.3	46.6
IC-DAPO	1.5B	**75.5**	**31.9**	**10.7**	**40.6**	**48.3**
DAPO	7B	90.5	48.5	13.3	56.7	65.7
IC-DAPO	7B	**92.3**	**49.8**	**14.2**	**58.6**	**67.3**

Table 6: Cross-domain zero-shot evaluation on code generation (HumanEval, LCB), instruction following (IFBench, IFEval), and general knowledge (MMLU).

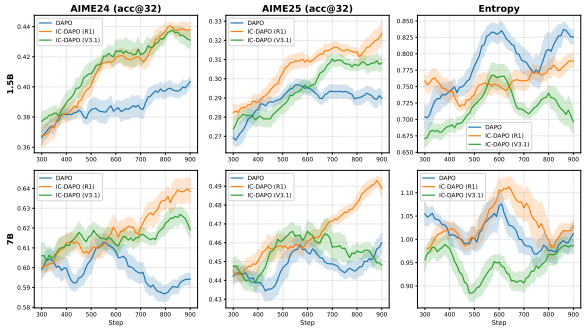

Figure 5: Training dynamics across 1.5B and 7B models. IC-DAPO variants consistently outperform DAPO on AIME24 and AIME25 while maintaining stable entropy throughout training.

supporting its reliability as an automatic evaluator in our experiments. We attribute this robustness to the comprehensive and fine-grained quality rubric, which decomposes reasoning quality into multiple distinguishable dimensions and thus enables more consistent judgements across different raters.

C.4 Cross-Domain Generalization

While our training focuses on mathematical reasoning, the theoretical mechanism is domain-agnostic. To test cross-domain generalization, we evaluate IC-DAPO and DAPO checkpoints on 5 additional benchmarks spanning code generation (HumanEval, LiveCodeBench), instruction following (IFBench, IFEval), and general knowledge (MMLU) in zero-shot mode. Table 6 shows that IC-DAPO consistently outperforms DAPO across all domains and both scales, suggesting that the benefits of quality-aware reweighting are not confined to mathematical reasoning.

C.5 Ablation Study on Demonstration Quality

Our main experiments in Section 4 construct the demonstration set $\mathcal{E}$ using the refined content following DeepSeek-R1's `</think>` tag. To investigate whether demonstration quality affects training outcomes, we construct an alternative set using DeepSeek-V3.1 (DeepSeek-AI, 2024), a strong non-reasoning model, for comparison. Without the reasoning capabilities that produce R1's refined traces, V3.1 is expected to generate solutions of lower quality, making it a suitable baseline for studying the effect of demonstration quality. To ensure fair comparison, we use the same 1,082 questions from the demonstration set $\mathcal{E}$ described in Section 4 and generate solutions using DeepSeek-V3.1. We generate multiple responses per question to ensure each question obtains a correct solution.

Table 5 shows that both IC-DAPO variants outperform DAPO, confirming In-Context RLVR's robustness over different demonstration sources. Crucially, IC-DAPO (R1) always surpasses IC-DAPO (V3.1), with gains of +1.1 on AIME24 and +1.9 on AIME25 at 1.5B scale. From a theoretical perspective, since the reference solutions in validation set are used to compute Evidence Gain (Eq. 1), the quality of these references may affect the accuracy of Δ as a quality signal, which in turn affects the effectiveness of implicit reweighting (Eq. 3).

C.6 Training Dynamics

We present extended training dynamics for both 1.5B and 7B models in Figure 5. Across both model scales, IC-DAPO variants (i.e., IC-DAPO (R1) and

IC-DAPO (V3.1)) consistently achieve higher accuracy on AIME24 and AIME25 throughout training while maintaining comparable entropy trajectories to the DAPO baseline. Notably, IC-DAPO (R1) demonstrates a clear advantage over standard DAPO, with the performance gap widening as training progresses. Comparing IC-DAPO (R1) and IC-DAPO (V3.1), we observe that R1-generated demonstrations yield better final performance, consistent with our findings in Section C.5 that higher-quality demonstrations lead to improved training outcomes. The entropy curves remain stable across all methods, indicating that In-Context RLVR does not compromise training stability while achieving superior accuracy, which aligns with the stability analysis in prior work (Yu et al., 2025a).

D Quality Evaluation Prompt

We present the prompt template used for reasoning quality evaluation with DeepSeek-V3.2 in Figure 6. The prompt assesses reasoning traces across eight dimensions, including Repetition, Redundancy, Logical Consistency, Relevance, CoT-Answer Alignment, Reasoning Rigor, Clarity, and Completeness, with each scoring on a 1–5 scale. The template uses placeholders `{question}` and `{response}` which are populated with the specific math problem and corresponding reasoning trace during evaluation.

```
# Mathematical Reasoning Quality Evaluation

## Task
Evaluate the **quality** of mathematical reasoning (answer already verified correct).
Assess whether this reasoning would be valuable as a learning reference.

**Key insight**: Correct answer ≠ good reasoning. Watch for warning signs:
- Frequent "wait", "hold on", "let me try again"
- Same calculation repeated multiple times
- Long explorations that don't contribute to the answer
- Answer appearing without clear derivation

Be strict. Reasoning that wanders, second-guesses itself, or reaches the right answer through
    messy exploration is not high-quality, regardless of correctness.

## Dimensions (1-5 each)

| Dimension           | Assess                                        |
|---------------------|-----------------------------------------------|
| Repetition          | Same steps or ideas repeated?                 |
| Redundancy          | Unnecessary or verbose content?               |
| Logical Consistency | Contradictions or backtracks?                 |
| Relevance           | Off-topic content or dead-end exploration?    |
| CoT-Answer Alignment| Answer clearly derived from reasoning?        |
| Reasoning Rigor     | All claims justified without leaps?           |
| Clarity             | Easy to follow and well-structured?           |
| Completeness        | All necessary steps present?                  |

## Scoring

| Score | Meaning                                                |
|-------|--------------------------------------------------------|
| 5     | Excellent: Perfect textbook quality                    |
| 4     | Good: Minor flaws, suitable as reference               |
| 3     | Average: Clear flaws, but followable                   |
| 2     | Weak: Major issues hurting pedagogical value           |
| 1     | Poor: Guessed answer, chaotic flow                     |

## Evaluate This

### Math Problem:
{question}

### Reasoning Process (Answer Verified Correct):
{response}

---

Output JSON only:
{
  "dimensions": {
    "repetition": {"score": <1-5>, "comment": "<evidence>"},
    "redundancy": {"score": <1-5>, "comment": "<evidence>"},
    "logical_consistency": {"score": <1-5>, "comment": "<evidence>"},
    "relevance": {"score": <1-5>, "comment": "<evidence>"},
    "cot_answer_alignment": {"score": <1-5>, "comment": "<evidence>"},
    "reasoning_rigor": {"score": <1-5>, "comment": "<evidence>"},
    "clarity": {"score": <1-5>, "comment": "<evidence>"},
    "completeness": {"score": <1-5>, "comment": "<evidence>"}
  },
  "overall_analysis": "<key findings, limiting factors>",
  "score": <1-5>
}
```

Figure 6: Quality Evaluation Prompt Template. Placeholders {question} and {response} are replaced with the actual math problem and reasoning trace during evaluation.

E Proofs of Main Results

[This section of appendix uses single-column format for mathematical readability.]

This appendix provides complete derivations for the theoretical results in Section 3. We first establish notation, then derive the reweighting identity and characterize its relationship to Evidence Gain.

E.1 Notations

We formalize the training setup. Let $q \sim \mathcal{D}$ denote a question from the training distribution. Let $\mathcal{E} = \{e^{(i)}\}_{i=1}^{|\mathcal{E}|}$ be the held-out validation set, where each demonstration $e = (e_q, e_r)$ consists of a question e_q and a high-quality reference reasoning trace e_r. During In-Context RLVR training, a demonstration e is sampled from $\mathcal{E}$ and prepended to q, after which the model generates a reasoning trace $r \sim \pi_\theta(\cdot|e, q)$.

E.2 Bayesian Identity

We establish the key identity relating the conditioned policy to the base policy. The derivation relies on the following assumption, which reflects the independent sampling structure in our data construction.

Assumption E.1. *Providing only the demonstration question e_q, without its reasoning trace e_r, does not alter the distribution over reasoning traces for a training question q. Conversely, providing only the training question q, without any reasoning trace, does not alter the distribution over reasoning traces for the demonstration question e_q. Formally:*

$$\pi_\theta(r|e_q, q) = \pi_\theta(r|q), \tag{A1}$$

$$\pi_\theta(e_r|q, e_q) = \pi_\theta(e_r|e_q). \tag{A2}$$

Remark. This assumption is natural given the independent sampling of demonstrations and training questions. To see why (A1) holds, suppose e_q is "Solve $x^2 - 5x + 6 = 0$" and q is "Compute $\int \sin x \, dx$." The bare statement of e_q carries no information about integration techniques; it only indicates that the context involves math. Crucially, the model already knows this from observing q itself. Thus, conditioning on e_q alone provides no additional signal for solving q. A symmetric argument establishes (A2). While edge cases may exist where two questions happen to share methodological structure, so that e_q could bias the preferred style or method for solving q, such coincidences are rare under independent sampling and average out at scale. Thus (A1) and (A2) hold as statistical approximations.

Empirical Validation. We validate Assumption E.1 using DeepSeek-R1-Distill-Qwen-1.5B. We randomly select 100 question-reasoning pairs (q, r) from rollouts generated during training, and independently sample 100 additional questions $\{q'\}$ from the dataset. We prepend each q' to each (q, r) resulting in 10,000 samples (S). We measure the relative change in reasoning log-probability when prepending an randomly select question q':

$$\delta = \frac{1}{|S|} \sum_{(q,r,q') \in S} \frac{|\log \pi_\theta(r|q', q) - \log \pi_\theta(r|q)|}{|\log \pi_\theta(r|q)|}. \tag{4}$$

Notably, we obtain $\boldsymbol{\delta = 0.0384} < 5\%$, confirming that independently sampled question statements have negligible influence on reasoning distributions on average.

Although Assumption E.1 suggests that question statements alone provide negligible cross example influence, a complete demonstration (e_q, e_r) *does* provide transferable information: the reasoning trace e_r may exhibit problem-solving patterns (e.g., algebraic manipulation, problem decomposition) that generalize across problems. This distinction is precisely what Evidence Gain captures, and it explains why the policy model's ICL ability can serve as an effective quality signal.

Lemma E.2 (Bayesian Identity)**.** *Under Assumption E.1, the conditioned policy admits the decomposition:*

$$\pi_\theta(r|e, q) = \pi_\theta(r|q) \cdot \frac{\pi_\theta(e_r|q, r, e_q)}{\pi_\theta(e_r|e_q)}. \tag{5}$$

Proof. We begin by expanding the conditioned policy $\pi_\theta(r|e_q, e_r, q)$ using the definition of conditional probability. By Bayes' Rule, we have:

$$\pi_\theta(r|e_q, e_r, q) = \frac{\pi_\theta(r|e_q, q)\,\pi_\theta(e_r|q, r, e_q)}{\pi_\theta(e_r|q, e_q)}. \tag{6}$$

We now apply Assumption E.1 to simplify the right-hand side. By (A1), the numerator term $\pi_\theta(r|e_q, q)$ reduces to $\pi_\theta(r|q)$, since observing e_q alone provides no additional information for generating r. By (A2), the denominator term $\pi_\theta(e_r|q, e_q)$ reduces to $\pi_\theta(e_r|e_q)$, since observing q alone provides no additional information for generating e_r. Substituting these simplifications into Eq. (6), we obtain:

$$\pi_\theta(r|e_q, e_r, q) = \pi_\theta(r|q) \cdot \frac{\pi_\theta(e_r|q, r, e_q)}{\pi_\theta(e_r|e_q)}. \tag{7}$$

This completes the proof. □

E.3 Implicit Reward Reweighting

We now establish that In-Context RLVR implicitly performs reward reweighting, showing how the policy model's ICL mechanism naturally upweights high-quality reasoning traces. We present the theoretical results in two separate theorems.

Theorem E.3 (Implicit Reweighting)**.** *Under Assumption E.1, the In-Context RLVR objective*

$$J_{\text{IC}}(\theta) = \mathbb{E}_q \mathbb{E}_{e\sim\mathcal{E},\, r\sim\pi_\theta(\cdot|e,q)}\big[R(q, r)\big] \tag{8}$$

can be exactly rewritten as

$$J_{\text{IC}}(\theta) = \mathbb{E}_q \mathbb{E}_{r\sim\pi_\theta(\cdot|q)}\big[R(q, r) \cdot w(q, r)\big], \tag{9}$$

where the weight factor is defined as

$$w(q, r) = \mathbb{E}_{e\sim\mathcal{E}}[\exp(\Delta_e)] \tag{10}$$

with $\Delta_e = \log \pi_\theta(e_r|q, r, e_q) - \log \pi_\theta(e_r|e_q)$.

Proof. Invoking Lemma E.2 and assuming uniform sampling over $\mathcal{E}$, we expand:

$$J_{\text{IC}}(\theta) = \mathbb{E}_q \left[\frac{1}{|\mathcal{E}|} \sum_e \sum_r \pi_\theta(r|e, q) \cdot R(q, r)\right] \tag{11}$$

$$= \mathbb{E}_q \left[\frac{1}{|\mathcal{E}|} \sum_e \sum_r \pi_\theta(r|q) \cdot \frac{\pi_\theta(e_r|q, r, e_q)}{\pi_\theta(e_r|e_q)} \cdot R(q, r)\right] \tag{12}$$

$$= \mathbb{E}_q \left[\sum_r \pi_\theta(r|q) \cdot R(q, r) \cdot \underbrace{\frac{1}{|\mathcal{E}|} \sum_e \frac{\pi_\theta(e_r|q, r, e_q)}{\pi_\theta(e_r|e_q)}}_{w(q,r)}\right]. \tag{13}$$

Writing $\Delta_e = \log \pi_\theta(e_r|q, r, e_q) - \log \pi_\theta(e_r|e_q)$, the weight becomes $w(q, r) = \mathbb{E}_e[\exp(\Delta_e)]$. □

Interpretation. Theorem E.3 establishes an *exact* equivalence: the In-Context RLVR objective is mathematically identical to standard RLVR with rewards reweighted by $w(q, r)$. No approximation is involved. The weight $w(q, r) = \mathbb{E}_e[\exp(\Delta_e)]$ measures how much the candidate trace r improves the model's ability to generate reference solutions on average.

The next theorem characterizes the relationship between the implicit weight $w(q, r)$ and the Evidence Gain $\Delta(q, r) = \mathbb{E}_e[\Delta_e]$.

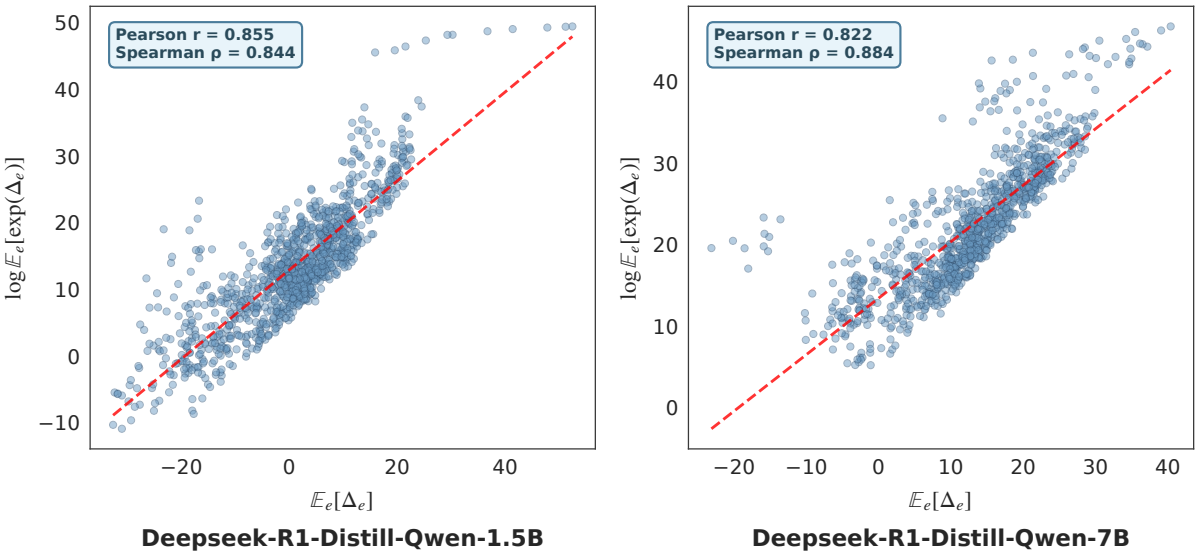

Figure 7: Empirical verification of the log-linear relationship between $\log w(q,r) = \log \mathbb{E}_e[\exp(\Delta_e)]$ and $\Delta(q,r) = \mathbb{E}_e[\Delta_e]$. Strong Pearson correlations ($r = 0.855$ for 1.5B, $r = 0.822$ for 7B) confirm that Evidence Gain serves as a reliable proxy for the implicit weight.

Theorem E.4 (Log-Linear Approximation)**.** *The weight factor $w(q,r)$ from Theorem E.3 and the Evidence Gain $\Delta(q,r) = \mathbb{E}_e[\Delta_e]$ satisfy:*

1. ***(Lower bound)** By Jensen's inequality: $w(q,r) \geq \exp\big(\Delta(q,r)\big)$.*
2. ***(Refined bound)** $\log w(q,r) = \Delta(q,r) + \log\big(1 + \frac{1}{2}\mathrm{Var}_e[\Delta_e]\big) + o(\mathrm{Var}_e[\Delta_e])$.*

In particular, when $\mathrm{Var}_e[\Delta_e]$ is approximately constant across (q,r) pairs, the relationship simplifies to:

$$\log w(q,r) \approx \Delta(q,r) + c, \tag{14}$$

for some constant c that depends on the average variance $\mathbb{E}_{q,r}[\mathrm{Var}_e[\Delta_e]]$.

Proof. Since $\exp(\cdot)$ is convex, Jensen's inequality gives $\mathbb{E}_e[\exp(\Delta_e)] \geq \exp(\mathbb{E}_e[\Delta_e]) = \exp(\Delta(q,r))$, establishing (i).

For (ii), we expand $\exp(\Delta_e)$ around $\Delta := \Delta(q,r)$ via Taylor series:

$$\exp(\Delta_e) = \exp(\Delta) \cdot \exp(\Delta_e - \Delta) \tag{15}$$

$$= \exp(\Delta) \cdot \Big(1 + (\Delta_e - \Delta) + \tfrac{1}{2}(\Delta_e - \Delta)^2 + o\big((\Delta_e - \Delta)^2\big)\Big). \tag{16}$$

Taking expectations and using $\mathbb{E}_e[\Delta_e - \Delta] = 0$:

$$w(q,r) = \exp(\Delta) \cdot \big(1 + \tfrac{1}{2}\mathrm{Var}_e[\Delta_e] + o(\mathrm{Var}_e[\Delta_e])\big). \tag{17}$$

Taking logarithms on both sides and applying $\log(1+x) = x + o(x)$ yields (ii). □

Interpretation. Result (i) shows that $\exp(\Delta(q,r))$ serves as a *lower bound* for the implicit weight $w(q,r)$, but does not quantify how tight this bound is. Result (ii) refines this by showing that the gap is controlled by $\frac{1}{2}\mathrm{Var}_e[\Delta_e]$. Specifically, $\log w(q,r)$ exceeds $\Delta(q,r)$ by approximately $\log(1 + \frac{1}{2}\mathrm{Var}_e[\Delta_e]) > 0$, a strictly positive correction.

Empirical Verification. Although Theorem E.4(ii) indicates that $\log w(q, r)$ and $\Delta(q, r)$ differ by a variance-dependent term, if $\mathrm{Var}_e[\Delta_e]$ remains relatively stable across different (q, r) pairs, the relationship simplifies to an approximate linear correspondence. To verify this, we conduct experiments using rollouts generated by DeepSeek-R1-Distill-Qwen at 1.5B and 7B (Guo et al., 2025a). Specifically, we randomly sample 1,100 (q, r) pairs for the 1.5B model and 1,000 pairs for the 7B model. For each (q, r) pair, we compute $\Delta(q, r) = \mathbb{E}_e[\Delta_e]$ and $\log w(q, r) = \log \mathbb{E}_e[\exp(\Delta_e)]$, then measure their correlation.

Figure 7 presents the results. We observe strong linear relationships. The 1.5B model yields Pearson $r = 0.855$ (Spearman $\rho = 0.844$), while the 7B model achieves $r = 0.822$ (Spearman $\rho = 0.884$). These high correlations confirm that the variance term contributes a near-constant offset across (q, r) pairs, validating the log-linear approximation in Theorem E.4.

Summary. Combining Theorems E.3 and E.4, we conclude that In-Context RLVR exactly reweights rewards by $w(q, r) = \mathbb{E}_e[\exp(\Delta_e)]$, and this weight is approximately log-linear in Evidence Gain: $\log w \approx \Delta + c$. This confirms that the policy model's intrinsic ICL ability provides an effective quality signal. Traces with higher Evidence Gain receive proportionally higher weights in the reweighted objective, without requiring any external evaluator. In this way, In-Context RLVR leverages the model's own capacity to distinguish reasoning quality, enabling implicit reward reweighting through a simple modification to the training procedure.