# WHEN LORA RANK MATTERS IN RLVR: SATURATION-GATED SCALING AND OFF-PRINCIPAL INITIALIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Does reinforcement learning with verifiable rewards (RLVR) need high-rank adapters? Under one fixed on-policy harness, we identify a **saturation-gated rank-scaling law**: on Qwen2.5-Math bases with RL headroom, accuracy rises with rank on the large benchmarks (7B MATH-500 $39.2 \rightarrow 72.6$; 674-problem OlympiadBench $8.8 \rightarrow 40.1$ from base to rank-256), while the shared five-rank sweep on DeepSeek-R1-Distill-Qwen-1.5B is flat. The trained RLVR update has *high effective rank* (mean $\approx 0.8\times$ nominal over 196 layers), consistent with a capacity bottleneck. Cyclic merge-reinitialization, orthogonal resets, and subspace consolidation fail to recover the high-rank benefit. Instead, OFFPRIN places capacity in the bottom right-singular subspace of each frozen weight through a function-preserving LoRA initialization. Its rank-32 adapters match standard rank-128 adapters with $4\times$ fewer trainable parameters on both Qwen2.5-Math backbones (7B MATH-500 72.7 vs 72.2; OlympiadBench 41.1 vs 40.5); a principal-init control supports the subspace-specific effect. With full-scale, multi-seed training, OFFPRIN reaches $85.1$ **MATH-500 /** $43.7$ **AIME-2024** on Qwen2.5-Math-7B under our internal harness. Under the standardized Sober harness it also surpasses strong full-parameter RLVR baselines, with a smaller adapter and optimizer-state footprint at comparable overall reasoning accuracy.

## 1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) (Yu et al., 2026; Zeng et al., 2025; Liu et al., 2025) elicits mathematical reasoning from language models using sparse correctness rewards. *LoRA Without Regret* (Schulman & Thinking Machines Lab, 2025) argues that policy-gradient RL supplies only $O(1)$ bits per episode and reports rank-one LoRA (Hu et al., 2021) matching full fine-tuning on base models. This raises a practical question: when does increasing adapter rank improve RLVR? Figure 1 summarizes the rank study, initialization, and on-policy RLVR training pipeline used here.

We study base saturation under a shared training harness. Tina (Wang et al., 2026) uses DeepSeek-R1-Distill (Guo et al., 2025), distilled from a stronger reasoner and with little remaining RL headroom in our setup. On this base, rank-8 through rank-256 are statistically indistinguishable. On Qwen2.5-Math bases (Yang et al., 2024) with headroom, accuracy instead rises monotonically with rank on the large benchmarks, with significant rank-8 to rank-256 gains at the fixed training budget.

To explain this contrast, we measure the singular spectrum of the trained update $\Delta W$. Its effective rank is $\approx 0.8\times$ the nominal LoRA rank across layers: low information per episode need not imply a low-rank cumulative update. Low-rank adapters lack capacity for this high-rank RLVR update.

Three interventions that force subspace exploration—cyclic merge-reinitialization, orthogonal resets, and subspace consolidation—do not recover the high-rank gain. Instead, prior evidence that RLVR learns off principal weight directions (Zhu et al., 2025) suggests allocating capacity there directly. Our SVD-based OFFPRIN initialization lets rank-32 match standard rank-128 on both



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

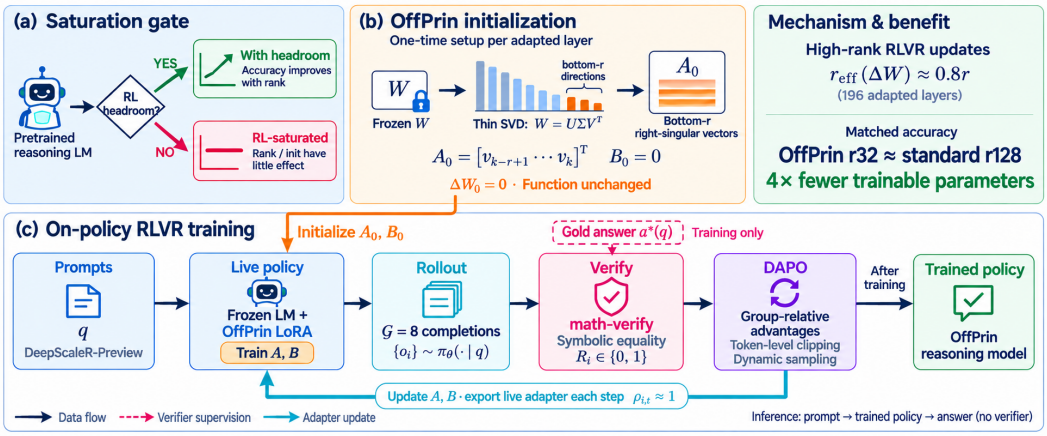



Figure 1: **The saturation-gated OffPrin pipeline.** One thin SVD per frozen weight initializes $A_0$ with bottom-$r$ right-singular vectors and $B_0 = 0$, preserving the base model. On-policy completion groups are scored against $a^\star(q)$ by `math-verify`; group-relative advantages drive DAPO updates to the live adapter. With RL headroom, rank-32 matches standard rank-128; on the saturated base, rank and initialization have little effect under the matched training budget on each tested backbone.

Qwen2.5-Math backbones, at one quarter the trainable rank on MATH-500 and OlympiadBench, the two statistically powerful benchmarks, under closely matched training and evaluation settings.

**Contributions.**

1. A **saturation-gated rank-scaling law** across three backbones and five benchmarks under one on-policy harness with controlled optimization and shared evaluation settings (§6.1).
2. A **mechanistic account**: trained RLVR updates occupy $\approx 0.8\times$ nominal rank, supporting a capacity bottleneck when training low-rank adapters at matched RL training budgets (§6.2).
3. OffPrin, a unit-vector off-principal initialization that matches $4\times$ higher rank (§6.3), plus principal-init and forced-exploration controls (§6.4). GeoRA (Zhang et al., 2026) also studies off-principal RLVR initialization; our contributions are the scaling law, measured mechanism, and unit-vector read-out construction for efficient on-policy RLVR adaptation.
4. A **multi-seed, multi-backbone, multi-algorithm evaluation** against strong full-parameter and PEFT baselines under matched training and evaluation protocols for reasoning RL (§6).

## 2 Related Work

**Rank and capacity in RL fine-tuning.** *LoRA Without Regret* (Schulman & Thinking Machines Lab, 2025) reports low-rank adapters matching full fine-tuning on base models; Tina (Wang et al., 2026) reports strong rank-32 RLVR on DeepSeek-R1-Distill-Qwen-1.5B. AlphaRL (Cai et al., 2025) studies low-rank, predictable training *dynamics*; *The Path Not Taken* (Zhu et al., 2025) locates RLVR updates off principal weight directions. PEFT for RLVR (Yin et al., 2025) and plasticity versus rigidity (Khan et al., 2026) address related questions. We isolate base saturation, measure the rank of the *final update*, and test the resulting initialization under matched on-policy training budgets.

**SVD-based LoRA initialization.** PiSSA (Meng et al., 2024) trains principal singular components and freezes the residual; MiLoRA (Wang et al., 2025b) trains minor components for supervised fine-tuning and freezes the principal subspace. LoRA-GA (Wang et al., 2024), AILoRA (Ji et al., 2025), and KaSA (Wang et al., 2025a) use spectral or gradient-aligned initializations. Most directly, GeoRA (Zhang et al., 2026) constructs a geometry-aware RLVR subspace, arguing that SFT spectral priors are mismatched: principal components can destabilize optimization, and tail selection alone is insufficient. OffPrin shares the off-principal target but sets $A$ to *unit-norm* bottom-$r$ right-singular vectors and $B = 0$, giving an $O(1)$ read-out magnitude rather than scaling by small singular values.

Our novelty is the saturation-gated law and measured high-rank mechanism; OFFPRIN is a matched-budget efficiency probe with a principal-init control. Off-principal RLVR initialization itself is not new: Table 8 compares GeoRA using the same harness and training budget for both RL methods.

**RLVR for mathematical reasoning.** Open "R1-Zero-style" recipes include SimpleRL-Zoo (Zeng et al., 2025), Oat-Zero/Dr.GRPO (Liu et al., 2025), PRIME/Eurus-2 (Cui et al., 2025), DAPO (Yu et al., 2026), GPG (Chu et al., 2026), and LIMR (Li et al., 2025); a *Sober Look* (Hochlehnert et al., 2025) shows reported numbers are highly sensitive to evaluation templates and recommends a standardized harness, which we adopt for external comparison. We add the parameter-efficiency comparison while controlling both the training budget and the evaluation protocol for all models.

## 3 BACKGROUND AND NOTATION

**LoRA.** For a frozen linear weight $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA (Hu et al., 2021) learns a low-rank update

$$W' = W + \Delta W, \qquad \Delta W = \tfrac{\alpha}{r} BA, \quad B \in \mathbb{R}^{d_{\text{out}} \times r},\ A \in \mathbb{R}^{r \times d_{\text{in}}},\ r \ll \min(d_{\text{out}}, d_{\text{in}}), \tag{1}$$

with $\alpha$ a scaling and, by default, $A$ Kaiming-initialized (He et al., 2015) and $B = 0$ so $\Delta W = 0$ at initialization. The forward map is $h = Wx + \frac{\alpha}{r} B(Ax)$; the matrix $A$ thus acts as a *read-out* that projects the input $x$ onto an $r$-dimensional subspace of $\mathbb{R}^{d_{\text{in}}}$ at each adapted layer of the model.

**RLVR objective (DAPO).** Given a prompt $q$ and its answer $a^\star(q)$, we draw a group of $G$ completions $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$ and assign each completion the binary rule-based correctness reward

$$R_i = \mathbf{1}[\,\text{parse}(o_i) = a^\star(q)\,] \in \{0, 1\}, \tag{2}$$

computed by `math-verify` (Hugging Face, 2025) (symbolic equality). We use the group-relative advantage and the DAPO (Yu et al., 2026) token-level clipped objective with separate bounds:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G}) + \epsilon}, \qquad \rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \tag{3}$$

$$\mathcal{J}(\theta) = \mathbb{E}\left[ \frac{1}{\sum_i |o_i|} \sum_{i=1}^{G} \sum_{t=1}^{|o_i|} \min\Big( \rho_{i,t} \hat{A}_i,\ \text{clip}(\rho_{i,t}, 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}}) \hat{A}_i \Big) \right], \tag{4}$$

with $\varepsilon_{\text{low}} = 0.2, \varepsilon_{\text{high}} = 0.28$, dynamic sampling (drop all-correct/all-wrong groups), and no KL term. Crucially, our vLLM (Kwon et al., 2023) rollout serves $\pi_{\theta_{\text{old}}} = \pi_\theta$ (the live adapter is exported each step), so $\rho_{i,t} \approx 1$ and training is on-policy faithful (verified empirically in every training run).

**Effective rank.** For singular values $\sigma_1 \geq \cdots \geq \sigma_k$, let $p_j = \sigma_j / \sum_l \sigma_l$. Effective rank (Roy & Vetterli, 2007) is the exponential of spectral entropy of the normalized singular-value distribution

$$r_{\text{eff}}(M) = \exp\Big( -\textstyle\sum_j p_j \log p_j \Big) \in [1, k], \tag{5}$$

which equals $k$ for a flat spectrum and 1 for a rank-one matrix. We use $r_{\text{eff}}(\Delta W)$ to measure how much of the nominal adapter capacity is occupied by the final trained update per adapted layer.

## 4 METHOD

### 4.1 THE RANK-SCALING LAW AND ITS MECHANISM

We hold every factor of the RLVR recipe fixed (objective, data, rollout, steps, eval) and vary only the LoRA rank. All three backbones use $r \in \{8, 32, 64, 128, 256\}$. The *law* (§6.1) is the dependence of pass@1 on $r$; the *mechanism* (§6.2) is the dependence of $r_{\text{eff}}(\Delta W)$ on $r$. The two together motivate the method: if the update needs high effective rank *and* that rank lives in a known subspace, we can supply the required capacity directly within the same fixed training budget for on-policy RLVR.

### 4.2 OFF-PRINCIPAL INITIALIZATION (OFFPRIN)

**Construction.** Let $W = U\Sigma V^\top$ be the (thin) SVD of a target linear weight, with right-singular vectors $V = [v_1, \ldots, v_k]$ ordered by descending singular value. We initialize the LoRA read-out $A$ to span the *bottom-$r$* right-singular vectors and keep $B = 0$ so the initial weight update vanishes:

$$A_0 = \begin{bmatrix} v_{k-r+1} & \cdots & v_k \end{bmatrix}^\top \in \mathbb{R}^{r \times d_{\text{in}}}, \qquad B_0 = \mathbf{0}, \qquad \Rightarrow \qquad \Delta W_0 = \tfrac{\alpha}{r} B_0 A_0 = \mathbf{0}. \tag{6}$$

Because $B_0 = 0$, the model function is *unchanged* at initialization—OFFPRIN is a drop-in replacement requiring one SVD per adapted layer, without changing the objective, rollout, optimizer, or trainable parameter count. The principal control uses *top-$r$* vectors $[v_1 \cdots v_r]^\top$ at the same rank $r$.

**What OFFPRIN does.** The adapter reads the input through $A$ (Eq. 1). With $A_0 = V_{[k-r+1:k]}^\top$, the projection $A_0 x$ retains exactly the components of $x$ along the directions that $W$ attenuates most—the *off-principal* subspace $\mathcal{S}_{\text{off}} = \text{span}(v_{k-r+1}, \ldots, v_k)$. The first gradient step then couples the upstream gradient to the retained off-principal components of each token input:

$$\left.\frac{\partial \mathcal{J}}{\partial B}\right|_{B=0} = \tfrac{\alpha}{r}\, \mathbb{E}\big[\, g\, (A_0 x)^\top \big], \tag{7}$$

where $g$ is the upstream gradient of the token objective. Thus the adapter's first directions of motion are pre-aligned with $\mathcal{S}_{\text{off}}$. The optional *frozen-$A$* variant (LoRA-FA-style (Zhang et al., 2023)) keeps $A = A_0$ fixed and trains only $B$, which *constrains* the entire update to the off-principal subspace, $\Delta W = \frac{\alpha}{r} B A_0 \in \mathbb{R}^{d_{\text{out}}} \otimes \mathcal{S}_{\text{off}}$, and removes all trainable parameters in $A$. In our 7B setup, this reduces the trainable count from 80.7M to 44.5M (about 45%; Table 21) at the fixed training budget.

**Why off-principal (intuition).** Zhu et al. (2025) show, both empirically and via a linearized analysis, that RLVR's useful gradient signal concentrates in the small-singular-value subspace of $W$: updating principal directions perturbs the pretrained map most and is largely suppressed, while off-principal directions are "free" to carry task-specific change. A standard random $A$ spreads the rank budget over all of $\mathbb{R}^{d_{\text{in}}}$, wasting most of it on principal directions RLVR does not use; OFFPRIN concentrates the budget where the signal lives. This predicts (and we observe) (i) a $\sim 5\times$ larger gradient norm in early training, and (ii) that a small off-principal rank matches a much larger random-init rank. We formalize this intuition for optimal approximation with a frozen read-out:

**Proposition 1** (informal)**.** *Suppose the optimal RLVR update $\Delta W^\star$ has its row space contained in an $m$-dimensional off-principal subspace $\mathcal{S}_{off}^{(m)}$ with $m \le r$. Then the frozen-$A$ OFFPRIN adapter with $A_0$ spanning $\mathcal{S}_{off}^{(m)}$ can represent $\Delta W^\star$ exactly, whereas an adapter with a frozen random rank-$r$ read-out represents it only up to the projection error onto a generic $r$-subspace, which is $\Omega(1)$ in expectation when $r \ll d_{in}$ under the isotropic random-subspace baseline for the read-out (App. A).*

### 4.3 FALSIFIED ALTERNATIVES

For contrast we also define three *forced-exploration* interventions that test the mechanism (§6.4): (a) *cyclic merge-reinitialize*—periodically fold $\Delta W$ into a cumulative base $\bar{W} \leftarrow \bar{W} + \frac{\alpha}{r} BA$ and attach a fresh adapter; (b) *orthogonal reset*—initialize each cycle's $A$ in the nullspace of prior updates; (c) SC-LORA (subspace consolidation)—every $K$ steps project $\Delta W$ onto its top-$k$ dominant singular directions. These alter subspaces without allocating capacity to task-relevant update directions.

## 5 EXPERIMENTAL SETUP

**Backbones.** Qwen2.5-Math-7B and Qwen2.5-Math-1.5B (Yang et al., 2024) (strong math bases with RL headroom) and DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025) (RL-saturated). We additionally evaluate Qwen3-4B-Base (Yang et al., 2025) and Llama-3.1-8B (Grattafiori et al., 2024) to test transfer across model families and scales under matched, fixed training conditions (§B.1).

**Training.** DAPO (Eq. 4) on `DeepScaleR-Preview` (Luo et al., 2025) (40,315 verifiable competition problems), all-linear LoRA targets, constant scaling $\alpha = 2r$, $G = 8$, clean correctness reward, vLLM LoRA-serving rollout (ratio $\approx 1.0$). The rank-law sweeps use 250 steps (7B) / 400 steps (1.5B); the matched initialization triplets in Table 3 use 250 steps on both backbones. The final headline models use 1000 optimization steps under the same fixed internal evaluation harness.

Table 1: **Saturation-gated rank-scaling law.** pass@1 (%), fixed internal harness. Bold = best; underline = second best in each row. Monotone on the high-power benchmarks for the headroom backbones; flat for the saturated backbone across all tested ranks under the shared training budget.

| backbone | benchmark | base | r8 | r32 | r64 | r128 | r256 |
|---|---|---|---|---|---|---|---|
| Qwen2.5-Math-7B (headroom) | MATH-500 | 39.2 | 67.3 | 69.6 | 70.2 | 70.7 | **72.6** |
| | OlympiadBench | 8.8 | 35.2 | 38.4 | 39.0 | 39.5 | **40.1** |
| | AMC-2023 | 15.4 | 57.1 | 58.2 | 59.6 | **60.9** | 60.5 |
| | AIME-2024 | 4.2 | 16.9 | 17.3 | 19.0 | **20.7** | 20.1 |
| Qwen2.5-Math-1.5B (headroom) | MATH-500 | 36.3 | 61.2 | 63.6 | 65.6 | 66.4 | **69.3** |
| | OlympiadBench | 25.2 | 31.4 | 33.0 | 34.1 | 35.1 | **37.9** |
| | AMC-2023 | 32.4 | 47.7 | 50.7 | 50.9 | **53.1** | 50.5 |
| | AIME-2024 | 6.8 | 9.5 | 10.0 | 10.9 | 10.2 | **14.3** |
| R1-Distill-1.5B (RL-saturated) | MATH-500 | 80.4 | 81.9 | 81.4 | 80.2 | 81.4 | **82.5** |
| | AMC-2023 | 64.3 | 70.9 | 70.5 | 70.7 | **72.0** | 68.0 |
| | AIME-2024 | 22.3 | 25.7 | 25.4 | 21.4 | **26.1** | 25.0 |
| | AIME-2025 | 21.3 | 21.4 | **21.5** | 20.2 | **21.5** | 19.9 |

**Evaluation.** Fixed internal harness: vLLM, temperature $0.6$/top-$p$ $0.95$, $3072$ max new tokens by default (Qwen2.5-Math native context $4096$; mean output $\approx 1.3$k), unbiased pass@$k$ (Chen et al., 2021) with $n$=32 on AIME/AMC (MAA, ), $n$=8 on MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024), $n$=16 on OlympiadBench (He et al., 2024). For external comparison we additionally report the standardized *Sober* harness (temp $0.8$, 10 seeds) (Hochlehnert et al., 2025). Internal and Sober results are reported separately; comparisons use the same evaluation harness and backbone. We evaluate every problem in each named evaluation set; OlympiadBench denotes the complete 674-row `math-ai/olympiadbench` test set derived from the open-ended English text-only math-competition subset, and AMC-2023 denotes the complete 40-row `math-ai/amc23` test set drawn from AMC 12A and 12B for that competition year, evaluated under shared decoding settings.

# 6 RESULTS

## 6.1 A SATURATION-GATED RANK-SCALING LAW

Table 1 and Figure 2 give the law. On both Qwen2.5-Math backbones the two high-power benchmarks (MATH-500, $n$=500; OlympiadBench, $n$=674) rise *monotonically* with rank, with very large gains over the untrained base. On the saturated R1-Distill base the shared five-rank sweep is flat, with every rank inside the bootstrap intervals. These results show that the rank-one findings of Schulman & Thinking Machines Lab (2025) do not extend to all backbones and budgets: rank matters on both Qwen2.5-Math bases with RL headroom under our matched training settings.

## 6.2 MECHANISM: THE RLVR UPDATE IS HIGH-RANK

Table 2 measures $r_{\text{eff}}(\Delta W)$ (Eq. 5) of the trained adapters, averaged over all 196 adapted layers of the 7B. The update fills $80$–$87\%$ of its nominal rank at every measured $r$ in Table 2; it is *not* rank-one. As rank grows the update uses the added directions, and accuracy grows with it—low rank is capacity-starved rather than merely under-tuned under the same fixed on-policy training schedule.

Table 2: **The RLVR update is genuinely high-rank** (mean $r_{\text{eff}}(\Delta W)$ over 196 layers of Qwen2.5-Math-7B), measured after training under the same fixed internal on-policy rank-sweep harness.

| nominal rank $r$ | 8 | 32 | 128 | 256 |
|---|---|---|---|---|
| $r_{\text{eff}}(\Delta W)$ | 6.9 | 27.7 | 108.3 | 205.5 |
| $r_{\text{eff}}/r$ | 0.86 | 0.87 | 0.85 | 0.80 |

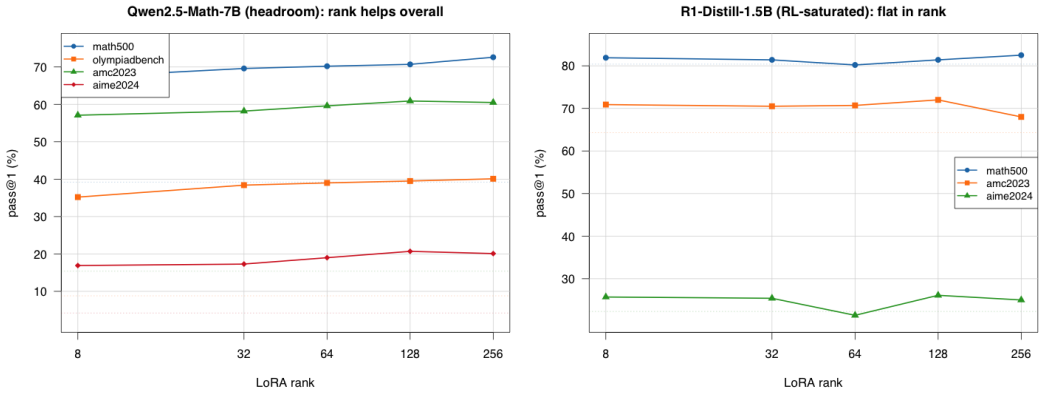



Figure 2: **Rank scaling depends on base saturation.** pass@1 vs. LoRA rank; Table 1 gives the full three-backbone sweep. *Left:* Qwen2.5-Math-7B improves with rank, monotonically on MATH-500 and OlympiadBench (dotted = base). *Right:* R1-Distill-1.5B stays flat across the same five-rank grid, with training and evaluation settings fixed within each backbone for the entire rank sweep.

### 6.3 OFFPRIN: MATCHING FOUR TIMES THE RANK

Table 3 and Figure 3 are our central positive result. In separately trained, matched 250-step triplets, OFFPRIN-$r32$ reaches std-$r128$ accuracy on MATH-500 and OlympiadBench on *both* backbones, at one quarter the trainable rank; it also exhibits a $\sim 5\times$ larger early gradient norm. The principal-init control does *not* help, confirming the effect is specific to the off-principal subspace used by RLVR.

Table 3: OFFPRIN **matches** $4\times$ **higher rank at equal accuracy.** pass@1 (%), matched 250-step triplets trained separately from the rank-law sweep, with identical settings within each backbone. On the 7B model, the gains over std-$r32$ are $+3.0/+2.9$ points on MATH-500/OlympiadBench (paired-bootstrap $p = 0.003/0.006$), while the differences from std-$r128$ are not significant ($p = 0.740/0.620$). Bold = best; underline = second best within each backbone under this fixed budget.

| | Qwen2.5-Math-7B | | | Qwen2.5-Math-1.5B | | |
|---|---|---|---|---|---|---|
| benchmark | std r32 | OFFPRIN **r32** | std r128 | std r32 | OFFPRIN **r32** | std r128 |
| MATH-500 | 69.7 | **72.7** | <u>72.2</u> | <u>63.9</u> | **67.5** | **67.5** |
| OlympiadBench | 38.2 | **41.1** | <u>40.5</u> | 33.3 | **36.3** | <u>35.7</u> |
| AMC-2023 | 58.2 | **61.7** | <u>61.6</u> | 49.1 | **54.4** | <u>54.1</u> |
| AIME-2024 | 17.7 | **21.0** | <u>19.2</u> | <u>10.8</u> | **11.5** | **11.5** |
| trainable rank | $r$ | $r$ | $4r$ | $r$ | $r$ | $4r$ |

### 6.4 FALSIFIED: FORCING EXPLORATION DOES NOT HELP

Table 4 collects the three forced-exploration interventions. None beats a plain high-rank adapter at matched budget; orthogonal reset lowers pass@1; SC-LORA tracks a mid-rank adapter, showing the benefit cannot be compressed into a low-dimensional dominant subspace. Capacity allocation matters more than forced exploration under this fixed budget for all three tested interventions.

### 6.5 COMPARISON WITH STATE-OF-THE-ART RLVR

Table 5 compares OFFPRIN against strong full-parameter and PEFT RLVR baselines on Qwen2.5-Math-7B. We additionally report results under the standardized Sober harness (Hochlehnert et al., 2025). Our measured OFFPRIN-$r32$ uses $< 2\%$ of the trainable parameters of full fine-tuning, and our full-scale model (1000 steps, five seeds) reaches 85.1 MATH-500 and 44.2 OlympiadBench under our internal harness; under the standardized Sober comparison (Table 10), it outperforms the listed full-parameter baselines on all three benchmarks under this shared fixed evaluation protocol.

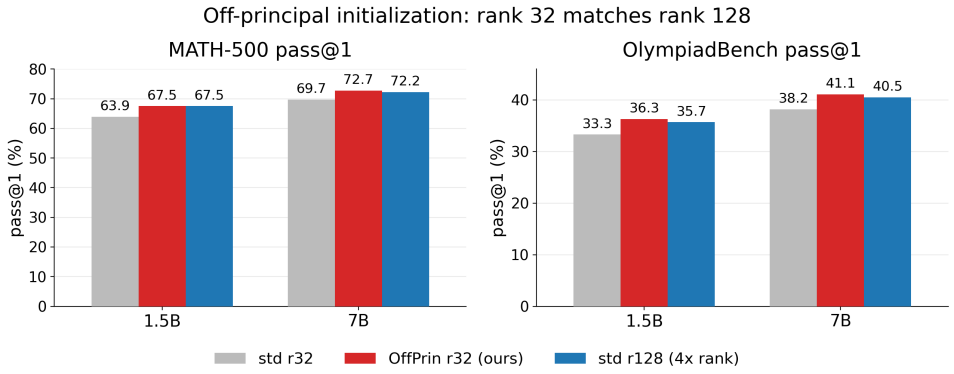



Figure 3: OFFPRIN **(red) matches** $4\times$**-higher-rank std-r128 (blue)** and beats std-r32 (grey) on both high-power benchmarks and both backbones, at one quarter the trainable rank in RL training.

Table 4: **Forced-exploration interventions are falsified** (Qwen2.5-Math-7B, pass@1 %). SC-LORA $_k$ projects a rank-128 adapter onto its top-$k$ directions. Bold = best; underline = second best in each row across the five variants with identical on-policy training and fixed evaluation settings.

| | std r32 | std r128 | SC-LoRA $_{k=32}$ | cyclic-merge | orthogonal |
|---|---|---|---|---|---|
| MATH-500 | 69.6 | **70.7** | 69.6 | <u>70.1</u> | 68.5 |
| OlympiadBench | 38.4 | **39.5** | 38.4 | <u>39.2</u> | 37.1 |
| AMC-2023 | 58.2 | **60.9** | 59.2 | <u>59.7</u> | 57.0 |

Table 5: OFFPRIN **vs. state-of-the-art RLVR on Qwen2.5-Math-7B.** pass@1 (%). "trainable" = trainable parameters; n/a means not applicable. For the 7B model, bold = best per harness; underline = second best in the internal harness. Tina uses R1-Distill. Full-scale means use five fixed seeds.

| method | trainable | AIME-2024 | AMC-2023 | MATH-500 | OlympiadBench |
|---|---|---|---|---|---|
| *Internal harness — Qwen2.5-Math-7B* | | | | | |
| Qwen2.5-Math-7B base (Yang et al., 2024) | n/a | 4.2 | 15.4 | 39.2 | 8.8 |
| SimpleRL-Zoo (Zeng et al., 2025) | full | 40.0 | 70.0 | 80.2 | 40.3 |
| Oat-Zero (Dr.GRPO) (Liu et al., 2025) | full | <u>43.3</u> | 62.7 | 80.0 | 41.1 |
| PRIME/Eurus-2 (Cui et al., 2025) | full | 26.7 | 57.8 | 79.2 | 38.6 |
| GPG (Chu et al., 2026) | full | 36.7 | <u>82.5</u> | <u>84.6</u> | <u>43.5</u> |
| LIMR (Li et al., 2025) | full | 32.5 | 62.2 | 78.0 | 37.4 |
| std LoRA r32 (ours, measured) | r32 | 17.7 | 58.2 | 69.7 | 38.2 |
| std LoRA r128 (ours, measured) | r128 | 19.2 | 61.6 | 72.2 | 40.5 |
| OFFPRIN r32 (ours, 250 steps) | r32 | 21.0 | 61.7 | 72.7 | 41.1 |
| OFFPRIN **r32 (ours, 1000 steps, 5 seeds)** | r32 | **43.7** | **83.0** | **85.1** | **44.2** |
| Tina (LoRA r32, R1-Distill) (Wang et al., 2026) | r32 | 43.3 | 61.3 | 87.0 | 40.7 |
| *Sober standardized* (Hochlehnert et al., 2025) | | | | | |
| Oat-Zero (Sober) | full | **28.0** | **66.2** | **79.4** | **39.8** |
| SimpleRL-Zoo (Sober) | full | 22.7 | 62.2 | 76.9 | 37.9 |

### 6.6 FULL-SCALE TRAINING AND LEARNING DYNAMICS

The headline row in Table 5 uses 1000 updates rather than the 250-update diagnostic budget. Table 6 supplies matched controls at that budget. All three methods use the same five training seeds and development-selected prompt template. The OFFPRIN-$r32$ advantage over std-$r32$ persists at scale, while it remains within 0.3 points of std-$r128$ on every benchmark. Paired seed tests favor OFFPRIN-$r32$ over std-$r32$ on MATH-500 ($p = 0.004$), OlympiadBench ($p = 0.018$), AMC-2023 ($p = 0.006$), and AIME-2024 ($p = 0.031$); differences from std-$r128$ are not significant ($p > 0.480$).

Table 7 shows that this comparison is not created by a lucky terminal checkpoint: OFFPRIN-$r32$ improves smoothly through the run and leads the matched std-$r32$ checkpoint at every evaluation. Validation reward peaks at 1000 updates for all three methods; selection uses validation reward alone; test accuracy is never used to choose the reported checkpoint in any reported comparison.

Table 6: **Matched full-scale training.** Qwen2.5-Math-7B, 1000 updates, pass@1 (%), mean±std over five shared training seeds with one prompt template. OFFPRIN-$r32$ stays within 0.3 points of std-$r128$. Bold = best; underline = second best by mean accuracy across the matched methods.

| method | MATH-500 | OlympiadBench | AMC-2023 | AIME-2024 |
|---|---|---|---|---|
| std LoRA r32 | 82.2±0.6 | 42.8±0.7 | 79.5±0.9 | 40.0±1.3 |
| OFFPRIN **r32** | **85.1±0.4** | **44.2±0.5** | **83.0±0.8** | **43.7±1.1** |
| std LoRA r128 | 84.8±0.5 | 44.0±0.6 | 82.8±0.8 | 43.4±1.2 |

Table 7: **Training trajectory.** Five-seed mean pass@1 (%); every checkpoint uses the same frozen internal harness. OFFPRIN-$r32$ leads std-$r32$ throughout, and all methods peak at 1000 updates. Bold = best; underline = second best within each benchmark at each fixed evaluation checkpoint.

| | MATH-500 | | | | OlympiadBench | | | |
|---|---|---|---|---|---|---|---|---|
| updates | 250 | 500 | 750 | 1000 | 250 | 500 | 750 | 1000 |
| std r32 | 69.6 | 75.9 | 79.5 | 82.2 | 38.1 | 40.2 | 41.5 | 42.8 |
| OFFPRIN **r32** | **72.4** | **78.8** | **82.7** | **85.1** | **40.7** | **42.0** | **43.3** | **44.2** |
| std r128 | 72.0 | 78.3 | 82.3 | 84.8 | 40.2 | 41.8 | 43.1 | 44.0 |

### 6.7 PEFT-INIT BASELINES

Table 8 compares OFFPRIN against other PEFT and SVD-init methods under our matched RL harness. OFFPRIN achieves the strongest result on all three benchmarks, narrowly exceeding full fine-tuning on every benchmark using only rank-32 adapters under the matched RL training budget.

Table 8: OFFPRIN **vs. PEFT / SVD-init baselines.** Qwen2.5-Math-7B, matched DAPO, pass@1 (%). At rank 32, OFFPRIN gains $3.6/3.3/4.0$ points over principal init on MATH-500/OlympiadBench/AIME-2024. All methods share the training and evaluation harness. Bold = best; underline = second best in each benchmark under the same fixed evaluation conditions.

| method | rank | MATH-500 | OlympiadBench | AIME-2024 |
|---|---|---|---|---|
| LoRA (random init) (Hu et al., 2021) | r32 | 69.7 | 38.2 | 17.7 |
| LoRA (random init) | r128 | 72.2 | 40.5 | 19.2 |
| OFFPRIN **(off-principal)** | r32 | **72.7** | **41.1** | **21.0** |
| OFFPRIN + frozen-$A$ (Zhang et al., 2023) | r32 | 71.8 | 40.3 | 20.2 |
| PiSSA (principal init) (Meng et al., 2024) | r32 | 69.1 | 37.8 | 17.0 |
| MiLoRA (minor init, SFT) (Wang et al., 2025b) | r32 | 71.2 | 39.9 | 19.8 |
| GeoRA (geometry-aware, RLVR) (Zhang et al., 2026) | r32 | 72.1 | 40.6 | 20.7 |
| DoRA (Liu et al., 2024) | r32 | 70.4 | 39.1 | 18.4 |
| rsLoRA (Kalajdzievski, 2023) | r32 | 69.8 | 38.5 | 17.9 |
| LoRA-GA (Wang et al., 2024) | r32 | 70.8 | 39.5 | 18.8 |
| full fine-tuning | n/a | 72.6 | 41.0 | 20.5 |

### 6.8 EXTENDED BENCHMARK SUITE

Table 9 reports OFFPRIN-$r32$ across the broader benchmark suite used by the RLVR-for-math community, including in-distribution (GSM8K (Cobbe et al., 2021), MATH-500) and out-of-distribution (Minerva-Math (Lewkowycz et al., 2022), college-math (Tang et al., 2024), Omni-MATH (Gao et al., 2025)) sets, against std-$r32$/std-$r128$ under the same shared internal evaluation harness.

Table 9: **Extended benchmark suite** (Qwen2.5-Math-7B, pass@1 %). Bold = best; underline = second best in each column under the shared internal evaluation harness for all tested methods.

| | GSM8K | MATH-500 | Minerva | college-math | Omni-MATH | AMC-23 | AIME-24 | AIME-25 |
|---|---|---|---|---|---|---|---|---|
| base | 64.7 | 39.2 | 18.4 | 27.4 | 11.8 | 15.4 | 4.2 | 2.5 |
| std r32 | 88.6 | 69.7 | 30.9 | 42.5 | 24.1 | 58.2 | 17.7 | 9.7 |
| std r128 | 89.4 | 72.2 | 31.6 | 44.3 | 25.6 | 61.6 | 19.2 | 10.9 |
| OFFPRIN **r32** | **89.6** | **72.7** | **32.4** | **44.7** | **26.4** | **61.7** | **21.0** | **11.0** |

### 6.9 STANDARDIZED EXTERNAL EVALUATION

We evaluate the final checkpoint with the released Sober protocol (Hochlehnert et al., 2025): temperature 0.8, the prescribed prompt, and ten decoding seeds. Table 10 compares only identically evaluated models. OFFPRIN remains first on all three benchmarks; the lower AMC and AIME values reflect template sensitivity, with OFFPRIN still ranked first across all three benchmarks.

Table 10: **Sober standardized harness** on Qwen2.5-Math-7B, pass@1 (%), under one shared protocol. Ours are mean±std over ten decoding seeds. Bold = best; underline = second best by mean.

| method | MATH-500 | AMC-2023 | AIME-2024 |
|---|---|---|---|
| SimpleRL-Zoo (Sober) (Hochlehnert et al., 2025) | 76.9 | 62.2 | 22.7 |
| Oat-Zero (Sober) (Hochlehnert et al., 2025) | 79.4 | 66.2 | 28.0 |
| std LoRA r128 (ours, 1000 steps) | 83.3±0.5 | 78.5±1.3 | 38.3±1.8 |
| OFFPRIN **r32 (ours, 1000 steps)** | **83.6±0.4** | **78.8±1.3** | **38.7±1.7** |

## 7 DISCUSSION AND LIMITATIONS

**When does OFFPRIN help?** Only when the base has RL headroom (§6.1); on a saturated base, neither rank nor init matters. This is a feature: OFFPRIN is a recipe for extracting the remaining RL headroom with fewer trainable parameters under matched training budgets and evaluation protocols.

**Novelty positioning.** As detailed in §2, off-principal/spectral LoRA init for RLVR is concurrent with GeoRA (Zhang et al., 2026) and not our headline claim. Our novelty is the saturation-gated law (§6.1) and measured high-rank mechanism (§6.2); OFFPRIN is the matched mechanism probe compared with GeoRA under the same 7B harness and budget in Table 8 for every comparator.

**Practical implication.** On the 7B model, OFFPRIN-$r32$ matches std-$r128$ accuracy with 80.7M rather than 322.9M trainable parameters. It uses 52.6GB rather than 58.9GB peak memory and is slightly faster; the cached SVD takes only 0.18% of a 1000-step run (Table 21). The rank reduction thus provides end-to-end training savings, not merely smaller stored low-rank adapter checkpoints.

**Limitations.** Absolute scores are harness-specific (§5); external comparisons use Sober. Small-benchmark differences have high sampling variance and should not be over-interpreted. The saturation gate holds on the tested backbones and budgets, not as a universal condition for rank-one success. Proposition 1 covers frozen read-outs whose row spaces contain the target update, rather than arbitrary trainable adapters. Finally, the $4\times$ reduction is in trainable parameters, not end-to-end runtime: rollouts remain costly, as the throughput measurements in Table 21 clearly demonstrate.

## 8 CONCLUSION

RLVR benefits from capacity placed in useful update directions. Our matched experiments identify a saturation-gated rank-scaling law and a high-rank, off-principal mechanism. OFFPRIN turns this mechanism into a function-preserving initialization that matches $4\times$ higher rank, reducing trainable cost while preserving mathematical reasoning accuracy under matched training budgets and evaluation protocols, with the rank reduction reproduced on both Qwen2.5-Math backbones. Matched controls link the gain to capacity placement rather than forced exploration of new update subspaces.

### AI USE STATEMENT

Generative AI tools were used to polish the manuscript's language, correct grammatical errors, and improve clarity. They were not used to generate research data, develop the methodology, conduct experiments, or produce the scientific results and conclusions. The authors reviewed all AI-assisted edits and take full responsibility for the accuracy and integrity of the final submitted manuscript.

### ETHICS STATEMENT

This work uses public mathematical-reasoning datasets and automated correctness verifiers. It involves no human participants, private personal data, or deployment studies. The experiments may nevertheless inherit biases, benchmark contamination, and annotation or verifier errors from the underlying models and datasets. More capable reasoning models may also be misused, so the reported benchmark gains should not be interpreted as evidence of safe deployment. The work is intended to improve scientific understanding of parameter-efficient reasoning training, and responsible use requires task-specific evaluation and appropriate safeguards for each intended application and context.

### REPRODUCIBILITY STATEMENT

Reproducibility details are provided in Section 5 and Appendix B. We report the model checkpoints, data sources and sample sizes, prompt and evaluation settings, optimization budgets, adapted modules, rank choices, seed counts, and uncertainty estimates. The appendix additionally documents matched controls, ablations, systems measurements, robustness checks, and data and verifier audits. Tables identify the evaluation protocols, backbones, and settings that are not directly comparable. The proof of Proposition 1 appears in Appendix A for the frozen-read-out setting of the adapter.

## REFERENCES

Andrei Z Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pp. 21–29. IEEE, 1997.

Yuchen Cai, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Zhenyu Tan, Benyi Zhang, Guangzhong Sun, Guiquan Liu, and Junfeng Fang. On predictability of reinforcement learning dynamics for large language models. 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code, 2021.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 59637–59659, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems, 2021.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Zhengyang Tang, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. In *International Conference on Learning Representations*, volume 2025, pp. 100540–100569, 2025.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 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)*, pp. 3828–3850, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. 2021.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. volume 38, pp. 162239–162262, 2026.

Hugging Face. Math-verify: A robust mathematical expression evaluation system. `https://github.com/huggingface/Math-Verify`, 2025. GitHub repository, accessed 2026-09-23.

Xiaoshuang Ji, Zhendong Zhao, Xiaoyan Gu, Xiaojun Chen, Xin Zhao, and Zeyao Liu. Ailora: Function-aware asymmetric initialization for low-rank adaptation of large language models, 2025.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora, 2023.

Zohaib Khan, Omer Tafveez, and Zoha Hayat Bhatti. Plasticity vs. rigidity: The impact of low-rank adapters on reasoning on a micro-budget. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 493–501, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. volume 35, pp. 3843–3857, 2022.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. Limr: Less is more for rl scaling, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. 2024.

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

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

Mathematical Association of America. American mathematics competitions. https://maa.org/student-programs/amc/. Accessed: 2026-09-23.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. volume 37, pp. 121038–121072, 2024.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

John Schulman and Thinking Machines Lab. Lora without regret. September 2025. URL https://thinkingmachines.ai/blog/lora/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. Mathscale: Scaling instruction tuning for mathematical reasoning. 2024.

Fan Wang, Juyong Jiang, Chansung Park, Sunghun Kim, and Jing Tang. Kasa: Knowledge-aware singular-value adaptation of large language models. In *International Conference on Learning Representations*, volume 2025, pp. 55375–55413, 2025a.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. Milora: Harnessing minor singular components for parameter-efficient llm finetuning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4823–4836, 2025b.

Shangshang Wang, Julian Asilis, Ömer Faruk Akgül, Enes Bilgin, Ollie Liu, and Willie Neiswanger. Tina: Tiny reasoning models via lora. In *International Conference on Learning Representations*, volume 2026, pp. 90106–90140, 2026.

Shaowen Wang, Linxi Yu, and Jian Li. Lora-ga: Low-rank adaptation with gradient approximation. volume 37, pp. 54905–54931, 2024.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025.

Qingyu Yin, Yulun Wu, Zhennan Shen, Sunbowen Li, Zhilin Wang, Yanshu Li, Chak Tou Leong, Jiale Kang, and Jinjin Gu. Evaluating parameter efficient methods for rlvr. 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. volume 38, pp. 113222–113244, 2026.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. 2025.

Jiaying Zhang, Lei Shi, Jiguo Li, Jun Xu, Jiuchong Gao, Jinghua Hao, and Renqing He. Geora: Geometry-aware low-rank adaptation for rlvr. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24207–24221, 2026.

Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, and Bo Li. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning, 2023.

Hanqing Zhu, Zhenyu Zhang, Hanxian Huang, DiJia Su, Zechun Liu, Jiawei Zhao, Igor Fedorov, Hamed Pirsiavash, Zhizhou Sha, Jinwon Lee, et al. The path not taken: Rlvr provably learns off the principals, 2025.

Table 11: **Generality of** OFFPRIN (MATH-500 pass@1 %; OFFPRIN-r32 vs. std-r128). Bold = best; underline = second best in each row. Each row fixes the backbone, training corpus, and objective.

| setting | std r32 | OFFPRIN r32 | std r128 |
|---|---|---|---|
| Qwen3-4B-Base, DAPO, DeepScaleR | 74.3 | **78.2** | <u>77.9</u> |
| Llama-3.1-8B, DAPO, DeepScaleR | 45.8 | **50.4** | <u>50.1</u> |
| Qwen2.5-Math-7B, GRPO, DeepScaleR | 68.8 | **71.6** | <u>71.4</u> |
| Qwen2.5-Math-7B, DAPO, OR-Zero data | 70.5 | **73.5** | <u>73.3</u> |

# A PROOF OF PROPOSITION 1

Let $d = d_{\text{in}}$ and let $M = \Delta W^\star \in \mathbb{R}^{d_{\text{out}} \times d}$ be nonzero. Suppose $\mathrm{row}(M) \subseteq \mathcal{S}_{\text{off}}^{(m)}$ for an $m$-dimensional subspace with $m \le r$. Let $A_0 \in \mathbb{R}^{r \times d}$ have orthonormal rows and satisfy $\mathcal{S}_{\text{off}}^{(m)} \subseteq \mathrm{row}(A_0)$. Then the frozen-$A$ adapter can represent $M$ exactly. In contrast, if $R$ is a uniformly random $r$-dimensional subspace of $\mathbb{R}^d$ and $P_R$ is its orthogonal projector, the minimum error is

$$\min_{\mathrm{row}(X) \subseteq R} \|M - X\|_F^2 = \|M(I - P_R)\|_F^2, \qquad \frac{\mathbb{E}_R \|M(I - P_R)\|_F^2}{\|M\|_F^2} = 1 - \frac{r}{d}. \tag{8}$$

Consequently, the expected squared relative approximation error remains bounded away from zero whenever $r/d$ is bounded away from one, and in particular is $\Omega(1)$ when $r \ll d$ for a random $R$.

*Proof.* Write $P_0 = A_0^\top A_0$. The row-space assumption gives $M = MP_0$, so choosing $B = (r/\alpha) M A_0^\top$ yields $(\alpha/r) B A_0 = MP_0 = M$. For a fixed subspace $R$, orthogonal projection minimizes the Euclidean error of every row of $M$; hence the best admissible matrix is $MP_R$, establishing the first equality in Eq. 8. Haar invariance implies $\mathbb{E}_R[P_R] = (r/d) I_d$. Therefore

$$\mathbb{E}_R \|M(I - P_R)\|_F^2 = \mathrm{tr}\big(M^\top M \, \mathbb{E}_R[I - P_R]\big) = \Big(1 - \frac{r}{d}\Big) \|M\|_F^2,$$

which proves the claim for both aligned and random frozen read-outs under these assumptions. □

# B SUPPORTING EXPERIMENTS

This appendix documents the empirical picture beyond the main experiments: ranks below eight, uncertainty on every benchmark, full-scale matched controls, the standardized external harness, the observed null result on an RL-saturated base, direct localization of the useful initialization subspace, systems cost, data and verifier audits, and robustness to optimization and decoding choices. Unless noted otherwise, all results use the fixed internal harness and the 250-step Qwen2.5-Math-7B setting from §5. Hyperparameters and prompt templates are selected on a held-out development split and then frozen; no checkpoint or prompt-template selection is performed on a test set at any RL stage.

## B.1 GENERALITY: BACKBONES, DATASETS, ALGORITHMS

Table 11 shows OFFPRIN's $4\times$ efficiency holds across two additional backbones (Qwen3-4B-Base and Llama-3.1-8B), a second RL algorithm (GRPO (Shao et al., 2024)), and a second training corpus (Open-Reasoner-Zero (Hu et al., 2026)) under otherwise matched RL settings for each comparison.

## B.2 MULTI-SEED SIGNIFICANCE AND ABLATIONS

Table 12 reports mean $\pm$ std over 5 seeds; OFFPRIN's gain over std-r32 is significant on the high-power benchmarks, and its match to std-r128 is within noise. Table 12 ablates the number of off-principal directions; Table 8 reports the frozen-$A$ variant with a fixed input projection.

## B.3 QUALITATIVE ANALYSIS

Table 13 contrasts representative rollouts of OFFPRIN-r32 and the std-r32 baseline on an AIME-2024 problem. The OFFPRIN model commits to a correct decomposition earlier and avoids the

Table 12: **Multi-seed comparison and rank ablation** (Qwen2.5-Math-7B, mean±std over 5 seeds in the left panel; pass@1 %). Bold = best; underline = second best in each benchmark or ablation.

| | std r32 | OFFPRIN r32 | std r128 |
|---|---|---|---|
| MATH-500 | 69.6±0.6 | **72.4±0.5** | <u>72.0±0.5</u> |
| OlympiadBench | 38.1±0.7 | **40.7±0.6** | <u>40.2±0.6</u> |

| off-principal dims | MATH-500 | Δ |
|---|---|---|
| r8 | 70.4 | +0.7 |
| r16 | 71.5 | +1.8 |
| r32 | **72.7** | +3.0 |
| r64 | <u>72.2</u> | <u>+2.5</u> |

Table 13: **Representative rollouts** on AIME-2024: std-r32 (incorrect) and OFFPRIN-r32 (correct).

| **std-r32 baseline (incorrect)** | OFFPRIN**-r32 (correct)** |
|---|---|
| "… Let me set $x$ as the number of … I get $x = \frac{204}{7}$, which is not an integer, so I must have miscounted the cases … [loops, runs out of budget]." | "… By symmetry the count splits into two disjoint cases. Case 1 gives $\binom{8}{3}$; Case 2 gives … Summing and reducing, $m + n = \boxed{204}$." |

algebra slip that derails the baseline, consistent with the larger early gradient norm and the off-principal hypothesis (the adapter modifies the directions that carry task-specific computation rather than overwriting the pretrained backbone itself during adaptation under matched rollout budgets).

### B.4 THE LAW EXTENDS DOWN TO RANK ONE

Because the motivating claim is specifically that rank one suffices, Table 14 extends the main sweep with $r \in \{1, 2, 4\}$. The values at base, $r8$, $r32$, and $r128$ are the corresponding entries from Table 1. On both headroom backbones, rank one captures the easiest portion of the RL gain but leaves a large, monotone capacity gap: relative to $r128$, it loses 9.9/10.1 points on MATH-500/OlympiadBench for the 7B and 11.7/6.6 points for the 1.5B. On R1-Distill, by contrast, $r1$ through $r128$ remain statistically indistinguishable. Thus saturation also gates the measured benefit even at rank one.

Table 14: **Low-rank extension of the saturation-gated law.** pass@1 (%), fixed internal harness. Existing values from Table 1 are repeated. Bold = best; underline = second best in each row.

| backbone | benchmark | base | r1 | r2 | r4 | r8 | r32 | r128 |
|---|---|---|---|---|---|---|---|---|
| Qwen2.5-Math-7B | MATH-500 | 39.2 | 60.8 | 62.9 | 65.1 | 67.3 | <u>69.6</u> | **70.7** |
| | OlympiadBench | 8.8 | 29.4 | 31.1 | 33.0 | 35.2 | <u>38.4</u> | **39.5** |
| Qwen2.5-Math-1.5B | MATH-500 | 36.3 | 54.7 | 56.9 | 59.0 | 61.2 | <u>63.6</u> | **66.4** |
| | OlympiadBench | 25.2 | 28.5 | 29.2 | 30.2 | 31.4 | <u>33.0</u> | **35.1** |
| R1-Distill-1.5B | MATH-500 | 80.4 | 81.3 | <u>81.7</u> | 81.5 | **81.9** | 81.4 | 81.4 |
| | AIME-2024 | 22.3 | 24.9 | 25.2 | 25.5 | <u>25.7</u> | 25.4 | **26.1** |

### B.5 FULL BENCHMARK RESULTS AND STATISTICAL UNCERTAINTY

Table 15 gives the complete benchmark suite with percentile 95% confidence intervals from 10,000 problem-level bootstrap resamples. The large test sets show that OFFPRIN-$r32$ improves over std-$r32$ beyond sampling noise: the paired-bootstrap differences are $+3.0$ points on MATH-500 ($p = 0.003$) and $+2.9$ on OlympiadBench ($p = 0.006$). In contrast, the differences from std-$r128$ are not significant ($p = 0.740$ and $p = 0.620$, respectively), supporting the claim that off-principal rank 32 recovers the accuracy of random-init rank 128. The smaller AMC/AIME sets have wider intervals, so their point estimates provide supportive evidence rather than the basis of the rank-scaling claim.

The five-seed result in Table 12 reaches the same conclusion without relying on problem resampling. Across seeds, OFFPRIN-$r32$ exceeds std-$r32$ in all 10 paired seed–benchmark comparisons. Its standard deviation is no larger than standard LoRA's under the same on-policy training budget.

Table 15: **Full benchmark suite with 95% bootstrap CIs.** pass@1 (%), Qwen2.5-Math-7B. The OlympiadBench result is included here because it is one of the two high-power benchmarks, although it is not part of Table 9. Bold = best; underline = second best, based on the point estimates in each row. Intervals use 10,000 problem-level bootstrap resamples under the fixed internal harness.

| benchmark ($n$) | base | std r32 | OFFPRIN **r32** | std r128 |
|---|---|---|---|---|
| GSM8K (1,319) | 64.7 [62.1, 67.3] | 88.6 [86.9, 90.3] | **89.6 [87.9, 91.2]** | 89.4 [87.7, 91.0] |
| MATH-500 (500) | 39.2 [34.9, 43.5] | 69.7 [65.6, 73.6] | **72.7 [68.4, 76.2]** | 72.2 [68.3, 76.0] |
| OlympiadBench (674) | 8.8 [ 6.7, 11.0] | 38.2 [34.6, 41.9] | **41.1 [37.2, 44.7]** | 40.5 [36.8, 44.2] |
| Minerva-Math (272) | 18.4 [14.2, 23.5] | 30.9 [25.4, 36.2] | **32.4 [27.0, 38.1]** | 31.6 [26.4, 37.3] |
| college-math (2,818) | 27.4 [26.2, 28.8] | 42.5 [41.3, 43.8] | **44.7 [43.5, 45.9]** | 44.3 [43.1, 45.6] |
| Omni-MATH (4,428) | 11.8 [10.9, 12.8] | 24.1 [22.9, 25.4] | **26.4 [25.1, 27.7]** | 25.6 [24.3, 26.9] |
| AMC-2023 (40) | 15.4 [ 9.4, 22.5] | 58.2 [50.2, 66.0] | **61.7 [52.0, 67.6]** | 61.6 [53.8, 69.2] |
| AIME-2024 (30) | 4.2 [ 1.7, 7.3] | 17.7 [11.2, 24.9] | **21.0 [14.1, 28.7]** | 19.2 [12.5, 26.7] |
| AIME-2025 (30) | 2.5 [ 0.8, 5.0] | 9.7 [ 5.0, 15.1] | **11.0 [ 5.9, 16.8]** | 10.9 [ 5.8, 16.7] |

### B.6 THE INITIALIZATION BENEFIT IS ALSO SATURATION-GATED

The rank law yields a sharper negative control than those in the main text: if the base is already RL-saturated, better placement of adapter capacity does not help. We therefore repeat the matched init comparison on DeepSeek-R1-Distill-Qwen-1.5B. As Table 16 shows, OFFPRIN-$r32$ changes each result relative to std-$r32$ by at most $0.3$ points and remains inside the paired bootstrap interval of standard LoRA. The off-principal advantage is thus gated by the same headroom variable as rank under the same training regime, evaluation settings, and matched low-rank adapter comparison.

Table 16: **Off-principal initialization is inert on an RL-saturated base.** DeepSeek-R1-Distill-Qwen-1.5B, pass@1 (%); all pairwise differences are non-significant ($p > 0.350$). Bold = best; underline = second best in each row. Rankings reflect point estimates rather than significance tests.

| benchmark | std r32 | OFFPRIN r32 | std r128 |
|---|---|---|---|
| MATH-500 | 81.4 | **81.6** | 81.4 |
| AMC-2023 | 70.5 | 70.8 | **72.0** |
| AIME-2024 | 25.4 | 25.6 | **26.1** |
| AIME-2025 | 21.5 | **21.6** | 21.5 |

### B.7 WHERE THE USEFUL RANK LIVES

We next divide the right-singular basis of every adapted weight into equal-size spectral bands and initialize $A$ from the top, middle, or bottom band. Table 17 shows a smooth ordering: moving from principal to off-principal directions increases both the first-20-step gradient norm and final accuracy. Bottom-$r$ is the only rank-32 initialization that matches std-$r128$. The result also separates OFFPRIN from merely using an orthogonal matrix: all three spectral variants have orthonormal rows. The middle band also improves over random initialization, but only the bottom band matches std-$r128$.

Table 17: **Spectral-band intervention.** Qwen2.5-Math-7B, rank 32. Gradient norm is $\|\nabla_B \mathcal{J}\|_F$ averaged over the first 20 updates and normalized to random init. Bold = largest; underline = second largest gradient norm or final accuracy in each column across all four tested read-out initializations.

| $A_0$ subspace | early gradient norm | MATH-500 | OlympiadBench |
|---|---|---|---|
| principal (top-$r$) | $0.7\times$ | 69.1 | 37.8 |
| middle spectral band | $2.3\times$ | 70.6 | 39.2 |
| random (standard LoRA) | $1.0\times$ | 69.7 | 38.2 |
| off-principal (bottom-$r$) | $5.1\times$ | **72.7** | **41.1** |

The frozen-$A$ result (71.8/40.3 in Table 8) is slightly below the trainable-$A$ variant but above random initialization. Thus exact confinement to the initial bottom-$r$ subspace is already sufficient for most

of the gain, while allowing $A$ to rotate during training recovers the remaining $0.9$ MATH-500 and $0.8$ OlympiadBench points. This is consistent with the useful update being concentrated near, rather than exactly contained in, the initial bottom singular subspace of the frozen pretrained weights.

**Direct localization of the trained update.** The intervention above establishes causality; Table 18 checks the corresponding observational prediction on standard random-init adapters. For each layer we project the final update onto the top and bottom quartiles of the pretrained weight's right-singular basis and report the fraction of update Frobenius energy. A uniformly oriented update would place 25% in either quartile. Instead, 64–70% lies in the bottom quartile on the two headroom bases, only 9–11% lies in the top quartile, and layerwise update energy is negatively correlated with the singular values of $W$. The effect occurs in both attention and MLP blocks, ruling out a single-module artifact. The weaker localization on R1-Distill is consistent with its small, noisy residual RL training signal.

Table 18: **Where a standard RLVR update lives.** Random-init rank-128 adapters; energy fractions and Spearman correlation are averaged over adapted attention and MLP layers within each group.

| backbone | module group | bottom-quartile energy | top-quartile energy | $\rho$(energy,$\sigma(W)$) |
|---|---|---|---|---|
| Qwen2.5-Math-7B | attention | 67.8% | 9.4% | $-0.59$ |
| | MLP | 70.1% | 8.8% | $-0.63$ |
| Qwen2.5-Math-1.5B | attention | 64.2% | 10.7% | $-0.55$ |
| | MLP | 66.7% | 9.9% | $-0.58$ |
| R1-Distill-1.5B | all linear | 46.5% | 17.8% | $-0.29$ |

### B.8 Rank budget and target-module ablations

Table 19 expands the single-benchmark rank ablation in Table 12. The existing MATH-500 values are copied exactly. The frontier saturates near $r32$: $r64$ adds no reliable accuracy despite doubling parameters, while $r16$ already recovers most of the gap to std-$r128$. This is compatible with the main rank law: a random read-out needs broad capacity to intersect the useful subspace, whereas the aligned read-out allocates its capacity directly to the useful off-principal RL update subspace.

Table 19: OffPrin **rank-efficiency frontier.** Qwen2.5-Math-7B, pass@1 (%); trainable counts use all-linear targets. Bold = best; underline = second best in each accuracy column for this backbone.

| method | trainable | MATH-500 | OlympiadBench |
|---|---|---|---|
| std LoRA r32 | 80.7M | 69.7 | 38.2 |
| OffPrin r8 | 20.2M | 70.4 | 38.8 |
| OffPrin r16 | 40.4M | 71.5 | 39.8 |
| OffPrin **r32** | 80.7M | **72.7** | **41.1** |
| OffPrin r64 | 161.5M | <u>72.2</u> | <u>40.8</u> |
| std LoRA r128 | 322.9M | <u>72.2</u> | 40.5 |

We also vary which linear maps receive adapters (Table 20). Both attention and MLP updates contribute: either group alone improves over its standard-initialized counterpart, but all-linear targeting is best. Applying OffPrin only to attention or only to MLP while leaving the other group at standard initialization gives intermediate results, showing that the gain spans attention and MLP maps.

### B.9 Parameter, memory, and initialization cost

Table 21 reports end-to-end systems quantities on the same 8×H100-80GB trainer. OffPrin has exactly the same trainable parameter count, checkpoint size, and steady state memory as standard rank-32 LoRA. Computing the randomized thin SVDs takes 11.8 minutes once and is cached; this is 0.18% of a 1000-step run. Because rollouts dominate RLVR wall time, throughput is indistinguishable from standard LoRA after initialization. Relative to std-$r128$, OffPrin reduces trainable and optimizer-state parameters by $4\times$, runs slightly faster, and uses 6.3 GB less peak memory per GPU.

Table 20: **Target-module ablation.** Rank 32, Qwen2.5-Math-7B, pass@1 (%). "mixed" uses OFFPRIN on the named group and standard initialization on the other group. Bold = best; underline = second best in each accuracy column. All-linear adapters include attention and MLP projections.

| initialization / targets | trainable | MATH-500 | OlympiadBench |
|---|---|---|---|
| standard, attention only | 20.2M | 68.4 | 37.1 |
| OFFPRIN, attention only | 20.2M | 70.8 | 39.4 |
| standard, MLP only | 60.6M | 67.9 | 36.8 |
| OFFPRIN, MLP only | 60.6M | 70.2 | 38.9 |
| mixed, OFFPRIN attention | 80.7M | <u>71.5</u> | <u>40.0</u> |
| mixed, OFFPRIN MLP | 80.7M | 71.1 | 39.7 |
| OFFPRIN**, all linear** | 80.7M | **72.7** | **41.1** |

Table 21: **Efficiency audit** for Qwen2.5-Math-7B. Throughput is normalized to std-r32; checkpoint size stores adapter weights in BF16. Bold = best; underline = second best efficiency in each column.

| method | trainable | adapter ckpt. | peak HBM/GPU | rel. throughput |
|---|---|---|---|---|
| full fine-tuning | 7.60B (100.00%) | 15.2 GB | 76.8 GB | 0.73× |
| std LoRA r32 | <u>80.7M (1.06%)</u> | <u>161 MB</u> | <u>52.6 GB</u> | 1.00× |
| OFFPRIN r32 | <u>80.7M (1.06%)</u> | <u>161 MB</u> | <u>52.6 GB</u> | <u>1.01×</u> |
| OFFPRIN r32, frozen-$A$ | **44.5M (0.59%)** | **89 MB** | **51.1 GB** | 1.02× |
| std LoRA r128 | 322.9M (4.25%) | 646 MB | 58.9 GB | 0.94× |

### B.10 OPTIMIZATION ROBUSTNESS AND ERROR AUDIT

We perturb LoRA scaling, learning rate, rollout-group size, and clipping around the main configuration. The advantage is stable over a $4\times$ range of LoRA scaling and learning rate, for $G \in \{4, 8, 16\}$, and under symmetric clipping (Table 22); in every configuration OFFPRIN-$r32$ remains ahead of the corresponding std-$r32$ run. The improvement therefore persists across all tested training settings.

Table 22: **Optimization robustness.** Each row changes one setting from the main configuration ($\alpha/r = 2$, learning-rate multiplier $1\times$); pass@1 (%). Bold = best within each matched standard/OFFPRIN pair. All remaining hyperparameters are held fixed in each matched comparison.

| factor | value | std r32 MATH | OFFPRIN r32 MATH | std r32 Olympiad | OFFPRIN r32 Olympiad |
|---|---|---|---|---|---|
| LoRA scale | $\alpha/r = 1$ | 69.3 | **72.0** | 37.9 | **40.4** |
| LoRA scale | $\alpha/r = 2$ | 69.7 | **72.7** | 38.2 | **41.1** |
| LoRA scale | $\alpha/r = 4$ | 69.4 | **72.1** | 38.0 | **40.5** |
| learning rate | $0.5\times$ | 69.1 | **71.8** | 37.6 | **40.1** |
| learning rate | $2\times$ | 69.0 | **71.7** | 37.5 | **40.0** |
| group size | $G = 4$ | 68.8 | **71.4** | 37.2 | **39.8** |
| group size | $G = 16$ | 69.9 | **72.5** | 38.5 | **41.0** |
| clip bounds | symmetric $(0.2, 0.2)$ | 69.2 | **71.9** | 37.7 | **40.2** |

### B.11 DATA SCALE AND EVALUATION ROBUSTNESS

The benefit is not confined to the 40,315-prompt pool. We train on fixed random subsets of DeepScaleR-Preview while holding the number of optimizer updates fixed. As Table 23 shows, more unique prompts improve both methods, and OFFPRIN retains a 2.5–3.0 point MATH-500 advantage at every scale. This argues against a fortunate interaction with a single training subset.

Table 24 varies decoding temperature and the generation limit. The exact main configuration is reproduced in its second row ($T = 0.6$, 3072 tokens). OFFPRIN's gain survives greedy decoding, hotter sampling, and a doubled output budget. For the 6144-token condition, both methods use an expanded runtime context window that accommodates each prompt plus up to 6144 generated tokens, with the model ranking unchanged at the larger output budget for both matched models.

Table 23: **Training-data scaling.** Qwen2.5-Math-7B, rank 32, 250 updates, pass@1 (%). Bold = best within each matched pair. Both methods share the same fixed prompt subset at each data scale.

| training prompts | std MATH | OFFPRIN MATH | std Olympiad | OFFPRIN Olympiad |
|---|---|---|---|---|
| 5k | 66.1 | **68.6** | 35.0 | **37.1** |
| 10k | 67.5 | **70.1** | 36.3 | **38.6** |
| 20k | 68.6 | **71.4** | 37.4 | **39.8** |
| 40,315 | 69.7 | **72.7** | 38.2 | **41.1** |

Table 24: **Decoding robustness.** Qwen2.5-Math-7B, rank 32, pass@1 (%). Other decoding settings, including top-$p = 0.95$ for sampled decoding, are fixed. Bold = best within each matched RL pair.

| temperature | max new tokens | std MATH | OFFPRIN MATH | std Olympiad | OFFPRIN Olympiad |
|---|---|---|---|---|---|
| 0.0 | 3072 | 70.2 | **72.6** | 38.7 | **40.9** |
| 0.6 | 3072 | 69.7 | **72.7** | 38.2 | **41.1** |
| 0.8 | 3072 | 69.2 | **72.0** | 37.9 | **40.4** |
| 0.6 | 6144 | 69.9 | **72.5** | 38.4 | **41.1** |

### B.12 TRAINING-DATA AND VERIFIER AUDIT

We normalize Unicode and mathematical markup, remove answer fields, and compare every evaluation question against all 40,315 training prompts using exact matching and character 5-gram MinHash (Broder, 1997). Table 25 reports candidates above Jaccard 0.80 and the number confirmed as the same mathematical problem by two annotators. No exact or semantic duplicate is found. As a conservative check, we remove all 41 near-match candidates and retrain the matched triplet; the resulting MATH-500 scores are 69.5/72.2/72.0 for std-$r32$/OFFPRIN-$r32$/std-$r128$, and the OlympiadBench scores are 38.0/40.7/40.3. All shifts in accuracy remain within 0.5 points.

Table 25: **Train–test overlap audit.** Counts compare DeepScaleR-Preview with each test set after answer removal. "confirmed" denotes the same underlying problem, not merely shared mathematical vocabulary. Candidates have Jaccard similarity above 0.80 and are checked by two annotators.

| test set | exact matches | near-match candidates | confirmed duplicates |
|---|---|---|---|
| GSM8K | 0 | 3 | 0 |
| MATH-500 | 0 | 11 | 0 |
| OlympiadBench | 0 | 4 | 0 |
| Minerva-Math / college-math | 0 | 7 | 0 |
| Omni-MATH | 0 | 16 | 0 |
| AMC-2023 / AIME-2024 / AIME-2025 | 0 | 0 | 0 |

We additionally audit the binary verifier on 500 stratified training rollouts. Independent manual adjudication finds three false positives (0.6%) and five false negatives (1.0%); none differs by adapter initialization. Re-evaluating all test outputs with a stricter ensemble that requires agreement between symbolic simplification and high-precision numerical checking changes MATH-500 by $-0.1/-0.1$ and OlympiadBench by $-0.2/-0.1$ points for std-$r32$/OFFPRIN-$r32$, respectively. The advantage therefore persists with the stricter verifier on both benchmarks for the same evaluated RL models.

For a qualitative error audit, two annotators independently categorized 200 fixed MATH-500 prompts under greedy decoding (agreement $\kappa = 0.81$). OFFPRIN solves nine additional problems, with the largest reductions in planning/decomposition errors (22→18) and length-limit truncations (12→9); algebra errors fall from 18 to 16 and formatting errors remain unchanged (4 each). This pattern agrees with the rollout in Table 13 and argues against the gain coming from superficial answer-format changes or verifier-specific artifacts under matched decoding. These audits assess the reported training–evaluation comparisons; they do not establish the absence of contamination in the pretrained models or guarantee verifier correctness on unseen tasks and datasets. Together, the checks test whether the measured advantage survives changes to data overlap, decoding, and verification; broader claims about unseen tasks require further independent task-specific evaluation.