SEED VARIANCE AT THE STABILITY BOUNDARY CONFOUNDS CROSS-SCALE COMPARISONS IN RL POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning rates tuned on small models are often reused at larger scales in reinforcement-learning post-training. We ask whether the learning-rate optimum and the onset of instability transfer in the same way. Using GRPO on Qwen3-Base models at 0.6B, 1.7B and 4B parameters, we sweep learning rate and replicate configurations near and away from the transition to degradation. Across these scales, the 0.6B optimum remains near-optimal at 4B: the loss in final accuracy is 0.008 for the matched seed and 0.013 when the transferred setting is averaged over three seeds. The stability boundary is much less reproducible. Two grid steps below it, seed-to-seed variation is comparable to evaluation noise; at 10^{-5}, the standard deviation rises from 0.007–0.010 to 0.116–0.160, a 15–20-fold increase relative to the reference seed variation. This variance coincides with a qualitative failure mode: runs abruptly lose the ability to terminate outputs, but the onset occurs at different training steps for different seeds. Consequently, a one-seed sweep produces a non-monotone scale trend that disappears under replication. On our grid, the practical distinction is sharp: a single run is enough to select a near-optimal learning rate, whereas measuring the stability boundary requires substantially more replication.

1 INTRODUCTION

Learning rates for reinforcement-learning post-training are expensive to tune at large scale. A common workaround is to tune on a smaller model and reuse the result, while scaling studies often hold the learning rate and other training hyperparameters fixed across model sizes (Tan et al., 2025; Hu et al., 2025). Hyperparameter optimisation methods for RL post-training also use small proxy models for the same reason (Chen et al., 2026). In pretraining, such transfer can be justified by parametrization: under maximal update parametrization, the optimal learning rate can become approximately width-invariant (Yang et al., 2022; Hayou, 2025). RL post-training has no comparable guarantee.

There are two different transfer questions. The first is whether a learning rate that performs well on a small model remains near-optimal at larger scales. The second is whether the point at which training becomes unstable moves with scale. The literature is mixed on the latter. Ouyang et al. (2022) report the same PPO divergence threshold at 1.3B and 6B parameters, and Song et al. (2026) find similar staleness–learning-rate trends at 1B and 3B. Other studies report greater instability at larger scale (Prime Intellect Team et al., 2025), or a non-monotone pattern across model sizes (Çağatan and Akgün, 2025). The latter study notes that its single-seed runs cannot cleanly separate a scale effect from run-to-run variation. We study that ambiguity directly.

We sweep the learning rate for GRPO on Qwen3-Base models at 0.6B, 1.7B and 4B parameters, then replicate settings at the stability boundary and two grid steps below it. Three results are central:

- **Near the optimum, transfer is forgiving.** The 0.6B optimum costs only 0.008 in final accuracy when applied to the matched 4B run, and 0.013 when that transferred setting is averaged over three seeds (§4.1).

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.

- **At the boundary, variance spikes because termination fails at seed-dependent times.** Seed standard deviations increase from 0.007–0.010 in the controls to 0.116–0.160 at the boundary, while individual runs abruptly begin hitting the completion-length limit (§4.2–§4.3).
- **One-seed boundary sweeps can misstate the scale trend.** The non-monotone ordering seen in the original sweep disappears under replication, and a simple power calculation shows why boundary comparisons need many more runs than near-optimum selection (§4.4, §5.1).

The main point is not that learning-rate transfer fails. In our experiments it works well for choosing a high-performing value. What does not transfer cleanly from a one-seed sweep is the apparent location of the stability boundary, because that is precisely where run-to-run variability becomes large.

2 Related Work

Hyperparameter transfer across scale. Maximal update parametrization was introduced to make hyperparameters such as the learning rate transferable across width (Yang et al., 2022); recent work gives a theoretical account of this effect (Hayou, 2025) and connects it to size-stable landscape geometry (Noci et al., 2024). In standard parametrization, however, learning-rate stability can change with model size. Pretraining studies have shown that small models can reproduce large-model instabilities (Wortsman et al., 2023), that maximal stable learning rates depend on depth and width (Iyer et al., 2023), and that hyperparameter sensitivity itself can change with scale (Lourie et al., 2026). Related work in value-based RL also finds scale-dependent interactions among model size, batch size and optimization (Fu et al., 2025). We focus on the post-training setting, where standard parametrization is common and the transfer properties are less understood.

Stability and variance in RL post-training. Recent RLVR studies identify several routes to instability, including step-size thresholds (Suk and Duan, 2025), stale-policy dynamics (Song et al., 2026; Ghosh et al., 2026), learning-rate-gated collapse (Gan et al., 2026), clipping effects (Liu et al., 2025), and other forms of policy collapse (Deng et al., 2025; He et al., 2026). Scale can also change the ranking of post-training algorithms (Li, 2026). Separately, deep-RL methodology has long shown that conclusions can be sensitive to random seeds (Henderson et al., 2018; Colas et al., 2018; Agarwal et al., 2021) and to other sources of experimental variation (Bouthillier et al., 2021). Our result links these two literatures: the seed variance of an RL post-training recipe is not constant across learning rates, but rises sharply near the point where training becomes unstable.

3 Experimental Setup

Training. We fine-tune Qwen3-Base models (Qwen Team, 2025) at 0.6B, 1.7B and 4B parameters using GRPO (Shao et al., 2024) in TRL (von Werra et al., 2020), with rollouts generated by vLLM (Kwon et al., 2023). Each run uses 2,000 GSM8K training prompts (Cobbe et al., 2021) and 150 optimizer steps with 8-bit Adam in bfloat16. The reference recipe has KL coefficient $\beta{=}0$, clipping range $\epsilon{=}0.2$, group size 8, rollout temperature 1.0, and a maximum completion length of 1,024 tokens. The effective batch size is fixed at 32 sequences per optimizer step for all three models by changing the per-device batch size and gradient accumulation. We sweep $\{3{\times}10^{-7}, 10^{-6}, 3{\times}10^{-6}, 5{\times}10^{-6}, 10^{-5}, 3{\times}10^{-5}\}$ and vary the remaining hyperparameters one at a time (Appendix A). All models use standard parametrization.

Evaluation. We evaluate 200 held-out GSM8K problems before training and every 50 steps using the same code path. At each checkpoint we sample eight completions per problem at temperature 1.0, for 1,600 completions in total, with the same 1,024-token limit used during training. Accuracy is the mean symbolic-answer correctness over those completions. We also record the fraction of completions that reach the length limit. Re-evaluating the same untrained checkpoint across runs gives standard deviations of 0.007, 0.003 and 0.004 for 0.6B, 1.7B and 4B, respectively; these values provide a direct estimate of evaluation-sampling noise.

Table 1: Final accuracy across the learning-rate sweep (seed 0). "Pre" is the model before training. Bold marks the best setting for each model.

		learning rate					
model	pre	3×10^{-7}	10^{-6}	3×10^{-6}	5×10^{-6}	10^{-5}	3×10^{-5}
0.6B	0.074	0.383	0.620	**0.671**	0.605	0.000	0.000
1.7B	0.463	0.671	**0.813**	0.769	0.794	0.537	0.001
4B	0.423	0.879	**0.931**	0.923	0.814	0.085	0.000

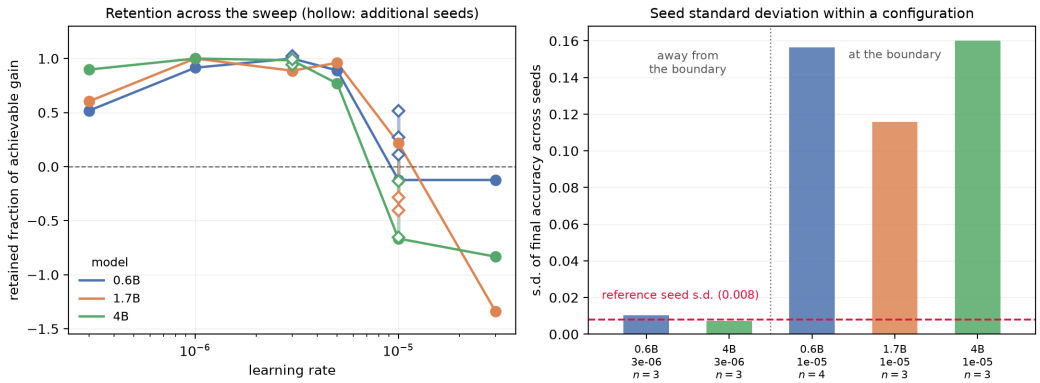


Figure 1: Left: retention (Eq. 1) across the learning-rate sweep. Filled markers are seed 0; hollow markers are additional seeds of the same configuration. Right: standard deviation of final accuracy across seeds for the five replicated configurations. Configurations two grid steps below the boundary match the reference seed standard deviation (dashed line); configurations at the boundary exceed it by a factor of 15 to 20.

Retention. Because the three models start from different accuracies, raw accuracy changes are not comparable across scales. We therefore report, for each run, the fraction of its model's best achievable gain that the run retains:

$$\text{retention}(\eta) \;=\; \frac{\text{acc}_{\text{post}}(\eta) - \text{acc}_{\text{pre}}}{\text{acc}_{\text{post}}(\eta^\star) - \text{acc}_{\text{pre}}}, \tag{1}$$

where $\eta^\star$ is the best learning rate for that model. Retention equals one at the optimum, zero when training leaves accuracy unchanged, and is negative when training degrades the model.

Replication. The initial grid uses one training seed per setting. We then add seeds to five configurations: 10^{-5} at all three scales (three to four seeds each) and 3×10^{-6} at 0.6B and 4B (three seeds each), the latter serving as controls two grid steps below the boundary. Separately, ten-seed runs of the reference recipe at 0.6B and 1.7B give a final-accuracy standard deviation of 0.008 at both scales. We use this value as a reference for ordinary run-to-run variation away from the boundary.

4 Results

4.1 The optimum transfers across the three scales

Table 1 shows the seed-0 sweep. The best learning rate is 3×10^{-6} at 0.6B and 10^{-6} at 1.7B and 4B, but the peak is broad. Applying the 0.6B optimum to 4B gives 0.923 for seed 0 and 0.918 averaged over three seeds, versus 0.931 at the 4B optimum. The corresponding loss is only 0.008–0.013. Across the interval from 10^{-6} to 3×10^{-6}, all three models retain at least 88% of their best observed gain (Figure 1, left). Thus, within the resolution of this grid, the near-optimal region transfers well enough that separate tuning at every scale is unnecessary.

Table 2: Final accuracy of replicated configurations. The control configurations lie two grid steps below the boundary. The last column is the within-configuration seed standard deviation divided by the reference value of 0.008.

model	learning rate	seeds	final accuracy	s.d.	ratio
Control					
0.6B	3×10^{-6}	3	0.656, 0.671, 0.676	0.010	1.3
4B	3×10^{-6}	3	0.910, 0.922, 0.923	0.007	0.9
Boundary					
0.6B	10^{-5}	4	0.000, 0.131, 0.219, 0.372	0.156	19.6
1.7B	10^{-5}	3	0.319, 0.361, 0.537	0.116	14.5
4B	10^{-5}	3	0.085, 0.086, 0.362	0.160	20.0

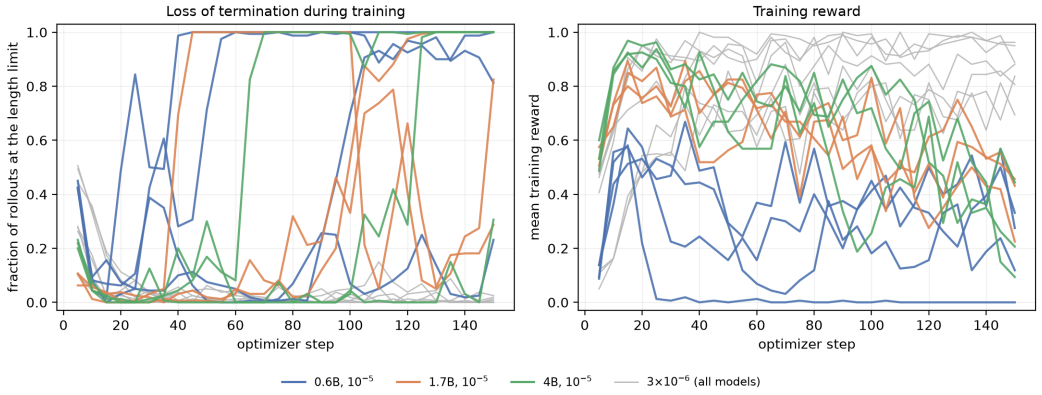


Figure 2: Training trajectories of the replicated configurations. Left: fraction of training rollouts that reach the 1,024-token limit. Right: mean training reward. Colored lines are individual seeds at 10^{-5}; gray lines are the seven runs at 3×10^{-6} across all three models.

4.2 Variance spikes at the stability boundary

All three models degrade between 5×10^{-6} and 3×10^{-5}, and at 3×10^{-5} every model collapses to near-zero accuracy. The boundary therefore lies in the same grid interval at every scale. What changes at the boundary is the reproducibility of individual runs.

Table 2 shows the replicated settings. At 3×10^{-6}, the seed standard deviation is 0.010 at 0.6B and 0.007 at 4B, close to the 0.008 reference value and to the 0.003–0.007 evaluation noise measured on identical checkpoints. At 10^{-5}, the standard deviation rises to 0.116–0.160, or 15–20 times the reference seed variation.

The separation is visible without a distributional model. After centering each configuration on its own mean, the ten absolute residuals from the boundary runs are all larger than the six control residuals. If these residuals are exchangeable between the two groups, this complete ordering has probability $1/\binom{16}{6}=1.25\times10^{-4}$. The variance increase therefore occurs over a single factor-of-three step in learning rate and appears at all three model sizes.

4.3 Boundary runs lose output termination

Figure 2 reveals a common failure mode behind the spread in final accuracy. In every control run, the fraction of rollouts that hit the length limit falls quickly and then remains at or below 0.15. At 10^{-5}, eight of ten runs instead lose termination during training. The fraction rises above 0.5 at seed-dependent onsets between steps 25 and 125 and, in most cases, exceeds 0.9 within another 20 steps. At the final evaluation, seven of the ten boundary runs have at least 97% of completions at the limit; every control run is at or below 9%.

Table 3: Runs per setting needed to resolve a difference Δ in mean final accuracy (α=0.05, power 0.8).

	Δ=0.05	Δ=0.10	Δ=0.20
away from the boundary (σ=0.008)	1	1	1
at the boundary (σ=0.147)	136	34	9

Losing termination does not force accuracy to zero: a completion can contain the correct answer before reaching the limit. The 1.7B seed-0 run, for example, has 99.8% of final completions at the limit but still scores 0.537; a 4B run loses termination at step 65 and ends at 0.362. Another 1.7B run loses termination near step 40 and largely recovers by step 150. Final accuracy therefore depends on both the onset of the failure and how much useful answering behaviour remains. That produces a continuum of outcomes rather than a clean split between "healthy" and "collapsed" runs: the replicated final accuracies span 0.000–0.372 at 0.6B and 0.319–0.537 at 1.7B.

4.4 One-seed sweeps can suggest a scale trend that does not replicate

At 10^{-5}, the seed-0 sweep gives final accuracies of 0.000, 0.537 and 0.085 for 0.6B, 1.7B and 4B. Taken by itself, that column suggests a non-monotone scale effect in which 1.7B is markedly more stable than either neighbour. The same 1.7B run also finishes above its pre-training accuracy, even though 99.8% of its final completions hit the length limit (§4.3). Replication changes the picture. The two additional 1.7B seeds end at 0.319 and 0.361, both below pre-training accuracy, and the 1.7B retention range $[-40.6\%, +21.7\%]$ overlaps the 0.6B range $[-12.4\%, +51.7\%]$. The non-monotone ordering in the original sweep is therefore not stable to changing the seed.

This matters for cross-scale comparisons based on one run per setting. With a per-run standard deviation of roughly 0.12–0.16 at the boundary, an apparent shift—or lack of a shift—can reflect which seed happened to be run rather than a reproducible difference between model sizes.

4.5 What remains under replication

Replication does leave a directional difference between the endpoints. At 10^{-5}, retention spans $[-12.4\%, +51.7\%]$ at 0.6B, $[-40.6\%, +21.7\%]$ at 1.7B and $[-66.5\%, -13.5\%]$ at 4B. The 0.6B and 4B ranges do not overlap, and each of the 4×3 pairwise comparisons places a 0.6B run above a 4B run. With four versus three observations, however, the exact two-sided Mann–Whitney test is p=0.057, the smallest attainable value for complete separation at these sample sizes. We therefore treat this as directional evidence rather than a resolved scale effect. Neither adjacent pair separates.

The same ordering is visible in the single runs at 5×10^{-6}: retention is 89% at 0.6B, 96% at 1.7B and 77% at 4B. At 4B, that run loses termination at step 95 and ends with every evaluation completion at the length limit; the 0.6B and 1.7B runs end at 2% or less. These observations are consistent with a smaller stability margin at 4B than at 0.6B. The present number of seeds does not resolve the adjacent-scale differences.

5 Analysis

5.1 Replication cost near the boundary

The practical cost of a sweep depends on the seed variance at the learning rate being tested. For planning purposes, a two-sided normal approximation with α=0.05 and 80% power gives $n \approx 2(z_{0.975} + z_{0.8})^2\sigma^2/\Delta^2$ runs per setting to resolve a mean difference Δ. Table 3 evaluates this expression at the pooled boundary standard deviation σ=0.147 and at the reference value σ=0.008.

The contrast is large. A difference of 0.20 in mean final accuracy requires about nine runs per setting at the boundary under this approximation, while the same calculation rounds to one run at the reference variance. With one run per boundary setting, the difference between two observations has standard deviation $\sigma\sqrt{2} \approx 0.21$. This is why coarse one-seed sweeps are adequate for finding the broad optimum in Figure 1, but not for estimating the edge of the stable region.

Table 4: Flagging a run when its accuracy at step s is below its pre-training accuracy. Twenty-nine runs; nine end below their pre-training accuracy.

step s	degraded runs flagged	degraded runs missed	healthy runs flagged
50	4	5	1
100	6	3	1
150	9	0	0

Table 5: Initial accuracy (averaged over runs) and mean completion length at the start of training.

model	initial accuracy	mean completion length (tokens)
0.6B	0.074	633
1.7B	0.464	347
4B	0.423	436

5.2 Early checks do not reliably catch degradation

One possible shortcut is to run each scale briefly and stop settings whose accuracy falls below the pre-training value. We test this rule on the 29 runs with complete evaluation trajectories; nine end below their pre-training accuracy.

At step 50, this rule misses five of the nine runs that eventually degrade (Table 4). Some look strong for much longer: a 4B run at 10^{-5} reaches 0.812 at step 100 and ends at 0.362, while a 1.7B run reaches 0.749 at step 50 and ends at 0.361. The reverse error also occurs: the 0.6B run at $3{\times}10^{-7}$ is below its pre-training accuracy at step 50 but finishes at 0.383. A length-based canary is only slightly better. At step 50, flagging runs for which more than half of training rollouts hit the length limit identifies five of the nine eventually degraded runs and one of the twenty other runs. For these simple checks, a short run is therefore not a reliable substitute for observing the full training trajectory.

5.3 Initial accuracy and response length

Suk and Duan (2025) derive a step-size threshold that depends on policy success rate and response length. Table 5 reports both quantities at initialization. The 0.6B model, which has the largest retention at 10^{-5} under replication, also has the lowest initial accuracy and the longest responses; 1.7B and 4B are closer on both quantities. These two observables therefore do not by themselves explain the cross-model ordering in our experiment.

5.4 Architecture of the model family

The Qwen3-Base series is not a pure width-scaling ladder (Table 6). From 0.6B to 1.7B, hidden size doubles while depth and the number of query heads stay fixed. From 1.7B to 4B, hidden size, depth and head count all increase. Pretraining experiments have found different learning-rate-sensitivity trends for width and depth (Wortsman et al., 2023). We therefore interpret the 0.6B–4B contrast in §4.5 as a comparison of these model configurations, not as a clean one-dimensional law in parameter count.

6 Implications for Fixed-Hyperparameter Scaling Studies

Tan et al. (2025) study RL post-training from 0.5B to 72B parameters with the learning rate fixed at 10^{-6}. On our grid, that value is optimal at 1.7B and 4B and only 0.051 below the best final accuracy at 0.6B. This is the benign regime suggested by our results: once a learning rate lies in the broad near-optimal region, keeping it fixed across scale introduces little error.

The boundary is different. At 10^{-5}, which the original 1.7B run would have made look acceptable by final accuracy alone, the replicated outcomes across our three scales range from 0.000 to 0.537, and most runs lose output termination. A scaling curve through one run per model would therefore

Table 6: Architecture of the three models.

model	hidden size	layers	query heads	FFN size
0.6B	1024	28	16	3072
1.7B	2048	28	16	6144
4B	2560	36	32	9728

mix model-size effects with unusually large seed variation. For fixed-hyperparameter studies, the useful check is not to re-tune every scale, but to verify that the chosen learning rate sits comfortably inside the stable region rather than near the first sign of degradation.

7 Conclusion

Across three Qwen3-Base scales on GSM8K, learning-rate selection is forgiving near the optimum and fragile near the stability boundary. The 0.6B optimum remains within 0.013 of the 4B optimum in final accuracy, while the seed standard deviation near the boundary rises to 0.116–0.160. The high variance is accompanied by a concrete behavioral failure: runs lose output termination at different points in training. This makes a one-seed sweep a poor instrument for locating the boundary even when it is adequate for selecting a high-performing learning rate.

Our experiments cover one model family, one task, GRPO, 150 training steps and a 1,024-token completion limit. Within that setting, the main empirical contrast is large and consistent across scales, while the remaining 0.6B–4B difference at the boundary is directional rather than statistically resolved. Thus, cross-scale stability claims need replication at the boundary; the one-seed sweep that is sufficient to find the broad optimum is not sufficient to locate that boundary.

Reproducibility Statement

All hyperparameters are listed in §3 and Appendix A. The supplementary material contains the run configurations, raw records, software and accelerator metadata, and the analysis scripts used for the tables and figures.

AI Use Statement

GPT was used only for writing and language polishing.

References

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

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, et al. Accounting for variance in machine learning benchmarks. *arXiv preprint arXiv:2103.03098*, 2021.

Ömer Veysel Çağatan and Barış Akgün. Failure modes of maximum entropy rlhf. *arXiv preprint arXiv:2509.20265*, 2025.

Minping Chen, Bowen Xiao, Du Liang, Chuxuan Zeng, and Zeyi Wen. Efficient hyperparameter optimization for llm reinforcement learning. *arXiv preprint arXiv:2606.03073*, 2026.

Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv:1806.08295*, 2018.

Wenlong Deng, Yushu Li, Boying Gong, Yi Ren, Christos Thrampoulidis, and Xiaoxiao Li. On group relative policy optimization collapse in agent search: The lazy likelihood-displacement. *arXiv preprint arXiv:2512.04220*, 2025.

Preston Fu, Oleh Rybkin, Zhiyuan Zhou, Michal Nauman, Pieter Abbeel, Sergey Levine, and Aviral Kumar. Compute-optimal scaling for value-based deep rl. *arXiv preprint arXiv:2508.14881*, 2025.

Chengguang Gan, Zhixi Cai, Yunhao Liang, Hanjun Wei, Shiwen Ni, and Tatsunori Mori. A learning-rate-gated failure of grpo in a small language and vision-language model web agent: A controlled null and its mechanism. *arXiv preprint arXiv:2607.12640*, 2026.

Rajat Ghosh, Datta Nimmaturi, Aryan Singhal, Vaishnavi Bhargava, Henry Wong, and Johnu George. Predictable grpo: A closed-form model of training dynamics. *arXiv preprint arXiv:2606.30789*, 2026.

Soufiane Hayou. A proof of learning rate transfer under μp. *arXiv preprint arXiv:2511.01734*, 2025.

Bingxiang He, Yuxin Zuo, Zeyuan Liu, Shangziqi Zhao, Zixuan Fu, and Junlin Yang. How far can unsupervised rlvr scale llm training? *arXiv preprint arXiv:2603.08660*, 2026.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI*, 2018.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, et al. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Gaurav Iyer, Boris Hanin, and David Rolnick. Maximal initial learning rates in deep relu networks. In *ICML*, 2023.

Woosuk Kwon et al. Efficient memory management for large language model serving with pagedattention. In *SOSP*, 2023.

Xiaoyi Li. Do post-training algorithms actually differ? a controlled study across model scales uncovers scale-dependent ranking inversions. *arXiv preprint arXiv:2603.19335*, 2026.

Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, et al. Part i: Tricks or traps? a deep dive into rl for llm reasoning. *arXiv preprint arXiv:2508.08221*, 2025.

Nicholas Lourie, Kyunghyun Cho, Karen Ullrich, and Sanae Lotfi. Small-scale experiments: Are we there yet? *arXiv preprint arXiv:2608.11859*, 2026.

Lorenzo Noci, Alexandru Meterez, Thomas Hofmann, and Antonio Orvieto. Super consistency of neural network landscapes and learning rate transfer. *NeurIPS*, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Prime Intellect Team, Sami Jaghouar, Justus Mattern, Jack Min Ong, and Jannik Straube. Intellect-2: A reasoning model trained through globally decentralized reinforcement learning. *arXiv preprint arXiv:2505.07291*, 2025.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jingwei Song, Haofeng Xu, Jie Xiao, Chengke Bao, Jingwei Shi, and Pengbin Feng. Staleness-learning rate scaling laws for asynchronous rlhf. *arXiv preprint arXiv:2607.01083*, 2026.

Joe Suk and Yaqi Duan. On the optimization dynamics of rlvr: Gradient gap and step size thresholds. *arXiv preprint arXiv:2510.08539*, 2025.

Zelin Tan, Hejia Geng, Xiaohang Yu, Mulei Zhang, Guancheng Wan, Yifan Zhou, et al. Scaling behaviors of llm reinforcement learning post-training: An empirical study in mathematical reasoning. *arXiv preprint arXiv:2509.25300*, 2025.

Leandro von Werra et al. Trl: Transformer reinforcement learning. *GitHub repository, https://github.com/huggingface/trl*, 2020.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023.

Greg Yang, Edward J. Hu, et al. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

A Other Hyperparameters

Table 7 varies the KL coefficient, clipping range, group size and sampling temperature one at a time around the reference recipe. The effects are small compared with the learning-rate sweep. The KL coefficient spans 0.026 at 0.6B and 0.052 at 1.7B, while clipping range and group size stay within two reference seed standard deviations. In this TRL setup, one temperature parameter controls both rollout and evaluation sampling, so changing temperature also changes the measured pre-training accuracy.

Table 7: Final accuracy when varying one hyperparameter from the reference configuration. For temperature, pre-training accuracy also changes and is given in the last column.

parameter	model	setting			note
		0.001	0.01	0.04	
KL β	0.6B	0.621	0.632	0.606	
	1.7B	0.809	0.784	0.757	
		0.1	0.4		
clip ϵ	0.6B	0.628	0.640		
	1.7B	0.816	0.804		
		4	16		
group size	0.6B	0.641	0.616		
	1.7B	0.811	0.804		
		0.7	1.2		pre-training
temperature	0.6B	0.661	0.611		0.099 / 0.026
	1.7B	0.827	0.804		0.727 / 0.081

B Retention for All Runs

Table 8: Retention (Eq. 1) for every run. Multiple entries in a cell are separate seeds.

model	3×10^{-7}	10^{-6}	3×10^{-6}	5×10^{-6}	10^{-5}	3×10^{-5}
0.6B	51.7%	91.5%	100, 100, 103%	89.0%	-12, 11, 27, 52%	-12.4%
1.7B	60.3%	100%	88.8%	96.0%	-41, -28, 22%	-134%
4B	89.8%	100%	94, 98, 99%	76.9%	-67, -66, -14%	-83.3%

C Evaluation Details

Completion length. With limits of 512–768 tokens, 53–60% of base-model completions hit the cap, motivating the 1,024-token budget used in all reported runs. Before training, the corresponding rates are 47%, 8% and 28% for 0.6B, 1.7B and 4B. By step 150, runs at 10^{-6} to 3×10^{-6} are at 9% or less. Boundary runs that degrade end between 16.5% and 100% (§4.3). The 1,024-token limit is shared by training rollouts and evaluation.

Reward logging. TRL's evaluation call returns the loss by default. We record the evaluation reward through the logged training history with an evaluation at step 0, which gives the pre-training accuracy through the same code path as the post-training accuracy.

Statistical comparisons. For the boundary-versus-control comparison in §4.2, we center each replicated configuration on its own mean and compare absolute residuals. All ten boundary residuals exceed all six control residuals. Under exchangeability of these residuals between the two groups, the probability of this complete ordering is $1/\binom{16}{6}=1.25\times10^{-4}$. The 0.6B–4B comparison in §4.5 uses the exact Mann–Whitney test on four and three retention values. Complete separation gives the smallest possible two-sided value at these sample sizes, $2/35=0.057$.