CircuitLens: Reasoning Circuits as Data Selection Signals for Reinforcement Learning with Verifiable Rewards

Zhuofan Chen[1] **Ziqian Jiao**[1] **Yikai Cui**[1] **Zhixin Cai**[1]
Jun Bai[2] **Wenge Rong**[1]
[1]School of Computer Science and Engineering, Beihang University, Beijing, China
[2]Beijing Institute for General Artificial Intelligence, China
{zhuofanchen, jiaoziqian, cuiyikai, caizhixin, w.rong}@buaa.edu.cn,
baijun@bigai.ai

Abstract

Reinforcement learning with verifiable rewards (RLVR) is sensitive to which problems a model trains on, yet existing selection criteria—difficulty filtering, hand-curation, reward-trajectory scoring—assess data value as an intrinsic property of problems, independent of the model that will learn from them. We introduce Circuit Reasoning Score (CRS), a selection signal derived from 46 reasoning-sensitive attention heads identified via contrastive ablation, computed in a single forward pass on the frozen base model without reward labels or rollouts. CRS runs against the intuitive hypothesis that stronger reasoning-circuit engagement produces better training data: on Qwen2.5-Math-7B, the lowest-engagement decile improves over random selection on three medium-difficulty benchmarks (GSM8K $+2.0$ pp, OlympiadBench $+1.6$ pp, Minerva $+2.9$ pp), while the highest-engagement decile gains less and is indistinguishable from the middle decile. The advantage has boundary conditions: on a domain-curated pool no selection method separates from the others; at 1.5B scale the useful direction differs; and the lowest-reward training condition produces the strongest downstream generalization. Within the Qwen2.5-Math settings tested, RLVR data selection appears regime-dependent rather than reducible to a static ranking of problem quality.

1 Introduction

If a language model's internal reasoning circuits can be identified, the most intuitive use for data selection is clear: train on problems that activate those circuits most strongly. We test that intuition on the Qwen2.5-Math family and find that it does not hold. On Qwen2.5-Math-7B, training on the problems that *least* engage 46 heads identified by contrastive ablation improves over both random selection and the high-engagement alternative on three medium-difficulty benchmarks. The result is not a simple reversal: the useful direction of selection depends on model capacity, source-pool diversity, and the target evaluation regime, in ways that a static ranking of problem quality does not capture.

Existing RLVR data selection methods rely on external criteria: difficulty filtering, hand-curation for "hard but solvable" problems (Li et al., 2025), reward-trajectory scoring, or single-example selection (Wang et al., 2025b). What these approaches share is that they treat a problem's training value as an *intrinsic property*—a function of the problem alone, independent of the model that will learn from it. Yet the same problem can lie at the frontier of one model's competence and far beyond another's, and the same training prompt can produce informative gradients for one policy and zero gradients for another. A selection signal that reflects the *relationship* between a problem and the model's current reasoning state could provide more principled guidance. We propose *Circuit Reasoning Score* (CRS), a per-problem signal derived from the model's own reasoning circuitry. Through contrastive ablation—comparing LM loss increases on reasoning versus trivial probes—we identify 46 reasoning-sensitive heads (bootstrap Jaccard = 0.91). CRS measures how strongly these heads activate during a single forward pass, residualized against token length, requiring no reward labels, rollouts, or reference model.

The core empirical finding runs against the naïve high-activation hypothesis. CRS-bottom selection improves over random on GSM8K, OlympiadBench, and Minerva Math by $+2.0$, $+1.6$, and $+2.9$ pp; CRS-top's gain is smaller and not separable from CRS-middle. This pattern has boundary conditions: it requires pool diversity (no detectable effect on a domain-curated Olympiads pool), the useful direction differs at 1.5B, and it decouples from training reward (the lowest-reward condition generalizes best).

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 51487–51504
October 24-29, 2026 ©2026 Association for Computational Linguistics

Our contributions are:

(1) We propose CRS, a training-free, model-internal data selection signal computed from 46 reasoning-sensitive attention heads via a single forward pass on the frozen base model (§3).

(2) Across 58 GRPO training runs on two model scales, we find evidence against the hypothesis that higher reasoning-circuit engagement produces better RLVR training data (§4).

(3) We show that, within the settings tested, RLVR data selection is regime-dependent—modulated by pool diversity, model capacity, and evaluation difficulty—rather than reducible to a static ranking of problem quality (§4–5).

2 Related Work

RLVR for mathematical reasoning. RLVR builds on verifier-based reasoning and process supervision (Luo et al., 2024; Setlur et al., 2025; Yuan et al., 2025; Zhang et al., 2025). Methods such as GRPO (Shao et al., 2024; DeepSeek-AI, 2025) use within-group reward variance to train chain-of-thought reasoning, with improvements to stability (Yu et al., 2025) and gradient bias (Liu et al., 2025). A key property is that prompts whose completions all receive identical rewards provide no gradient signal, making prompt selection central. Yue et al. (2025) show RLVR narrows the output distribution; Wang et al. (2025b) show a single example can recover most gains. Answer-level rewards can implicitly incentivize correct reasoning (Wen et al., 2025), yet gains arise even under spurious rewards (Shao et al., 2025), suggesting performance depends on the model-data regime.

Data selection for RLVR. Li et al. (2025) select 1,389 of 8,523 Olympiad problems by reward-trajectory alignment; others use difficulty filtering (Lin et al., 2024) or dataset scaling (Yu et al., 2025); for SFT, Ye et al. (2025) show 817 examples suffice. Model-aware selection has been studied via influence-based methods (Xia et al., 2024; Engstrom et al., 2024; Gu et al., 2025) and weak-to-strong filtering (Li et al., 2024; Mekala et al., 2024), but under supervised losses rather than RLVR. Our work differs from curriculum learning (Bengio et al., 2009) and RLVR curricula that adjust sampling by reward variance or environment difficulty (Jiang et al., 2025; Yang et al., 2025; Zeng et al., 2025): CRS is not a difficulty proxy ($R^2 < 2\%$; §5.1), and its useful direction differs across the two scales tested. Two RLVR-specific gaps remain: pool quality vs. selection quality are confounded without same-pool controls (§4.5), and optimal selection may depend on model capacity (§4.6).

Circuit-based data selection. Mechanistic interpretability has identified task-specific heads for induction (Olsson et al., 2022), factual recall (Meng et al., 2022), and reasoning (Elhage et al., 2021), with causal-intervention methods providing a broader toolkit (Hsu et al., 2025; Heimersheim and Nanda, 2024). Recent work examines head-level causal roles (Kissane et al., 2024; Nam et al., 2025) and has uncovered interpretable circuits for chain-of-thought, syllogistic inference, and number representation (Dutta et al., 2024; Kim et al., 2025; Kantamneni and Tegmark, 2025). Wang et al. (2025a) applied circuit signals to SFT data selection via attention-pattern variance. We depart in three ways: targeting RLVR rather than SFT; using contrastive ablation to isolate reasoning-specific heads; and testing both CRS directions rather than assuming higher engagement is better.

Position of this work. Prior RLVR selection asks *which* data to choose; circuit-based selection asks *whether* internal signals can rank examples (answered for SFT). We ask *when* a model-internal ranking helps, fails, or reverses—finding that the answer depends on pool diversity, model capacity, and evaluation regime. This reframes RLVR data selection from a static ranking problem to a regime-dependent matching problem.

3 Method

Our starting point is a simple observation: if certain attention heads specialize in mathematical reasoning, then how strongly a problem engages those heads should reflect the problem's relationship to the model's reasoning capacity. A problem that barely activates reasoning heads may lie outside the model's usual repertoire; one that strongly activates them may be well within it. Both ends could carry useful information for RLVR, but extracting a clean signal requires solving three problems: distinguishing *reasoning-specific* heads from generally important heads (§3.1), aggregating head-level activations into a per-problem score (§3.2), and removing a length confound that otherwise dominates the signal (§3.3). The result is CRS: a real-valued per-problem score from a single forward pass on the frozen base model, with no reward labels or

rollouts required. Figure 1 illustrates the pipeline.

3.1 Reasoning-Head Identification

Qwen2.5-Math-7B (Yang et al., 2024) has 784 attention heads (28×28, d_h=128). To isolate heads specifically important for reasoning (not just general language modeling), we use contrastive ablation. We curate 50 reasoning probes (Olympiad-style) and 50 trivial probes (single-step arithmetic), held out from all training and evaluation data. For each head, we zero its output and measure LM loss increases on each probe set (Δ_r, Δ_t):

$$s_{l,h} = \Delta_r(l,h) - 0.5\,\Delta_t(l,h). \quad (1)$$

The contrastive term is necessary because Δ_r and Δ_t correlate at $r \approx 0.75$; Eq. 1 extracts reasoning-specific importance beyond general-purpose function, unlike Wang et al. (2025a) who ablate against reasoning probes only.

We repeat the scan three times with resampled probes. The intersection yields 47 stable heads (pairwise Jaccard 0.89–0.92, mean 0.91; cross-round score correlation $r = 0.9998$). Excluding one global-bottleneck outlier (L0H3, score $4.8\times$ the runner-up), the final set $\mathcal{H}_r$ contains **46 heads** spanning 21 of 28 layers.

We use "reasoning-sensitive heads" to denote heads identified by contrastive ablation on held-out probes; we do not claim a verified causal pathway.

3.2 CRS Computation

We opt for activation magnitude—the ℓ_2-norm of each head's output—rather than attention-weight statistics, since activation norms directly measure how much information a head contributes to the residual stream (Kissane et al., 2024; Nam et al., 2025). For each problem x with n tokens, a single forward pass records each reasoning head's output activation at every position. The raw score averages per-head maxima:

$$\mathrm{CRS}_{\mathrm{raw}}(x) = \frac{1}{|\mathcal{H}_r|} \sum_{h \in \mathcal{H}_r} \max_{t \in [1,n]} \left\| \mathbf{a}_h^{(t)}(x) \right\|_2. \quad (2)$$

Taking the maximum captures peak reasoning engagement rather than diluting signal across boilerplate tokens.

3.3 Length Residualization

Raw CRS correlates with input length at $r = 0.568$. Since length is a $2.8\times$ stronger difficulty predictor than CRS ($R^2 = 4.36\%$ vs. 1.57%; §5.1), we residualize via OLS on $\log n$:

$$\mathrm{CRS}(x) = \mathrm{CRS}_{\mathrm{raw}}(x) - \hat{\beta}_0 - \hat{\beta}_1 \log n_x, \quad (3)$$

yielding $r = 0.013$ with length. The Jaccard overlap between the top-10% by residualized CRS and the longest 10% drops from 0.286 to 0.057, near chance. All subsequent references to CRS denote this residualized version. Appendix E reports sensitivity to the residualization form and to the contrastive weight in Eq. 1.

3.4 Selection and Controls

We rank DAPO-Math-17K (Yu et al., 2025) (17,186 problems) by CRS and extract **CRS-top** (highest 10%), **CRS-middle** (percentiles 45–55), and **CRS-bottom** (lowest 10%; 1,718 each), plus per-seed **Random** baselines.

Random-heads control. We repeat scoring with 46 randomly sampled non-reasoning heads and train on the resulting top-10% (**RandHeads**), testing whether the specific 46-head identity matters.

Cross-pool control. We re-score the 8,523-problem Olympiads pool—from which LIMR (Li et al., 2025) was curated—and construct size-matched subsets (1,389 each). Jaccard with LIMR's selection is 0.15–0.17, confirming independence.

Cross-scale control. We replicate on Qwen2.5-Math-1.5B, adding `xmod`: data selected by 7B's CRS, used to train 1.5B. `xmod` is a scale-transfer variant of the weak-to-strong direction (Li et al., 2024; Mekala et al., 2024), not a full weak-to-strong filtering baseline.

4 Experiments

4.1 Setup

All experiments use GRPO (Shao et al., 2024) via TRL v1.3.0 with the hyperparameters in Table 1, following defaults from DeepSeek-AI (2025) and Yu et al. (2025) on the Qwen2.5-Math family (Yang et al., 2024). Binary reward: 1 if the `\boxed{}` answer matches ground truth (SymPy), 0 otherwise. Under TRL's `RepeatSampler`, 1,000 steps cover ∼58% of each 1,718-prompt subset.

Conditions. **Tier 1** (5 seeds): CRS top / mid / bottom-10%, Random 10% of DAPO-17K. **Tier 2** (3 seeds): LIMR (Li et al., 2025) (1,389 from the 8,523-problem Olympiads pool), Full data (17K),

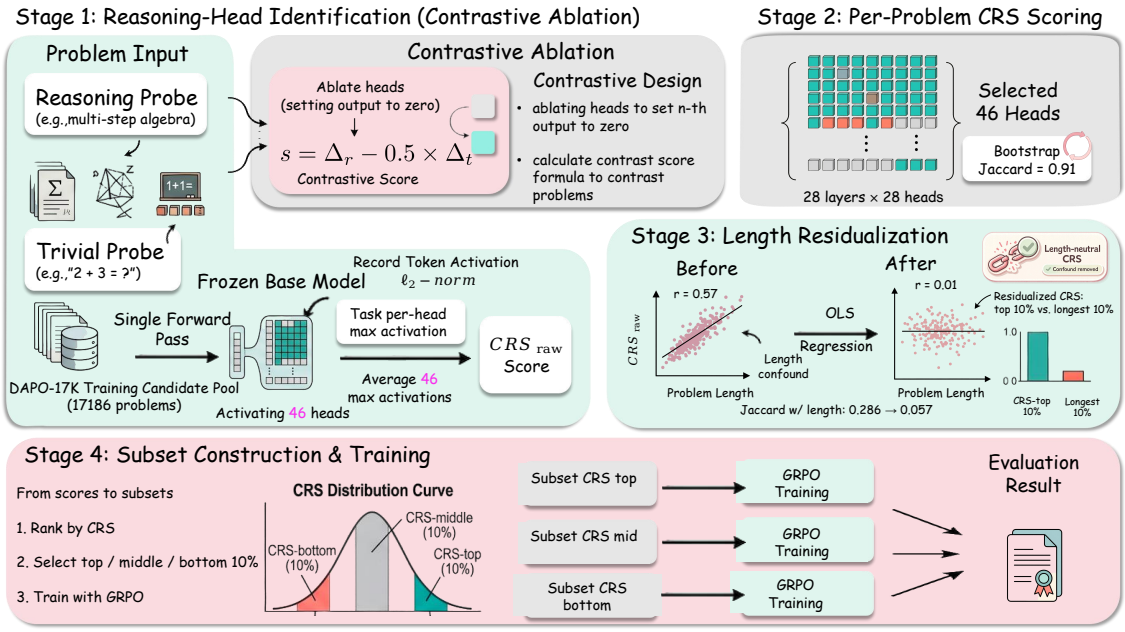


Figure 1: CRS pipeline. **Stage 1**: Contrastive ablation identifies 46 reasoning-sensitive heads (Jaccard = 0.91). **Stage 2**: Single forward pass records activation magnitudes; per-head max averaged to CRS_{raw}. **Stage 3**: OLS residualization removes length confound ($r : 0.57 \rightarrow 0.01$). **Stage 4**: Top/middle/bottom 10% deciles form GRPO training subsets.

Parameter	Value
Learning rate	1×10^{-6}, cosine, warmup 0.1
Max gradient norm	1.0
Temperature / Generations	0.7 / 16 per prompt
Max completion length	4,096 tokens
Batch / Grad. accum.	4 / 4 (eff. 16 prompts/step)
Training steps	1,000
Hardware	8×A100-SXM4-80GB

Table 1: GRPO hyperparameters (shared across all conditions and scales).

RandHeads top-10%, and a difficulty-filtering baseline (Diff-hard, 3 seeds; Diff-easy, 2 seeds; §4.4). **Tier 4** (3 seeds): Oly-CRS-top / bottom / Random (1,389 each from the Olympiads pool). **1.5B** (3 seeds): CRS top / mid / bottom-10%, Random 10%, `xmod`. Total: 58 runs.

Evaluation. Five primary benchmarks spanning a difficulty gradient (Table 2); OlympiadBench (He et al., 2024) anchors the competition tier. Two additional benchmarks (AMC 2023, n=40; AIME 2025, n=30) are in Appendix A. Benchmarks with $n \leq 60$ use avg@8 (τ=0.6); others use greedy pass@1. Pairwise comparisons are reported at the seed unit (§4.3).

Benchmark	Difficulty	N	Metric
GSM8K	Grade-school	1,319	pass@1
Minerva Math	Undergrad STEM	272	pass@1
MATH-500	Competition (mixed)	500	pass@1
OlympiadBench	Competition (mixed)	562	pass@1
AIME 2024	Competition (hard)	60	avg@8

Table 2: Primary benchmarks (no overlap with training pool).

4.2 CRS Is Not a Monotonic Quality Score

If CRS captured a monotonic notion of data quality, higher-CRS data should produce uniformly better models. Table 3 shows this is not the case.

CRS-bottom achieves the highest Tier 1 accuracy on GSM8K, OlympiadBench, and Minerva, while CRS-top's margin over Random is small on every benchmark. Figure 2 summarizes this: CRS-bottom's gains (green) concentrate on medium-difficulty benchmarks; its deficit (red) appears on AIME. No condition dominates all benchmarks.

4.3 CRS-Bottom Improves Medium-Difficulty Reasoning

We compare conditions at the seed unit, since what matters for a selection rule is whether its advantage reproduces across training runs. CRS-bottom's advantage over Random is +2.0 pp on GSM8K,

	Condition	GSM8K	Minerva	MATH-500	Olymp.	AIME '24
	Base	81.20	16.18	66.80	37.76	11.46
T1	CRS-top	84.26 ± 0.64	20.81 ± 1.90	69.20 ± 1.51	37.45 ± 0.47	16.63 ± 1.79
T1	CRS-mid	83.52 ± 0.57	20.88 ± 2.13	68.52 ± 1.51	37.87 ± 0.88	17.08 ± 1.97
T1	CRS-bot	$\mathbf{85.81 \pm 2.48}$	$\mathbf{22.50 \pm 3.05}$	68.56 ± 3.88	$\mathbf{38.36 \pm 1.91}$	14.88 ± 1.97
T1	Random	83.82 ± 0.58	19.63 ± 1.09	69.12 ± 1.29	36.75 ± 0.86	16.50 ± 1.26
T2	LIMR	85.39 ± 1.03	21.20 ± 2.61	70.93 ± 0.64	38.00 ± 0.66	15.41 ± 0.91
T2	Full (17K)	83.73 ± 0.61	17.53 ± 2.76	67.93 ± 0.70	36.30 ± 0.86	17.50 ± 1.46
T2	RandHeads	85.65 ± 2.72	20.47 ± 2.97	69.67 ± 2.72	37.18 ± 2.33	15.76 ± 2.65

Table 3: 7B main results: 7 conditions $\times$ 5 benchmarks (mean $\pm$ std). Tier 1: n=5; Tier 2: n=3. **Bold**: highest Tier 1 mean. LIMR uses a different pool; see §4.5.

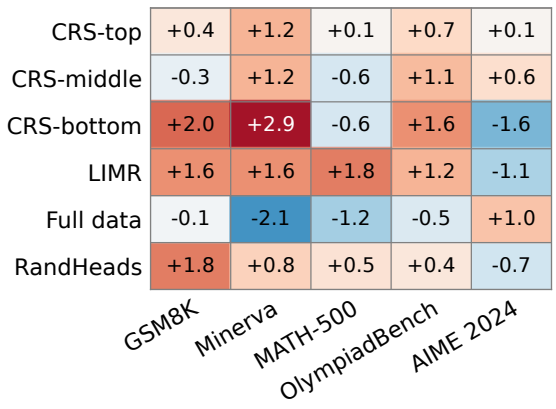


Figure 2: Gain over Random (pp) per condition per benchmark, as unpaired differences of condition means. CRS-bottom's gains cluster on medium-difficulty benchmarks; LIMR shows a deficit on AIME.

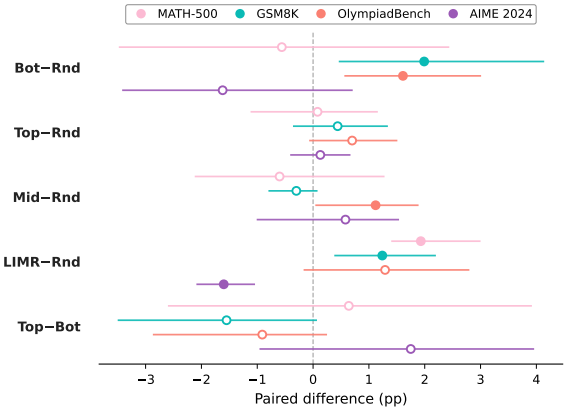


Figure 3: Seed-level paired differences for key comparisons across five benchmarks, with resampling envelopes. CRS-bottom exceeds Random on GSM8K, OlympiadBench, and Minerva; LIMR trails Random on AIME.

	GSM8K	Minerva	MATH	Olymp.	AIME	Pos.
CRS-bot	+2.0	+2.9	−0.6	+1.6	−1.6	13/15
CRS-top	+0.4	+1.2	+0.1	+0.7	+0.1	10/15
CRS-mid	−0.3	+1.3	−0.6	+1.1	+0.6	8/15

Table 4: Seed-level paired differences against Random (pp, n=5 seeds). "Pos.": seed-by-benchmark cells favouring the condition, out of 15, over the three target benchmarks.

+2.9 pp on Minerva, and +1.6 pp on OlympiadBench (Table 4), positive on 13 of 15 seed-by-benchmark cells and on all five seed composites, for a composite of +2.15 pp. A hierarchical binomial model of the same outcomes, treating the condition-by-seed interaction as a free parameter, places 91.3% of the posterior for the condition effect above zero (Appendix C.5). An item-level McNemar test on GSM8K ($p = 3.3 \times 10^{-4}$; Appendix C.1) confirms consistency across items within a run.

The effect has clear boundaries: on MATH-500 no Tier 1 condition separates from Random, consistent with a ceiling effect at 66.8% base accuracy; on AIME, CRS-bottom is −1.6 pp against Random. CRS-top's point estimate also exceeds Random on the three target benchmarks, so the relationship is not monotone in CRS, though the two upper conditions are not separable at this seed count (Appendix C.6). The CRS-bottom advantage is specific to medium-difficulty evaluation—a boundary condition we revisit in §4.6. Figure 3 provides a visual summary of all key comparisons.

These results raise a natural question: is the CRS-bottom effect genuinely attributable to circuit-based selection, or could simpler explanations account for it?

4.4 Ruling Out Alternative Explanations

We consider four alternative hypotheses, each addressed by a dedicated baseline.

"Reward-trajectory selection is all you need." LIMR (Li et al., 2025) selects by reward-trajectory

alignment and serves as our reward-trajectory baseline. It leads on MATH-500 (+1.9 pp over Random) but trails Random on AIME (−1.6 pp), where CRS-top exceeds it by +1.7 pp. LIMR's pool differs from ours (the 8,523-problem Olympiads pool vs. DAPO-17K), confounding selection quality with pool quality—a confound we isolate in §4.5. Selection criteria optimized for intermediate difficulty do not transfer to the hardest evaluation tier.

"Any head subset works." RandHeads achieves similar means to CRS-top (GSM8K: 85.65 vs. 84.26) but with 2–4× higher cross-seed variance (GSM8K std: 2.72 vs. 0.64; MATH: 2.72 vs. 1.51). Arbitrary head subsets can produce effective selections by chance in some seeds, but the 46 identified heads provide a more *consistent* signal. A post-training ablation against six layer-matched random head sets (Appendix C.4) approaches the same question causally.

"More data is better." Full-data training (17K problems, 5.8% coverage at 1,000 steps) trails all 10% subsets on MATH-500 (67.93 ± 0.70). A larger pool does not help when training budget is fixed; focused selection outperforms diluted sampling.

"Higher CRS is better." CRS-top's margin over Random is small on every benchmark and not separable from CRS-middle (§4.3; full difference table in Appendix A). The larger and more seed-stable effect at 7B lies in the bottom decile, not the top—a finding that runs against the naïve expectation that problems engaging the most reasoning circuitry should be the best training data.

"Difficulty filtering is all you need." Using the rollout-estimated success rates of §5.1 (K=4; valid estimates for 16,478 pool problems), we built two disjoint 1,647-problem subsets at the two ends of the SR distribution—**Diff-hard** and **Diff-easy**, selected by rank and labelled by measured mean SR—and trained them under the identical protocol (Appendix C.7). CRS-bottom remains the strongest condition on GSM8K, Minerva, and OlympiadBench (Table 5), the three benchmarks on which our claim rests. The two selectors are nearly orthogonal (Jaccard 0.030–0.069 against a chance level of ≈0.051), independently reproducing the result of §5.1 that CRS is not a difficulty proxy. The cost asymmetry is large: difficulty filtering required 65,912 generations, against a single forward pass per problem for CRS.

Taken together, these baselines indicate that CRS-bottom's advantage is not reducible to reward-trajectory selection, arbitrary head signals, data volume, CRS magnitude, or difficulty filtering. But a deeper question remains: is the effect a property of CRS itself, or of the training pool?

4.5 Selection Effects Require Pool Diversity

On the homogeneous 8,523-problem Olympiads pool, we construct three size-matched subsets (1,389 each). Table 6 shows no detectable difference among the three, despite selecting nearly independent problem sets (Jaccard with LIMR: 0.15–0.17).

On diverse DAPO-17K, CRS stratification produces up to 2.9 pp differences; on the narrow Olympiads pool, it produces none. The contrast is informative: on the diverse pool, CRS stratification separates grade-school arithmetic from Olympiad reasoning, and training outcomes diverge; on the homogeneous pool, CRS stratification produces subsets that differ in identity but converge in outcome. This is consistent with adaptive-environment work showing that narrow distributions lose learning signal (Zeng et al., 2025; Stojanovski et al., 2025). Pool diversity is a prerequisite for selection effects—if the pool is already domain-curated, effort is better spent expanding it than optimizing within it.

This establishes that CRS effects are pool-dependent. But are they also model-dependent?

4.6 The Useful Direction Differs at 1.5B

At 1.5B (Table 7), CRS-top leads on MATH-500 (65.4), GSM8K (77.5), and OlympiadBench (31.4)—the same benchmarks where CRS-bottom led at 7B. CRS-bottom trails CRS-top by 3.0 pp on MATH-500, against −0.6 pp at 7B. Figure 4 visualizes the difference.

The reversal is consistent with GRPO's reward-variance mechanism (Jiang et al., 2025; Yang et al., 2025): GRPO produces nonzero gradients only when the G=16 completions exhibit reward variance—some succeed and others fail. At 7B, both CRS-top and CRS-bottom problems lie within the regime of useful reward variance, but CRS-bottom provides more novel reasoning patterns when successes occur, producing transferable skills for easier benchmarks. At 1.5B, CRS-bottom problems fall below the threshold where the weaker

Condition	n	GSM8K	Minerva	MATH-500	Olymp.	AIME '24
Diff-hard (SR=.25)	3	84.53±0.69	21.08±1.50	68.60±1.10	37.47±0.90	17.57±2.10
Diff-easy (SR=.92)	2	83.36±0.48	18.75±0.00	68.10±0.14	36.98±0.37	15.94±1.00
CRS-bottom	5	85.81±2.48	22.50±3.05	68.56±3.88	38.36±1.91	14.88±1.97
Random	5	83.82±0.58	19.63±1.09	69.12±1.29	36.75±0.86	16.50±1.26

Table 5: Difficulty-filtering baseline (mean ± std). Conditions labelled by measured mean success rate. CRS-bottom and Random rows are the Tier 1 references.

	MATH	GSM8K	Olymp.
Oly-CRS-top	70.5 ± 2.5	86.0 ± 2.7	37.9 ± 0.8
Oly-CRS-bot	71.4 ± 3.1	86.6 ± 2.9	38.1 ± 3.3
Oly-Random	69.7 ± 1.4	86.8 ± 1.0	38.1 ± 1.6
LIMR (ref.)	70.9 ± 0.6	85.4 ± 1.0	38.0 ± 0.7

Table 6: Tier 4: Olympiads pool (n=3 seeds).

	MATH	GSM8K	Olymp.	AIME
Base	63.20	76.19	30.94	7.50
CRS-top	**65.4±3.0**	**77.5±4.5**	**31.4±3.3**	9.9±1.3
CRS-bot	62.4±0.6	76.5±3.1	30.6±0.5	10.2±1.3
Random	63.6±0.8	76.4±0.9	30.2±0.5	9.5±0.5
xmod	63.1±1.2	77.3±0.9	30.3±0.2	9.7±1.2

Table 7: 1.5B results (n=3 seeds). **Bold**: best per benchmark. xmod: 7B's CRS top-10% used for 1.5B.

model can generate correct completions, yielding predominantly zero-advantage steps. CRS-top problems remain within the 1.5B model's productive zone, explaining why high-CRS data becomes more valuable at smaller scale. This interpretation predicts that the optimal CRS direction should shift continuously with model capacity—a prediction we leave to future verification with intermediate-scale models.

Cross-model portability. The xmod condition matches native 1.5B CRS-top (GSM8K: 77.3 vs. 77.5, std 0.87 vs. 4.48) and trends positive over Random (+1.0 pp), echoing weak-to-strong filtering results (Li et al., 2024; Mekala et al., 2024) and suggesting CRS captures a partially transferable pool-level signal. CRS subsets computed once on a strong base model can be reused across moderate scale shifts.

4.7 Training Reward Does Not Predict Generalization

If training reward were a sufficient proxy for useful learning, the highest-reward condition should generalize best. The opposite emerges: CRS-bottom converges to the lowest reward (0.265) yet achieves

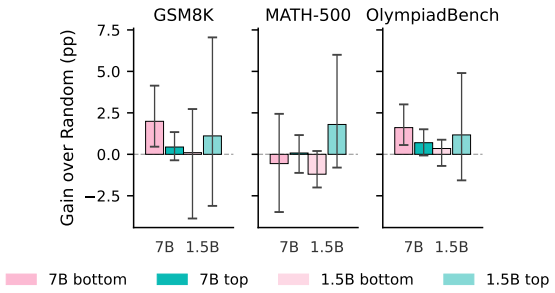


Figure 4: Useful direction at the two scales tested. At 7B, CRS-bottom (pink) is positive; at 1.5B, CRS-top (teal) leads. Error bars: resampling envelopes.

Condition	Reward	Zero-std %
LIMR	0.775	∼7%
CRS-middle	0.485	∼24%
Random	0.470	∼40%
CRS-top	0.362	∼44%
RandHeads	0.296	∼33%
CRS-bottom	**0.265**	∼28%
Full data	0.200	∼53%

Table 8: Training dynamics at step 1,000 (7B, seed-averaged).

the highest GSM8K (85.81) and Minerva (22.50); LIMR reaches the highest reward (0.775) but underperforms Random on AIME (Table 8). Across all runs the correlation between training reward and downstream accuracy is $r = -0.18$; across condition means it is +0.03 (Figure 5).

This decoupling resolves an apparent paradox: how can CRS-bottom—a condition where the model mostly *fails* during training—produce the strongest downstream performance? Training reward measures how well the model solves its training problems; it does not measure whether the learned skills generalize. CRS-bottom forces exploration of reasoning patterns outside the model's existing repertoire, and the resulting skills transfer effectively to evaluation benchmarks where those patterns are useful. Practitioners should not optimize data selection for training reward: in RLVR, the conditions that look worst during training may



Figure 5: Training reward vs. generalization. Across all runs $r = -0.18$; across the seven condition means $r = +0.03$. Optimizing for training reward does not predict downstream capability.

Signal	ρ	R^2 (%)
CRS (residualized)	$+0.125$	1.57
log token length	-0.209	4.36

Table 9: Correlation with base-model success rate (n=16,478).

generalize best.

5 Discussion

5.1 CRS Is Not a Difficulty Proxy

A natural concern is that CRS simply recapitulates problem difficulty: if low-CRS problems are easy and high-CRS problems are hard, the Tier 1 effects could reduce to "train on easy problems to improve easy benchmarks." We test this directly by computing rollout-estimated success rates (SR) for 16,478 DAPO-17K problems (K=4 rollouts at τ=0.6) and measuring their correlation with residualized CRS.

Residualized CRS explains only 1.57% of the variance in base-model success rate ($\rho = 0.125$, 95% CI $[+0.112, +0.140]$, n=16,478); token length explains 4.36%—a 2.8× stronger predictor (Table 9). The difficulty-filtering baseline of §4.4 reproduces this independently: its SR-selected subsets overlap the CRS subsets at Jaccard 0.030–0.069, at chance. Figure 6 makes the weak relationship visually apparent. The scatter (panel a) shows 16,478 problems as a diffuse cloud with a barely perceptible positive slope; SR takes all five discrete values at every CRS level. The per-decile violin plot (panel b) confirms this: the SR distribution within each CRS decile spans the full $[0, 1]$ range, with median SR shifting only from ∼0.15 (decile 1) to ∼0.30 (decile 10). As a further check, Jaccard overlap between CRS-top and SR-top (the easiest 10%) is 0.078, barely above the chance level of 0.053; CRS-bottom and SR-bottom overlap at 0.055—essentially random. CRS-bottom's advantage on medium-difficulty benchmarks (§4.3) therefore cannot be attributed to "CRS selects easy problems": CRS-bottom and the easiest-by-SR problems are almost entirely disjoint sets.

5.2 Three Axes of Selection Sensitivity

The experiments in §4 reveal that RLVR data selection effects are modulated by three factors, each of which can independently determine whether selection helps, hurts, or is irrelevant.

Pool diversity is the prerequisite. On diverse DAPO-17K, CRS-bottom outperforms Random by up to 2.9%; on the domain-narrow Olympiads pool, the same procedure produces no detectable effect (§4.5). If the training pool has already been curated to a narrow domain, effort is better spent expanding diversity than optimizing within-pool selection.

Evaluation regime determines the useful direction. CRS-bottom's advantage concentrates on medium-difficulty benchmarks (GSM8K, OlympiadBench, Minerva); on AIME, it provides no benefit. No single direction dominates all evaluation regimes—the practical guidance is to match selection direction to the target.

Model capacity shifts the productive zone. At 7B, the model benefits from low-CRS problems—unfamiliar territory that forces exploration. At 1.5B, the same problems fall below the productive reward-variance threshold (§4.6), and high-CRS problems become more valuable.

Unified framing. These three axes reframe the question from "which selection method is best?" to "which regime am I in?" A diverse pool with a capable model and medium-difficulty targets is the regime where CRS-bottom is most effective. Outside this regime, the signal either vanishes or reverses, and practitioners should adjust accordingly.

6 Conclusion

Within the Qwen2.5-Math settings tested, the training value of a problem for RLVR does not appear to be an intrinsic property of the problem. CRS serves less as a universal selector and more as a probe that exposes how selection effects depend on context—appearing on a diverse pool but not detectable on a

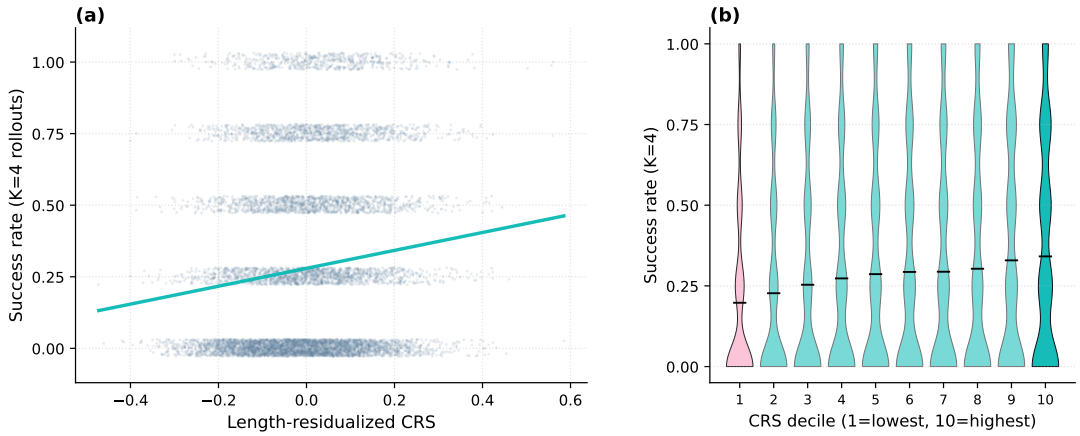


Figure 6: CRS is not a difficulty proxy. **(a)** Scatter of length-residualized CRS vs. rollout-estimated success rate (K=4, n=16,478). The regression line (Spearman ρ=0.125, R^2=1.57%) is barely visible against the spread. **(b)** Violin plots of SR per CRS decile. Distributions overlap heavily; median success rate shifts only from ∼0.15 to ∼0.30 across deciles. CRS and difficulty rank problems in nearly independent orders.

curated one, favouring low-engagement data at 7B but high-engagement data at 1.5B, and producing the strongest generalization from the lowest-reward condition. These interactions are invisible to criteria that treat data quality as problem-intrinsic, which is why selection optimized for intermediate difficulty can underperform random selection on the hardest tier. For practitioners working in a comparable setting: on a diverse pool with a 7B-scale model, CRS-bottom improved medium-difficulty reasoning in our runs; on curated pools or smaller models, the signal was not detectable or ran the other way.

Limitations

Scope. All experiments use the Qwen2.5-Math family at 7B and 1.5B, on mathematical data only. Generalization to other architectures and verifiable-reward domains is untested.

Statistical power and prompt coverage. The 7B analysis uses n=5 seeds and Tier 4 / 1.5B use n=3; we report effect sizes rather than significance labels. At 1,000 steps each run sees 58.2% of its 1,718-prompt subset, so labels denote a random sample of each decile, contributing to cross-seed variance (Appendix C.5).

Causal interpretation. CRS is correlational: activation magnitudes at heads identified via ablation. Zero-ablating these heads removes the CRS-bottom advantage but is equally destructive to the untrained base model, indicating a pathway already present in the base model rather than one built by GRPO. We present CRS as a selection-level predictive signal rather than a verified reasoning circuit.

Ethics Statement

This work uses publicly available mathematical benchmarks and open-weight models. No human subjects, private data, or dual-use concerns are involved.

Acknowledgments

During the preparation of this work, the authors used AI-assisted tools for language polishing, grammar checking, coding assistance, and brainstorming. All research ideas, experimental designs, data analyses, and conclusions were developed and verified by the authors. The authors take full responsibility for the content of the paper.

This work was supported in part by the National Natural Science Foundation of China under Grant 62477001.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, ACM International Conference Proceeding Series, pages 41–48. ACM.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *Trans. Mach. Learn. Res.*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. Transformer Circuits Thread.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. 2024. Dsdm: Model-aware dataset selection with datamodels. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 12491–12526. PMLR / OpenReview.net.

Yuxian Gu, Li Dong, Hongning Wang, Yaru Hao, Qingxiu Dong, Furu Wei, and Minlie Huang. 2025. Data selection via optimal control for language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3828–3850. Association for Computational Linguistics.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *CoRR*, abs/2404.15255.

Aliyah R. Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Y. Odisho, Peter R. Carroll, and Bin Yu. 2025. Efficient automated circuit discovery in transformers using contextual decomposition. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Guochao Jiang, Wenfeng Feng, Guofeng Quan, Chuzhan Hao, Yuewei Zhang, Guohua Liu, and Hao Wang. 2025. VCRL: variance-based curriculum reinforcement learning for large language models. *CoRR*, abs/2509.19803.

Subhash Kantamneni and Max Tegmark. 2025. Language models use trigonometry to do addition. *CoRR*, abs/2502.00873.

Geonhee Kim, Marco Valentino, and André Freitas. 2025. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 10074–10095. Association for Computational Linguistics.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. 2024. Interpreting attention layer outputs with sparse autoencoders. *CoRR*, abs/2406.17759.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14255–14273. Association for Computational Linguistics.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. LIMR: less is more for RL scaling. *CoRR*, abs/2502.11886.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. 2024. Rho-1: Not all tokens are what you need. *CoRR*, abs/2404.07965.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *CoRR*, abs/2503.20783.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. 2024. Improve mathematical reasoning in language models by automated process supervision. *CoRR*, abs/2406.06592.

Dheeraj Mekala, Alex Nguyen, and Jingbo Shang. 2024. Smaller language models are capable of selecting instruction-tuning training data for larger language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 10456–10470. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Andrew Nam, Henry Conklin, Yukang Yang, Thomas L. Griffiths, Jonathan D. Cohen, and Sarah-Jane Leslie. 2025. Causal head gating: A framework for interpreting roles of attention heads in transformers. *CoRR*, abs/2505.13737.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *CoRR*, abs/2209.11895.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2025. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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. 2025. Spurious rewards: Rethinking training signals in RLVR. *CoRR*, abs/2506.10947.

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. *CoRR*, abs/2402.03300.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. 2025. REASONING GYM: reasoning environments for reinforcement learning with verifiable rewards. *CoRR*, abs/2505.24760.

Shaobo Wang, Yongliang Miao, Yuancheng Liu, Qianli Ma, Ning Liao, and Linfeng Zhang. 2025a. Circuitseer: Mining high-quality data by probing mathematical reasoning circuits in llms. *CoRR*, abs/2510.18470.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. 2025b. Reinforcement learning for reasoning in large language models with one training example. In *Advances in Neural Information Processing Systems*, volume 38.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2025. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *CoRR*, abs/2506.14245.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 54104–54132. PMLR / OpenReview.net.

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. *CoRR*, abs/2409.12122.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Yongxin Wang, Dongchun Xie, Hanhui Li, Yiwei Wang, Xiaodan Liang, and Jing Tang. 2025. Depth-breadth synergy in RLVR: unlocking LLM reasoning gains with adaptive exploration. *CoRR*, abs/2508.13755.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. LIMO: less is more for reasoning. *CoRR*, abs/2502.03387.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2025. Free process rewards without process labels. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

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

Zhiyuan Zeng, Hamish Ivison, Yiping Wang, Lifan Yuan, Shuyue Stella Li, Zhuorui Ye, Siting Li, Jacqueline He, Runlong Zhou, Tong Chen, Chenyang Zhao, Yulia Tsvetkov, Simon Shaolei Du, Natasha Jaques, Hao Peng, Pang Wei Koh, and Hannaneh Hajishirzi. 2025. RLVE: scaling up reinforcement learning for language models with adaptive verifiable environments. *CoRR*, abs/2511.07317.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July*

27 - August 1, 2025, Findings of ACL, pages 10495–10516. Association for Computational Linguistics.

A Full Results Tables

A.1 7B Main Grid: All 7 Benchmarks

Table 10 extends Table 3 with two additional benchmarks: AMC 2023 (n=40, avg@8) and AIME 2025 (n=30, avg@8). These benchmarks have high per-problem variance due to small n and are excluded from the primary analysis in the main text.

A.2 Full Seed-Level Difference Table (Tier 1)

Table 11 reports all six Tier 1 pairwise comparisons across all 7 benchmarks. These are the exhaustive differences underlying the forest plot in Figure 3 and the summary in Table 4.

A.3 Tier 2 and 1.5B Seed-Level Differences

Table 12 reports Tier 2 pairwise comparisons (n=3 seeds). The LIMR–Random comparison on AIME 2024 (-1.6 pp) is the basis for the LIMR deficit reported in §4.4. The Top–LIMR row reports only the AIME comparison, as this is the only benchmark where the two differ appreciably.

Table 13 reports 1.5B mini-Tier 1 pairwise comparisons (n=3 seeds). The Bot–Top comparison on MATH-500 (-3.0 pp) is the basis for the cross-scale difference in §4.6. The xmod–Random comparison on GSM8K ($+1.0$ pp) is the cross-model portability trend discussed in the same section.

B Per-Seed Results

Figure 7 visualizes the per-seed spread for Tier 1 conditions across five primary benchmarks, providing full transparency on cross-seed variance. Tables 14–16 report the underlying per-seed accuracies. Seeds are ordered: s42, s3407, s31415, s2024, s7 (Tier 1) or s42, s3407, s31415 (Tier 2 / 1.5B).

C Supplementary Analyses

This section collects the analyses referenced in the main text: item-level McNemar tests, subset overlap statistics, the CRS distribution, the post-training head ablation, the hierarchical model of the condition effect, the two ends of the CRS ranking, and the construction of the difficulty-filtering subsets.

C.1 McNemar Item-Level Tests

For greedy-decoded benchmarks, we run McNemar's exact test on item-level binary outcomes for CRS-bottom vs. Random (using the median-accuracy seed). This is an item-level statement about a single pair of runs and is complementary to the seed-level analysis of §4.3. The GSM8K result ($p = 3.3 \times 10^{-4}$) reflects a net 38 additional problems solved by CRS-bottom that Random does not. MATH-500 and OlympiadBench do not reach item-level significance with a single seed pair, consistent with the smaller per-condition accuracy differences on these benchmarks.

C.2 Subset Overlap Analysis

Table 18 quantifies the independence of different selection methods. All Jaccard values are at or near chance, confirming that CRS, SR (success-rate-based), and LIMR rank problems in nearly independent orders.

C.3 CRS Distribution

Figure 8 shows the distribution of length-residualized CRS across all 17,186 DAPO-17K problems. The distribution is approximately Gaussian with mild right skew. Colored bands indicate the three selection deciles: bottom 10% (pink), middle 10% (gray), and top 10% (teal). The decile boundaries are well-separated, confirming that the three subsets represent genuinely different regions of the CRS ranking.

C.4 Post-Training Head Ablation

We zero-ablate the 46 identified heads in the trained models (CRS-bottom and Random, median seed, following the McNemar convention of §C.1), against six layer-matched random 46-head control sets and in the untrained base model, with the readout criterion fixed in advance. DR(S) denotes the reduction in CRS-bottom's advantage over Random when head set S is zero-ablated in both trained models.

Ablating the identified heads removes the CRS-bottom advantage on every benchmark, leaving a residual advantage near zero. DR uses the median seed, so it can slightly exceed the five-seed mean advantage (e.g. $+2.04$ vs. $+1.61$ on OlympiadBench). On Minerva the identified set yields the largest reduction of the seven sets tested; on GSM8K and OlympiadBench it ranks third, and the seven sets are not well separated there.

	Condition	GSM8K	Minerva	MATH-500	Olymp.	AIME '24	AMC '23	AIME '25
	Base	81.20	16.18	66.80	37.76	11.46	50.31	7.92
Tier 1	CRS-top	84.26±0.64	20.81±1.90	69.20±1.51	37.45±0.47	16.63±1.79	60.19±2.01	9.92±1.65
Tier 1	CRS-mid	83.52±0.57	20.88±2.13	68.52±1.51	37.87±0.88	17.08±1.97	61.56±3.17	9.75±2.14
Tier 1	CRS-bot	85.81±2.48	22.50±3.05	68.56±3.88	38.36±1.91	14.88±1.97	59.50±1.37	9.58±2.69
Tier 1	Random	83.82±0.58	19.63±1.09	69.12±1.29	36.75±0.86	16.50±1.26	58.81±1.42	9.08±1.00
Tier 2	LIMR	85.39±1.03	21.20±2.61	70.93±0.64	38.00±0.66	15.41±0.91	58.54±3.21	10.56±1.46
Tier 2	Full (17K)	83.73±0.61	17.53±2.76	67.93±0.70	36.30±0.86	17.50±1.46	59.27±1.09	10.00±0.42
Tier 2	RandHeads	85.65±2.72	20.47±2.97	69.67±2.72	37.18±2.33	15.76±2.65	59.48±0.18	9.58±2.09

Table 10: Full 7B results: 7 conditions × 7 benchmarks (mean ± std across seeds). Tier 1: n=5 seeds; Tier 2: n=3. AMC 2023 and AIME 2025 (rightmost two columns) are excluded from primary analysis due to small n.

Comparison	GSM8K	Minerva	MATH	Olymp.	AIME '24	AMC '23	AIME '25
Bot−Rnd	+2.0	+2.9	−0.6	+1.6	−1.6	+0.7	+0.5
Top−Rnd	+0.4	+1.2	+0.1	+0.7	+0.1	+1.4	+0.8
Mid−Rnd	−0.3	+1.3	−0.6	+1.1	+0.6	+2.8	+0.7
Bot−Top	+1.6	+1.7	−0.6	+0.9	−1.8	−0.7	−0.3
Mid−Top	−0.7	+0.1	−0.7	+0.4	+0.5	+1.4	−0.2
Bot−Mid	+2.3	+1.6	+0.0	+0.5	−2.2	−2.1	−0.2

Table 11: Tier 1 seed-level paired differences (n=5 seeds), all 7 benchmarks. Values are mean paired difference in pp.

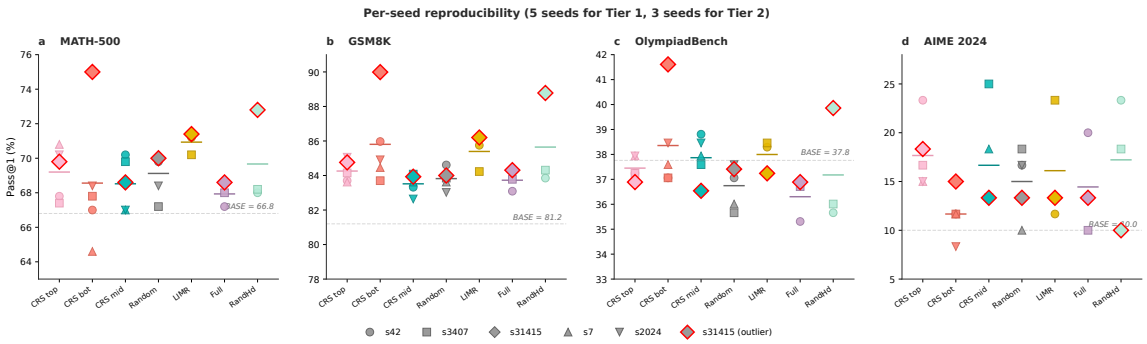


Figure 7: Per-seed accuracy for Tier 1 conditions (7B) across five primary benchmarks. Each dot is one (condition, seed) evaluation. CRS-bottom exhibits higher cross-seed variance than other conditions, particularly on MATH-500 and GSM8K.

Comparison	GSM8K	Minerva	MATH	Olymp.	AIME '24
LIMR−Rnd	+1.2	+0.9	+1.9	+1.3	−1.6
Top−LIMR	—	—	—	—	+1.7
Full−Rnd	−0.3	−2.5	−1.1	−0.4	+0.5
Full−LIMR	−1.5	−3.7	−3.0	−1.7	+2.1
RandH−Rnd	+2.0	+0.8	+0.1	+0.4	−0.9

Table 12: Tier 2 seed-level paired differences (n=3 seeds), 5 primary benchmarks.

Comparison	MATH	GSM8K	Olymp.	AIME '24
Top−Rnd	+1.8	+1.1	+1.2	+0.4
Bot−Rnd	−1.2	+0.1	+0.4	+0.8
Mid−Rnd	−1.1	−2.1	−0.7	+1.0
xmod−Rnd	−0.5	+1.0	+0.1	+0.2
Bot−Top	−3.0	−1.0	−0.8	+0.4

Table 13: 1.5B seed-level paired differences (n=3 seeds), 4 primary benchmarks.

Condition	Seed	GSM8K	Minerva	MATH	Olymp.	AIME
CRS-top	42	83.70	21.69	67.80	37.24	16.67
	3407	84.15	18.01	67.40	37.24	18.54
	31415	84.76	20.22	69.80	36.89	16.04
	2024	85.06	20.96	70.20	37.94	17.92
	7	83.62	23.16	70.80	37.94	13.96
CRS-mid	42	83.32	19.85	70.20	38.81	17.92
	3407	84.08	23.53	69.80	37.59	19.58
	31415	83.93	18.75	68.60	36.54	17.50
	2024	82.64	22.79	67.00	38.46	14.37
	7	83.62	19.49	67.00	37.94	16.04
CRS-bot	42	85.97	21.69	67.00	37.06	15.42
	3407	83.70	22.79	67.80	37.06	15.62
	31415	89.99	27.57	75.00	41.61	13.13
	2024	84.91	20.59	68.40	38.46	12.71
	7	84.46	19.85	64.60	37.59	17.50
Random	42	84.61	19.85	69.80	37.06	17.08
	3407	83.85	20.59	67.20	35.66	17.92
	31415	84.00	20.59	70.00	37.41	16.04
	2024	83.02	19.12	68.40	37.59	16.88
	7	83.62	18.01	70.20	36.01	14.58

Table 14: Tier 1 per-seed results (5 seeds), 5 primary benchmarks.

Cond.	Seed	GSM	Min.	M5	Oly.	AIME
LIMR	42	85.75	21.69	71.20	38.29	16.04
	3407	84.23	18.38	70.20	38.46	15.83
	31415	86.20	23.53	71.40	37.24	14.37
Full	42	83.09	19.12	67.20	35.31	19.17
	3407	83.78	19.12	68.00	36.71	16.88
	31415	84.31	14.34	68.60	36.89	16.46
RandH	42	83.85	17.28	68.00	35.66	17.08
	3407	84.31	20.96	68.20	36.01	17.50
	31415	88.78	23.16	72.80	39.86	12.71

Table 15: Per-seed results for LIMR, Full data, and RandHeads (3 seeds), 5 primary benchmarks.

Cond.	Seed	M5	GSM	Oly.	AIME
Top	42	63.60	76.27	29.90	8.54
	3407	68.80	82.41	35.14	10.00
	31415	63.80	73.69	29.20	11.04
Bot	42	62.40	78.32	30.59	11.46
	3407	63.00	78.09	31.12	8.96
	31415	61.80	72.93	30.07	10.21
Rnd	42	64.40	76.88	29.72	10.00
	3407	62.80	75.36	30.24	9.38
	31415	63.60	76.80	30.77	8.96
xmod	42	64.00	76.72	30.07	10.42
	3407	61.80	76.95	30.42	10.21
	31415	63.60	78.32	30.42	8.33

Table 16: 1.5B per-seed results (3 seeds), 4 primary benchmarks.

Benchmark	a	b	both	p
GSM8K (N=1319)	73	35	1081	3.3×10^{-4}
MATH-500 (N=500)	16	23	329	0.337
OlympiadBench (N=562)	23	14	193	0.188

Table 17: McNemar test: CRS-bottom vs. Random. a: bottom correct, random wrong; b: reverse; "both": both correct.

Pair	Jaccard	Chance
CRS-top ∩ CRS-bottom	0.000	0.053
CRS-top ∩ SR-top (easiest)	0.078	0.053
CRS-bot ∩ SR-bot (hardest)	0.055	0.053
Oly-CRS-top ∩ LIMR	0.153	0.163
Oly-CRS-bot ∩ LIMR	0.165	0.163

Table 18: Jaccard similarity between selection subsets (both at 10% of pool). Chance overlap ≈ 0.05 for same-pool pairs, ≈ 0.16 for cross-pool pairs.

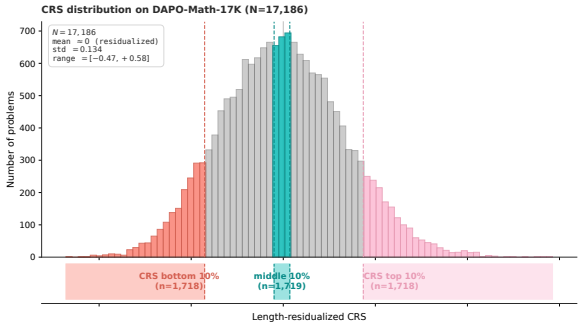


Figure 8: Distribution of length-residualized CRS (N=17,186). Colored bands: bottom-10% (pink), middle-10% (gray), top-10% (teal). The three selection deciles occupy distinct, non-overlapping regions of the distribution.

The same heads are critical to the untrained base model: on GSM8K, where the three models can be compared directly, ablation lowers base accuracy from 81.20% to 0.23%, a loss of 81.0 pp, against 86.0 and 84.2 pp for the two trained models. Had GRPO built new capability into these heads, ablation should harm the trained models far more than the base model; the matched losses instead indicate a pathway already present in the base model, which training uses but does not construct. Zeroing 46 whole heads is coarse and disables mathematical generation broadly (parse-failure rates rise to 73–88%), so this establishes that the heads are critical to mathematical generation, not that they form a reasoning-specific circuit.

	Minerva	GSM8K	Olymp.
DR(identified 46), pp	+1.23	+1.79	+2.04
rank among 7 sets	1 / 7	3 / 7	3 / 7
base acc., pre-ablation	16.18	81.20	37.76
base acc., post-ablation	0.74	0.23	0.00

Table 19: Post-training zero-ablation of the 46 identified heads.

C.5 Hierarchical Model of the Condition Effect

Seed-level pairing removes effects shared by both conditions within a seed, but not the condition-by-seed interaction. We therefore fit a binomial GLMM, `correct ~ condition + (1 + condition | seed) + (1 | problem)`, over 21,530 item-level outcomes on GSM8K, Minerva, and OlympiadBench (1,319 + 272 + 562 items, across five seeds and both conditions; MCMC, $\hat{R} = 1.002$). The condition coefficient is +0.256 log-odds, with 91.3% of the posterior above zero. The seed-level slope SD is $\sigma_t = 0.372$, of the same order as the condition coefficient, so the model treats the interaction as a measured quantity rather than an assumption.

Two further checks accompany this estimate. Removing any single seed leaves the sign of the CRS-bottom advantage unchanged on every benchmark. The exact sign-flip permutation test over five seed composites has a two-sided floor of 0.0625, so we report effect sizes rather than significance labels throughout.

C.6 The Two Ends of the CRS Ranking

CRS-top's point estimate exceeds Random on all three target benchmarks, so the relationship between CRS and downstream accuracy is not monotone. Read seed by seed, the two ends are asymmetric (Table 4): CRS-bottom's advantage is positive on 13 of 15 seed-by-benchmark cells, CRS-top's on 10, and CRS-middle's on 8. The pattern is best described as an asymmetric U—a large, seed-stable CRS-bottom effect alongside a weaker CRS-top effect not separable from CRS-middle at this seed count.

C.7 Difficulty-Filtering Subset Construction

The two difficulty bands of §4.4 are built from the rollout-estimated success rates of §5.1 (K=4 at τ=0.6; valid estimates for 16,478 of the 17,186 pool problems). Each band holds 1,647 problems, 10% of the SR-scored pool, matching the selection ratio used for the CRS deciles. Because K=4 quantizes SR onto five values, we select by rank rather than by threshold: **Diff-hard** takes the 1,647 lowest-SR problems with SR $>$ 0—at SR=0 every completion fails, the group reward has zero variance, and GRPO receives no gradient—and **Diff-easy** the 1,647 highest. Quantization makes Diff-hard entirely SR=0.25; conditions are labelled by measured mean SR. Jaccard overlap between these bands and the CRS deciles ranges from 0.030 to 0.069 across the four cross-pairings, against a chance level of ≈0.051, consistent with Table 18. Difficulty filtering therefore required $4 \times 16{,}478 = 65{,}912$ generations before any training run, against a single forward pass per problem for CRS.

D Circuit Identification Details

Figure 9 visualizes the bootstrap selection frequency across all 784 heads. L0H3 (red dashed box), excluded due to its outlier contrastive score (4.8× the runner-up), shows extreme loss increases on *both* probe types, suggesting a global information bottleneck rather than reasoning specialization.

Figure 10 shows the per-layer distribution of the 46 heads' contrastive scores. The signal is spread across the network, with concentrations near layer 0 and in the mid-to-late layers; layers 26–27 contribute few heads, suggesting that the final layers are more involved in output formatting than in reasoning computation.

E Sensitivity Analyses

The contrastive weight in Eq. 1. The weight 0.5 is a heuristic 50% discount on general-purpose

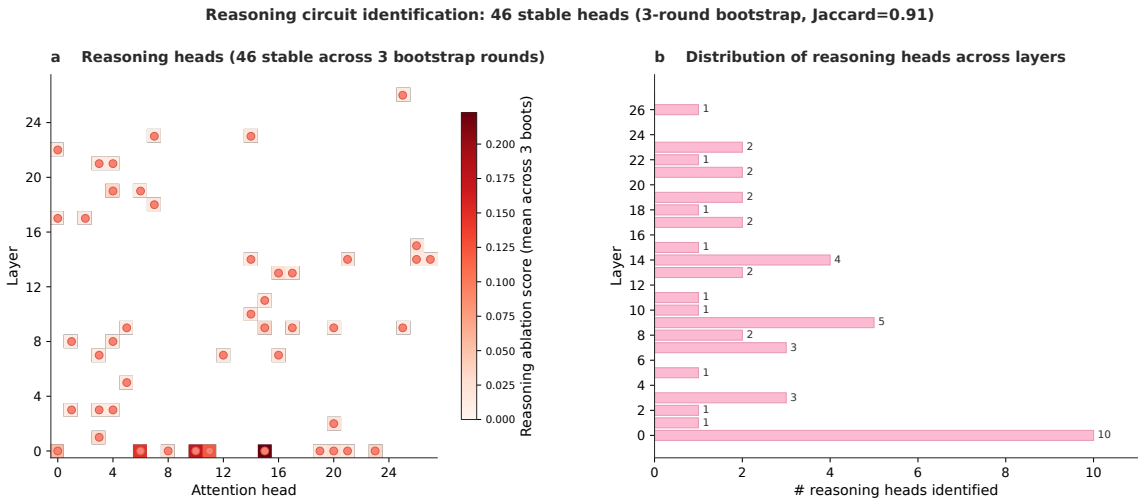


Figure 9: Bootstrap selection frequency across 784 heads (28×28). Dark cells: selected in 3/3 bootstrap rounds. Green boxes: 46 retained heads. Red dashed: L0H3 (excluded global bottleneck).

head importance, fixed before any downstream experiment. Varying w over $\{0, 0.25, 0.5, 0.75, 1.0\}$ and re-deriving both the head set and the full CRS ranking (Table 20), the head set and ranking are stable for $w \in [0.25, 1.0]$ (Jaccard ≥ 0.69, Spearman ≥ 0.94). The $w = 0$ row supports the design: with the contrastive term switched off the head set drifts substantially (Jaccard 0.48), evidence that the contrast against trivial-probe importance does real work.

w	**Jaccard (vs. .5)**	**Spearman (vs. .5)**	**heads**
0.00	0.484	0.605	46
0.25	0.745	0.937	43
0.50	1.000	1.000	46 (paper)
0.75	0.816	0.986	43
1.00	0.691	0.958	47

Table 20: Sensitivity of the head set and CRS ranking to the contrastive weight w.

The residualization form. We tested n, $\sqrt{n}$, and $\log n$ on all 17,186 problems (Table 21). $\log n$ gives the best fit and, by construction, zero residual correlation with log-length. The decisive quantity is downstream: bottom deciles under the three forms overlap pairwise at Jaccard 0.740–0.891 (chance ≈ 0.053), so the trained subsets would be substantially the same under any of the three.

Confidence intervals for Method-section correlations. Table 22 reports 10,000-resample bootstrap 95% CIs for the correlations quoted in §3. The interval on $\mathrm{Corr}(\Delta_r, \Delta_t)$ is wide because the covariance is heavy-tailed: a few high-magnitude

$f(n)$	R^2	$r(\text{resid}, n)$	$r(\text{resid}, \log n)$	**top-decile J.**
n	0.3223	0.000	0.056	0.048
$\sqrt{n}$	0.3412	-0.010	0.011	0.045
$\log n$	0.3436	0.013	0.000	0.058

Table 21: Residualization forms (N=17,186).

heads dominate it, and resamples omitting them pull the estimate toward zero—converging with the $w = 0$ result above, the reasoning-specific signal is concentrated in a few heads.

Quantity	r	**95% CI**	n
$\mathrm{Corr}(\Delta_r, \Delta_t)$	0.755	$[0.180, 0.921]$	784
between-round score corr.	0.9998	$[0.998, 0.9999]$	784
$\mathrm{Corr}(\mathrm{CRS}_{\mathrm{raw}}, n)$	0.568	$[0.557, 0.579]$	17,186
$\mathrm{Corr}(\mathrm{CRS}, n)$	0.013	$[-0.002, +0.028]$	17,186

Table 22: Bootstrap 95% CIs for Method-section correlations.

F Training Curves

Figure 11 shows the full training reward trajectories for the conditions plotted, with shaded bands indicating ± 1 standard deviation across seeds. LIMR's reward rises steeply and saturates near 0.78 by step 600, while CRS-bottom plateaus at $\sim$0.27. Despite this large gap in training reward, CRS-bottom achieves stronger downstream performance on three medium-difficulty benchmarks (§4.7). The divergence in reward curves is visible from early training (step $\sim$200), suggesting that the training-reward–generalization decoupling is a persistent

property of the data subsets, not a late-training artifact.

G Implementation Details

Circuit identification. Three bootstrap rounds of the full 784-head contrastive ablation scan, each with independently resampled probe sets (50 reasoning + 50 trivial). Total compute: $\sim$8 GPU-hours on a single A100-80GB.

CRS scoring. Single forward pass per problem on the frozen Qwen2.5-Math-7B base model. Total compute for 17,186 DAPO problems: $\sim$10 minutes on a single A100. For the 8,523 Olympiads pool: $\sim$5 minutes.

Difficulty-filtering scores. Rollout-estimated success rates for the difficulty baseline required $K{=}4$ generations on 16,478 problems, i.e. 65,912 generations in total.

GRPO training. Each run: $\sim$38 hours on 8$\times$A100-80GB (wall time varies with cluster preemption). 58 training runs $\times$ 38h $\times$ 8 GPUs $\approx$ 17,600 GPU-hours total.

Evaluation. Greedy benchmarks: $\sim$20 minutes per cell. avg@8 benchmarks (AIME, AMC): $\sim$60 minutes per cell. Total evaluation compute: $\sim$550 GPU-hours.

Overall compute budget. $\sim$18,100 GPU-hours (A100-80GB equivalent), dominated by GRPO training. CRS computation (circuit identification + scoring) accounts for less than 0.1% of total compute.

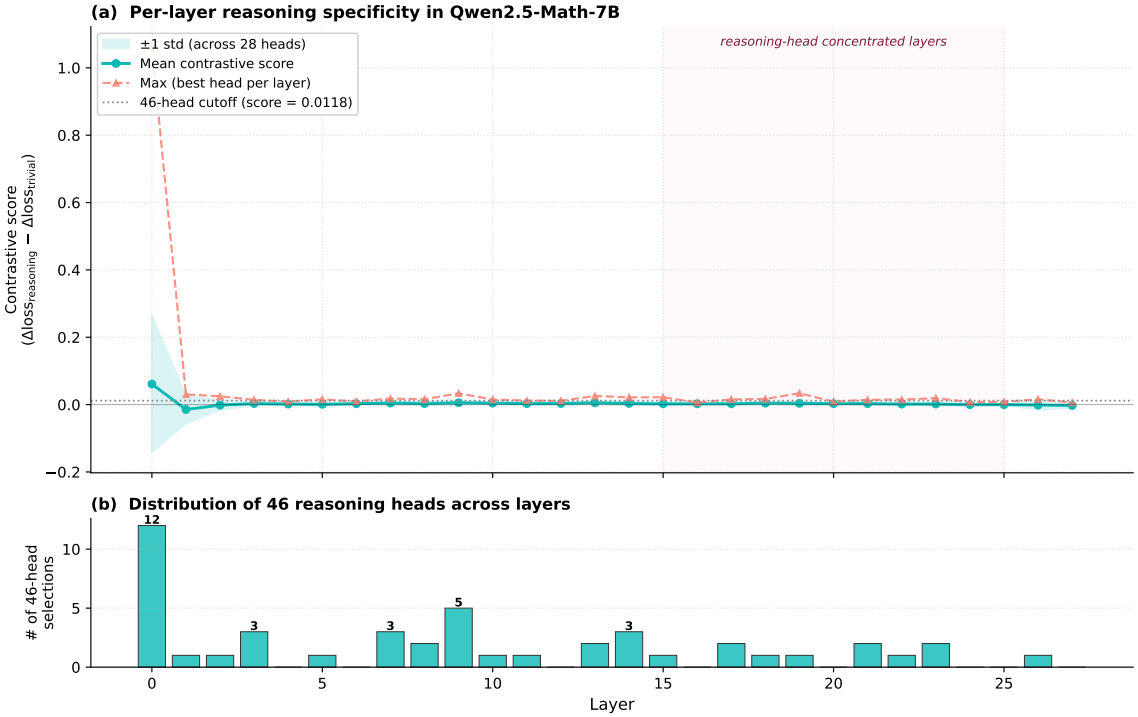


Figure 10: Per-layer contrastive scores. Each point is one attention head; highlighted points are the 46 selected heads.

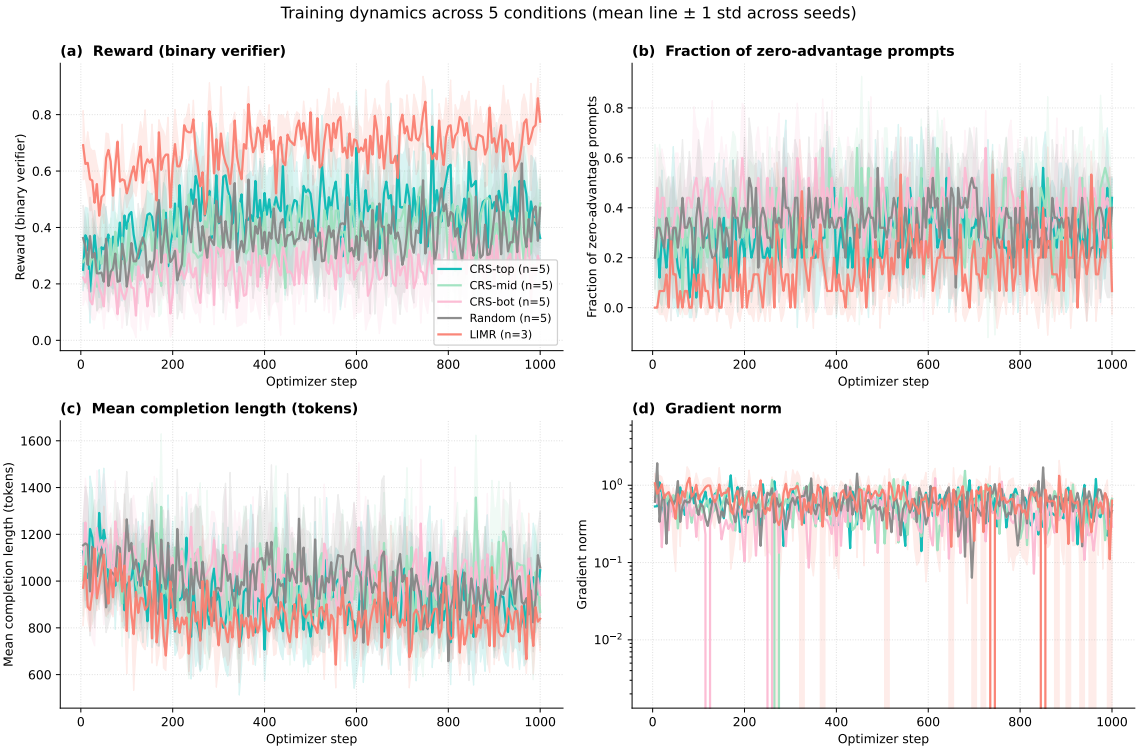


Figure 11: Training reward over 1,000 GRPO steps. Lines: seed-averaged reward; shaded bands: ±1 std. LIMR saturates at the highest reward (∼0.78); CRS-bottom at the lowest (∼0.27). Despite this gap, CRS-bottom produces the strongest downstream generalization (Table 3).