SAME ADAPTER FUNCTION, DIFFERENT GRPO DYNAMICS: FACTOR-SCALE CONFOUNDING IN LORA OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-Rank Adaptation (LoRA) factorizes each weight update into two trainable matrices, so the same adapter function admits multiple factor scales. Because standard optimizers update these factors separately, functionally identical initializations can receive different updates, confounding the attribution of gains from non-zero LoRA initializations. We study this ambiguity using function-preserving scalar gauges, an identical frozen-rollout control, and reciprocal factor learning rates. Under online Group Relative Policy Optimization (GRPO), shared-rate AdamW amplifies early adapter-product displacement by up to $6.49\times$ and separates otherwise identical trajectories, while reciprocal rates substantially reduce this sensitivity across Qwen3 models up to 8B. A teacher-forced SFT control reproduces the local coordinate effect without rollout, reward, or advantage estimation. Thus, factor scale is an optimizer-level confound rather than a GRPO-specific phenomenon, while online policy feedback makes it consequential for early RL trajectories.

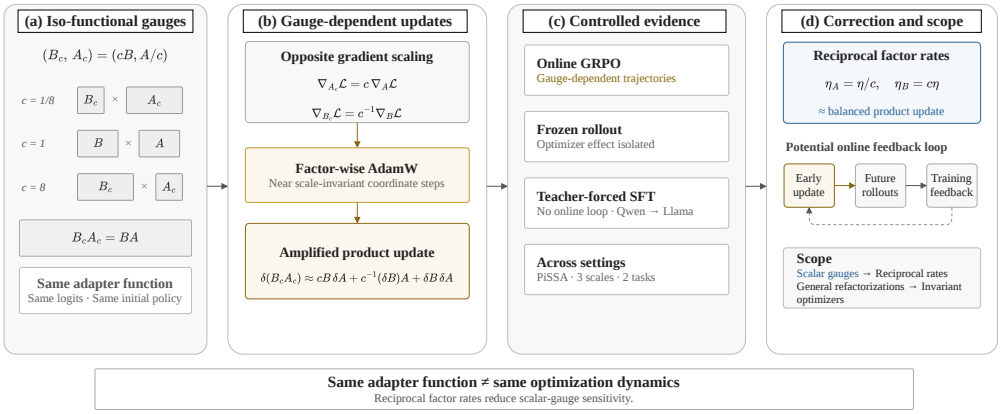

Figure 1: Overview of factor-scale confounding in LoRA optimization. (a) Iso-functional scalar gauges preserve the adapter product, logits, and initial policy while changing the factor coordinates. (b) The resulting opposite gradient scaling interacts with factor-wise AdamW to produce gauge-dependent product updates. (c) Online GRPO demonstrates divergent trajectories, frozen-rollout GRPO isolates the immediate optimizer-coordinate effect, and teacher-forced SFT shows that the effect does not require rollout or reward feedback. Evidence includes a PiSSA initialization, multiple Qwen3 model scales and tasks, and fixed-data and 200-update online replications on Llama-3.2. (d) Reciprocal factor rates reduce scalar-gauge sensitivity but do not provide exact practical invariance; general refactorizations require invariant optimizers. Under online GRPO, early coordinate-dependent updates may propagate through subsequent rollouts and training feedback.

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.

1 INTRODUCTION

Online reinforcement learning with verifiable rewards (RLVR) can improve language-model reasoning without annotated reasoning traces. Group Relative Policy Optimization (GRPO) estimates advantages from groups of sampled responses and is widely used in this setting (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025). Because online RL repeatedly samples from and updates a large policy, Low-Rank Adaptation (LoRA) is an attractive way to reduce trainable parameters and optimizer memory (Hu et al., 2022; Wang et al., 2026).

LoRA's efficiency introduces an attribution problem. The same update BA admits infinitely many factorizations $(cB, A/c)$, while standard optimizers act on A and B separately. Functionally identical adapters can therefore receive different product-space updates solely from factor coordinates. Comparisons of non-zero initializations may mix task-relevant information with incidental geometry. Prior work studies asymmetric factor rates, gradient corrections, and invariant or Riemannian low-rank optimization (Hayou et al., 2024b; Wang et al., 2025; Yen et al., 2025; Zhang & Pilanci, 2024; Bogachev et al., 2026), but this ambiguity is less studied in online reasoning RL, where an initial discrepancy can affect later samples and rewards. The mechanism is not tied to policy gradients: whenever a loss depends on BA, gauge-related factors expose differently scaled gradients. Online RL is especially consequential because the discrepancy can alter subsequent rollouts and advantages.

We isolate this confound with iso-functional scalar gauges that preserve the adapter product and initial policy. Under shared-rate AdamW, functionally identical Qwen3-4B adapters exhibit up to a $6.49\times$ difference in early product-space movement. Frozen-rollout GRPO and teacher-forced SFT localize the discrepancy to factor-wise optimization rather than sampling, reward, or advantage estimation. Evidence spans initialization schemes, model scales, two tasks, and both fixed-data and 200-update online replications in a second model family, while reciprocal factor rates substantially reduce scalar-gauge sensitivity. Our primary evidence is paired product-space movement under exact initial-function audits; task accuracy is a secondary outcome. Figure 1 summarizes the intervention and evidence.

Contributions.

1. **A threat-to-validity test for LoRA initialization studies.** We construct iso-functional scalar-gauge interventions and show that incidental factor scale changes optimization despite an identical initial adapter function, including for a standard PiSSA non-zero parent.
2. **Mechanism localization across data regimes and objectives.** Identical frozen-rollout data isolate the immediate GRPO discrepancy, while teacher-forced SFT reproduces it without rollout, rewards, or advantages.
3. **Persistence and bounded controls.** Three-seed, 200-update product-space trajectories show finite-horizon persistence across Qwen3 and Llama model families, while reciprocal rates, scalar-covariance diagnostics, and LoRA-RITE delimit what a lightweight scalar correction can and cannot address.

2 RELATED WORK

Parameter-efficient adaptation and LoRA variants. Parameter-efficient fine-tuning includes adapters, prompt- and prefix-based methods, selective parameter updates, and low-rank adaptation (Houlsby et al., 2019; Li & Liang, 2021; Lester et al., 2021; Ben Zaken et al., 2022; Ding et al., 2023; Hu et al., 2022). LoRA variants modify rank, scaling, quantization, or weight decomposition (Zhang et al., 2023; Kalajdzievski, 2023; Dettmers et al., 2023; Li et al., 2024; Liu et al., 2024). Most relevant here, PiSSA and LoRA-GA construct non-zero factors from spectral or gradient information, and initialization changes LoRA dynamics (Meng et al., 2024; Wang et al., 2024; Hayou et al., 2024a). We ask whether comparisons of such structured initializations isolate task information from arbitrary factor coordinates.

Optimization geometry and scale-dependent parameterizations. The product BA is unchanged under $(BR, R^{-1}A)$, so factor-wise optimization can depend on coordinates that do not

identify a distinct adapter function. Analogous scale sensitivities motivate path-normalized and weight-normalized optimization (Neyshabur et al., 2015; Salimans & Kingma, 2016); structured preconditioners such as K-FAC and Shampoo further illustrate that optimizer geometry depends on the parameter representation (Martens & Grosse, 2015; Gupta et al., 2018). Fixed-rank matrix optimization admits quotient-manifold formulations (Mishra et al., 2014; Chi et al., 2019). LoRA+ chooses an asymmetric rate ratio for efficient feature learning, whereas LoRA-Pro adjusts factor gradients to approximate the full fine-tuning gradient (Hayou et al., 2024b; Wang et al., 2025). Neither objective is an iso-functional attribution test. Riemannian methods and LoRA-RITE instead target low-rank geometry or invariance under the general refactorization $(BR, R^{-1}A)$ (Zhang & Pilanci, 2024; Bogachev et al., 2026; Yen et al., 2025). We use its scalar subgroup as an exactly auditable causal intervention and reciprocal rates as a lightweight scalar control, not as a general invariant optimizer. Table 1 separates these optimization objectives from the attribution control studied here.

Table 1: Scope of related LoRA optimization methods. "Matrix operations" refers to additional low-rank matrix solves, inverses, or preconditioners beyond standard factor-wise optimization.

Method	Primary objective	Coordinate scope	Relation to this work
LoRA+	Efficient feature-learning scaling	Fixed asymmetric factor rates	Rate heuristic; not an iso-functional attribution test
LoRA-Pro	Approximate the full-FT gradient	Factor-gradient correction	Gradient approximation; requires additional $r \times r$ operations
Riemannian LoRA	Optimize quotient geometry	General low-rank coordinates	Geometry-aware optimizer with matrix operations
LoRA-RITE	Transformation invariance	General $(BR, R^{-1}A)$	General invariant baseline with adaptive $r \times r$ preconditioning
Reciprocal rates	Control a known scalar gauge	Positive scalar c	Lightweight attribution control; no matrix operations

Online post-training for mathematical reasoning. Preference-based post-training commonly builds on PPO and RLHF (Schulman et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022). Verifiable mathematical rewards enable group-relative policy optimization in DeepSeekMath, DeepSeek-R1, DAPO, and related reasoning systems (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025; Havrilla et al., 2024). We evaluate on GSM8K (Cobbe et al., 2021) with Qwen3 models (Qwen Team, 2025) and use both online GRPO and fixed-data SFT. Unlike work that primarily asks whether RL improves final reasoning accuracy, we ask whether two identical initial policies receive different optimizer updates solely because of LoRA factor scale. Frozen-rollout and SFT controls separate this local coordinate effect from subsequent sampling, reward, and advantage feedback.

3 Factor-Scale Confounding and Reciprocal Control

3.1 Scalar-Gauge Equivalence

For a frozen weight W_0, LoRA learns

$$W = W_0 + sBA, \qquad s = \alpha/r, \tag{1}$$

with $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$ (Hu et al., 2022). **Definition 1 (scalar gauge).** For $c > 0$,

$$T_c(B, A) = (B_c, A_c) = (cB, A/c), \qquad B_c A_c = BA. \tag{2}$$

Thus an exactly constructed gauge preserves the adapter product, logits, and initial policy while changing factor coordinates. It is the scalar subgroup of the general equivalence $(BR, R^{-1}A)$ studied by invariant low-rank methods (Yen et al., 2025).

3.2 Why Shared-Rate AdamW Is Gauge Sensitive

Let $G = \partial \mathcal{L}/\partial(BA)$. Chain rule gives

$$\nabla_A \mathcal{L} = B^\top G, \quad \nabla_B \mathcal{L} = G A^\top, \qquad \nabla_{A_c} \mathcal{L} = c \nabla_A \mathcal{L}, \quad \nabla_{B_c} \mathcal{L} = c^{-1} \nabla_B \mathcal{L}. \tag{3}$$

Ignoring decay for one step, Adam's element-wise direction for gradient g is $u(g) = m/(\sqrt{v} + \epsilon)$ (Loshchilov & Hutter, 2019). Under $g_c = cg$ and correspondingly scaled moments,

$$u(cg) = \frac{cm}{c\sqrt{v} + \epsilon} = \frac{m}{\sqrt{v} + \epsilon/c}. \tag{4}$$

For $\epsilon \ll c\sqrt{v}$, the normalized parameter steps are nearly unchanged: $\delta A_c \simeq \delta A$ and $\delta B_c \simeq \delta B$. The exact product increment, however, is

$$\begin{aligned} B_c^+ A_c^+ - B_c A_c &= \delta B_c A_c + B_c \delta A_c + \delta B_c \delta A_c \\ &\simeq c^{-1} \delta B A + c B \delta A + \delta B \delta A, \end{aligned} \tag{5}$$

whereas the balanced update is $\delta B A + B \delta A + \delta B \delta A$. The two first-order terms therefore receive reciprocal amplification despite an identical initial function. This local argument applies to any objective that depends on the factors through BA; online GRPO can additionally feed the first split into later samples, rewards, and advantages.

3.3 Reciprocal Rates and Residual Terms

We assign

$$\eta_A(c) = \eta/c, \qquad \eta_B(c) = c\eta. \tag{6}$$

Then $\delta A_c^{\mathrm{rec}} \simeq c^{-1} \delta A$ and $\delta B_c^{\mathrm{rec}} \simeq c \delta B$, so all terms of the product increment match:

$$\delta B_c^{\mathrm{rec}} A_c + B_c \delta A_c^{\mathrm{rec}} + \delta B_c^{\mathrm{rec}} \delta A_c^{\mathrm{rec}} \simeq \delta B A + B \delta A + \delta B \delta A. \tag{7}$$

Proposition 1 (idealized scalar equivariance). If the loss depends on A, B only through BA, Adam uses $\epsilon = 0$ without decay or clipping, and optimizer states and functional gradients are gauge matched, then Eq. 6 preserves $B_c^+ = cB^+$ and $A_c^+ = A^+/c$ at every step; hence $B_c^+ A_c^+ = B^+ A^+$. This is not an invariance theorem for the full AdamW–GRPO stack.

AdamW exposes the principal residuals directly. A tighter one-step diagnostic covaries

$$\lambda_A = c\lambda, \quad \lambda_B = \lambda/c, \qquad \epsilon_A = c\epsilon, \quad \epsilon_B = \epsilon/c, \tag{8}$$

with Eq. 6. Global clipping remains non-equivariant because the LoRA contribution to its squared norm becomes

$$c^2 \|\nabla_A \mathcal{L}\|_F^2 + c^{-2} \|\nabla_B \mathcal{L}\|_F^2. \tag{9}$$

BF16 arithmetic, numerical order, unmatched optimizer history, and online feedback add further residuals. The scalar-covariance variant is therefore a mechanistic diagnostic; reciprocal rates are the lightweight intervention used in the main trajectories.

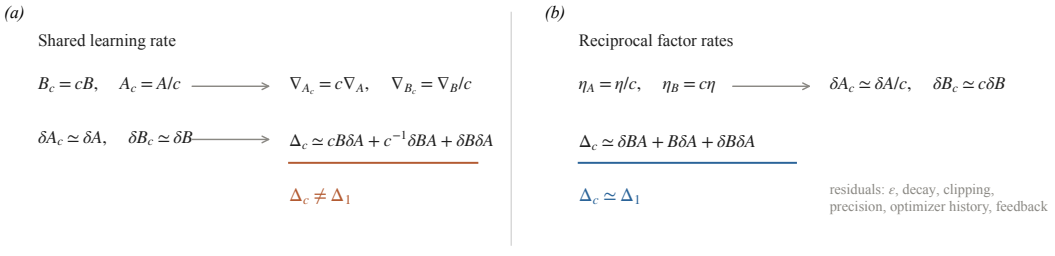

Figure 2: Scalar-gauge mechanism and reciprocal correction. A gauge leaves the adapter function unchanged but rescales the two factor gradients oppositely. Shared-rate Adam produces similarly scaled coordinate steps, which changes the first-order product terms. Reciprocal factor rates restore all terms of the idealized product increment, including the second-order cross term.

4 Experiments

4.1 Evaluation Design

We apply exact product-preserving gauges to non-zero adapters and compare a shared-rate optimizer with Eq. 6. Frozen-rollout GRPO holds responses, rewards, and advantages fixed; teacher-forced

SFT removes the online loop; PiSSA supplies a standard non-zero initialization; Qwen3-1.7B/4B/8B runs track GSM8K and DAPO-Math through 200 updates; and Llama-3.2-3B provides both a fixed-data SFT control and a three-seed, 200-update online-GRPO replication on DAPO-Math. We intentionally hold the base rate and all non-gauge settings fixed: this is a paired causal comparison, not a best-tuned performance contest. Complete objectives and configurations are in Appendix A.

The primary mechanistic endpoint is aggregate adapter-product displacement

$$U_{s\to t} = \left[\sum_{\ell} \left\| \frac{\alpha_\ell}{r_\ell}(B_{\ell,t}A_{\ell,t} - B_{\ell,s}A_{\ell,s}) \right\|_F^2 \right]^{1/2}. \tag{10}$$

Task trajectories use validation-accuracy AUC as a secondary descriptive outcome. Contrasts are paired within training seed; the two gauge directions share one balanced arm and are nested, not independent replicates. We report mean $\pm$ sample SD over seeds 42–44 and make no formal equivalence claim.

4.2 Main Results

4.2.1 Gauge-Only Effect and Frozen-Rollout Isolation

In the controlled Qwen3-4B intervention, shared-rate extreme gauges move $5.56\times$ and $6.49\times$ as far as the balanced arm by update 2 despite identical initial products and fixed-input logits. Reciprocal rates reduce the corresponding errors to $+2.65\%$ and -2.40%. The identical frozen-rollout control reuses response token IDs, rewards, and advantages, so its one-step split is an optimizer-coordinate effect rather than a sampling difference.

Figure 3 tests the predicted dependence on factor scale. Shared-rate displacement grows smoothly with $|\log_2 c|$, reaching 6.23–$6.87\times$ at the extreme gauges. Reciprocal rates remain within 4.40% of the balanced reference, while the optional full AdamW scalar-covariance diagnostic is within 2.06% across the sweep.

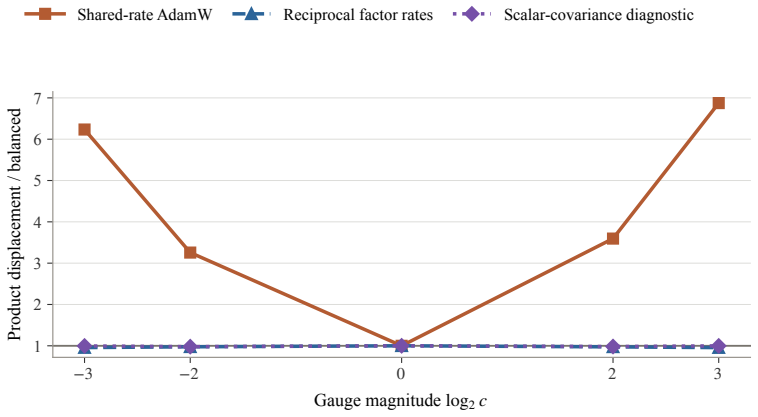

Figure 3: Controlled scalar-gauge stress profile on one Qwen3-4B frozen rollout and training seed. The vertical axis is one-step adapter-product displacement relative to the balanced gauge. Shared-rate AdamW becomes increasingly scale-sensitive, reciprocal factor rates remove most of the distortion, and covarying AdamW learning rate, weight decay, and ϵ provides a tighter one-step scalar-covariance diagnostic. The point at $c = 1$ is the common balanced reference; this mechanistic profile is not a population estimate.

4.2.2 Across Initialization, Objective, and Model Family

Table 2 removes distinct alternative explanations. PiSSA rules out dependence on the transported parent, teacher-forced SFT removes online sampling and reward feedback, and a three-seed Llama-3.2-3B replication transfers the fixed-data mechanism beyond the Qwen3 family. Across these settings, shared-rate optimization amplifies product movement, whereas reciprocal rates remain close to the balanced reference.

Table 2: Independent controls of the optimizer-coordinate mechanism. PiSSA entries are ranges over the two direction-wise three-seed means; SFT entries are mean ± sample SD over three independent training seeds. The Llama SFT block is a one-direction, fixed-data replication and by itself does not test cross-family online-feedback propagation.

Setting	Explanation ruled out	Shared AdamW / balanced	Reciprocal LR / balanced
Qwen3-4B PiSSA, GRPO	Transported-parent artifact	6.05–6.38×	0.964–1.012×
Qwen3-4B, SFT	Online rollout and reward feedback	$(4.442 \pm 0.007)\times$	$(0.9947 \pm 0.0024)\times$
Llama-3.2-3B, SFT ($c = 8$)	Qwen-family specificity	$(9.353 \pm 0.157)\times$	$(0.961 \pm 0.024)\times$

4.2.3 Across Scale, Task, and Horizon

A three-seed, 200-update Llama-3.2-3B replication extends the online product-space result to a second model family. Under the $c = 8$ shared-rate gauge, displacement is $(8.673 \pm 2.258)\times$, $(9.421 \pm 1.240)\times$, and $(3.162 \pm 0.344)\times$ the within-seed balanced reference at updates 2, 10, and 200. Reciprocal rates yield $(0.838 \pm 0.144)\times$, $(1.004 \pm 0.108)\times$, and $(0.966 \pm 0.108)\times$, removing 97.57%, 99.06%, and 96.32% of the shared-rate displacement deviation. Task-level AUC, endpoint-accuracy, and KL gaps shrink in the same direction; we treat them as descriptive because free-generation step-0 accuracy varies across iso-functional arms (Appendix F.2).

Across Qwen3 scales, reciprocal and balanced trajectories remain close through 200 GRPO updates on GSM8K (Figure 4). DAPO-Math provides the same paired design on a second task.

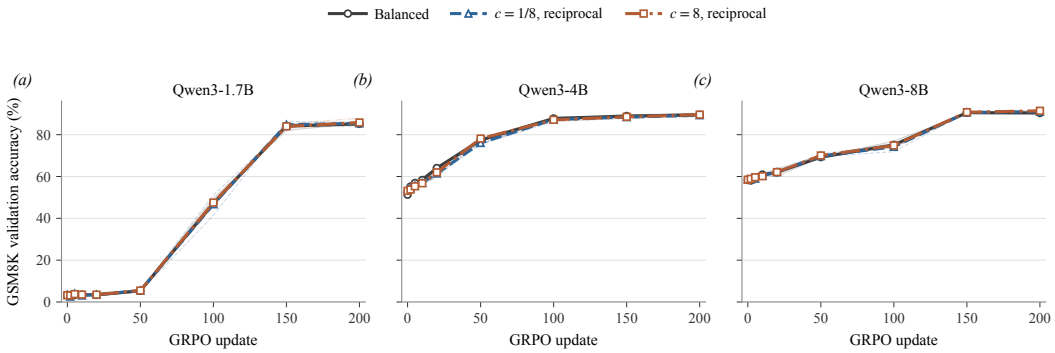

Figure 4: GSM8K validation trajectories through 200 GRPO updates. Thin translucent lines show the three individual training seeds, while thicker open-marker lines show their mean. All panels share the same vertical scale. Balanced and both reciprocal-rate gauges remain close across Qwen3 model scales; the displayed variability is descriptive and is not a confidence interval.

Figure 5 separates task-level agreement from residual product-space error. Reciprocal rates substantially suppress the early DAPO-Math discrepancy, although task- and scale-dependent residuals remain. Thus the intervention reduces scalar-gauge sensitivity without establishing exact trajectory equivalence.

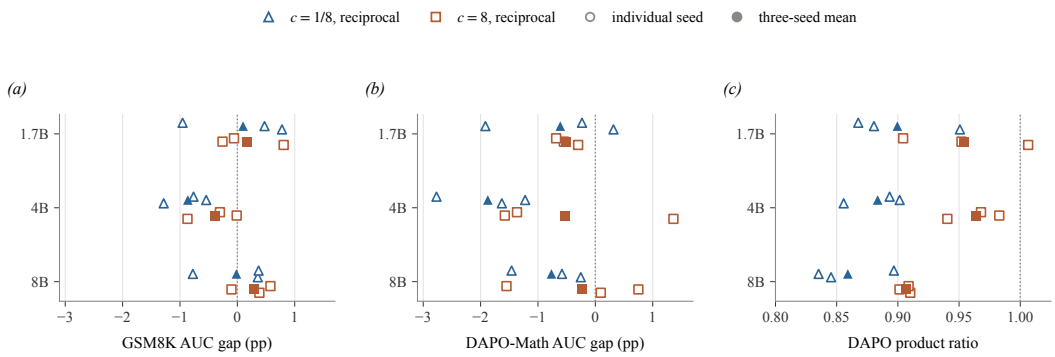

Figure 5: Cross-scale paired evidence for reciprocal factor rates. Panels (a) and (b) show reciprocal-minus-balanced validation-AUC differences through 200 GRPO updates on GSM8K and DAPO-Math. Panel (c) shows the early DAPO-Math adapter-product displacement ratio. Hollow markers are within-seed gauge comparisons; filled markers are direction-wise three-seed means. The two gauge directions share the balanced arm within each seed and are nested rather than independent replicates. Dotted reference lines mark zero AUC difference and unit product ratio.

Table 3 summarizes the long-horizon residuals after the two gauge directions are averaged within each seed. Their model- and task-dependent, mixed signs provide no evidence of uniform performance improvement.

Table 3: Long-horizon reciprocal-minus-balanced results. Within each seed, the two gauge contrasts are averaged first; entries are mean $\pm$ sample SD over three independent seeds. Values near zero indicate trajectory agreement, not a performance gain.

Task	Model	AUC gap, 0–200 (pp)	Accuracy@200 gap (pp)
GSM8K	1.7B	$+0.131 \pm 0.650$	$+0.456 \pm 0.963$
	4B	-0.629 ± 0.408	$+0.000 \pm 0.851$
	8B	$+0.135 \pm 0.500$	$+0.846 \pm 0.407$
DAPO-Math	1.7B	-0.561 ± 0.627	-0.600 ± 0.458
	4B	-1.201 ± 0.982	-0.200 ± 2.066
	8B	-0.500 ± 0.874	-0.433 ± 0.513

4.2.4 Boundary of Scalar Correction

Reciprocal rates are a scalar-coordinate control rather than a general invariant optimizer. Early DAPO-Math under-shoot and residual long-horizon differences rule out exact practical invariance. Scalar-covariance diagnostics isolate AdamW-specific residuals, while LoRA-RITE provides the broader refactorization-invariant baseline.

The scalar intervention is an attribution test, not a claim that realistic adapters generally differ by one global scale: native PiSSA is approximately balanced, and the audited adapter collection is better described by richer rank-space geometry. Fixed-token probes and an SGD ablation support the optimizer-coordinate mechanism beyond the primary product metric (Appendix L). The reciprocal control requires only separate A/B parameter groups and introduces no additional model parameters or forward computation.

5 Discussion

The practical issue is attribution: comparisons among non-zero LoRA initializations can mix task-relevant information with incidental factor coordinates. Initialization studies should therefore report factor geometry, include an iso-functional or invariant-optimizer control, and distinguish product-space, fixed-input output, and task-level endpoints.

SFT shows that GRPO does not create the ambiguity, while the Qwen3 and Llama online trajectories show that coordinate-dependent product movement can persist under policy feedback. Final task accuracy may converge even when optimization paths differ, so endpoint agreement alone does not validate an attribution about early training dynamics.

Reciprocal rates provide a low-overhead diagnostic when a positive scalar gauge is known and separate A/B optimizer groups are available. They complement, rather than replace, invariant or Riemannian methods designed for general rank-space transformations.

6 Conclusion and Limitations

Functionally identical LoRA factorizations can follow different optimization trajectories under factor-wise AdamW. Iso-functional gauges, frozen-rollout GRPO, and fixed-data SFT localize the discrepancy to optimizer coordinates; 200-update Qwen3 and Llama trajectories show finite-horizon persistence. Reciprocal factor rates substantially reduce scalar-gauge sensitivity but do not provide general refactorization invariance.

Limitations. Long-horizon online evidence spans Qwen3 and Llama-3.2, but the Llama replication covers one model size, DAPO-Math, and the $c = 8$ gauge direction. The experiments establish a controlled scalar-gauge confound; they do not imply that published initialization gains are generally artifacts or that realistic adapters are well described by one global scale.

Product displacement is a mechanistic proxy rather than a global function distance, and fixed-token probes sample only part of the output space. Free-generation variability makes task-level trajectory summaries descriptive; three-seed results are not formal equivalence tests. Reciprocal rates should therefore be interpreted as a lightweight attribution control, not as a universal performance improvement or a substitute for optimizers that handle general matrix refactorizations.

REPRODUCIBILITY STATEMENT

The appendices report the exact GRPO and SFT objectives, canonical optimizer and LoRA settings, task-specific horizons, evaluation procedures, and seed-level paired results. We provide an end-to-end protocol for balancing non-zero LoRA factors, applying function-preserving scalar gauges, auditing post-serialization adapter products and fixed-input logits, and creating disjoint factor-wise optimizer groups. All summary statistics treat training seed as the independent unit and gauge arms as comparisons nested within seed. Complete Qwen3 cross-scale trajectories, the Llama-3.2-3B SFT and online-GRPO replications, scalar-covariance diagnostics, negative coordinate audits, and optimizer controls appear in the corresponding appendices.

ETHICS STATEMENT

This work studies optimization behavior using publicly available language models and mathematical-reasoning benchmarks. It does not involve human subjects, private or personally identifiable data, or a new dataset. The main risks are incorrect attribution of empirical gains, overgeneralization beyond the evaluated settings, and accelerator use. Function-preserving interventions and optimizer controls should not be interpreted as guarantees of model safety, reliability, or downstream fairness.

AI USE STATEMENT

The authors used large language model (LLM) tools to assist with code development and debugging, literature retrieval and verification, feedback on experimental controls, manuscript drafting and revision, language editing, structural organization, and checks of mathematical derivations, internal consistency, and numerical calculations. The core research questions, main ideas, hypotheses, experimental objectives, and scientific conclusions were conceived and determined by the authors. All experimental data and quantitative results reported in this paper were obtained from experiments that were actually executed and were not generated, inferred, or fabricated by LLM tools. The authors reviewed, tested, verified, and revised all LLM-assisted code, text, derivations, and analyses as appropriate, and take full responsibility for the originality, accuracy, and entire content of this paper.

REFERENCES

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 1–9, 2022. doi: 10.18653/v1/2022.acl-short.1. URL https://aclanthology.org/2022.acl-short.1.

Vladimir Bogachev, Vladimir Aletov, Alexander Molozhavenko, Denis Bobkov, Vera Soboleva, Aibek Alanov, and Maxim Rakhuba. LoRA meets Riemannion: Muon optimizer for parametrization-independent low-rank adapters. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=WtbXgc9GVA.

Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019. doi: 10.1109/TSP.2019.2937282. URL https://doi.org/10.1109/TSP.2019.2937282.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5:220–235, 2023. doi: 10.1038/s42256-023-00626-4. URL https://doi.org/10.1038/s42256-023-00626-4.

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

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850, 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*, 2024. doi: 10.48550/arXiv.2403.04642. URL https://arxiv.org/abs/2403.04642.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on LoRA finetuning dynamics. In *Advances in Neural Information Processing Systems*, volume 37, pp. 117015–117040, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d4387c37b3b06e55f86eccdb8cd1f829-Abstract-Conference.html.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17783–17806, 2024b. URL https://proceedings.mlr.press/v235/hayou24a.html.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023. URL https://arxiv.org/abs/2312.03732.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353.

Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatziakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. LoftQ: LoRA-fine-tuning-aware quantization for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LzPWWPAdY4.

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121038–121072, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db36f4d603cc9e3a2a5e10b93e6428f2-Abstract-Conference.html.

Bamdev Mishra, Gilles Meyer, Silvere Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and Riemannian low-rank optimization. *Computational Statistics*, 29:591–621, 2014. doi: 10.1007/s00180-013-0464-z. URL https://doi.org/10.1007/s00180-013-0464-z.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2422–2430, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/eaa32c96f620053cf442ad32258076b9-Abstract.html.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29, pp. 901–909, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/ed265bc903a5a097f61d3ec064d96d2e-Abstract.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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

Shangshang Wang, Julian Asilis, Ömer Faruk Akgül, Enes Burak Bilgin, Ollie Liu, and Willie Neiswanger. Tina: Tiny reasoning models via LoRA. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=P2OXYO3bEe.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 54905–54931, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/62c4718cc334f6a0a62fb81c4a2095a1-Abstract-Conference.html.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. LoRA-Pro: Are low-rank adapters properly optimized? In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gTwRMU3lJ5.

Jui-Nan Yen, Si Si, Zhao Meng, Felix X. Yu, Sai Surya Duvvuri, Inderjit S. Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. LoRA done RITE: Robust invariant transformation equilibration for LoRA optimization. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VpWki1v2P8.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pp. 125532–125554, 2025. doi: 10.52202/085713-3775. URL https://doi.org/10.52202/085713-3775.

Fangzhao Zhang and Mert Pilanci. Riemannian preconditioned LoRA for fine-tuning foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59641–59669, 2024. URL https://proceedings.mlr.press/v235/zhang24ax.html.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. doi: 10.48550/arXiv.1909.08593. URL https://arxiv.org/abs/1909.08593.

A Complete GRPO Objective and Hyperparameters

For prompt x, the old policy samples responses y_i. With $\rho_{i,t}(\theta) = \pi_\theta(y_{i,t} \mid x, y_{i,<t}) / \pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})$, the actor uses the token-mean clipped surrogate

$$\min\Big(\rho_{i,t}\hat{A}_i, \text{clip}(\rho_{i,t}, 0.8, 1.2)\hat{A}_i\Big), \tag{11}$$

with group-normalized advantages. The low-variance KL term uses $d_t = \ell_q - \ell_p$, clamps d_t to $[-20, 20]$, computes $k_3 = \exp(d_t) - d_t - 1$, clamps k_3 to $[-10, 10]$, and averages over response tokens. The coefficient is 10^{-3}; no KL is included in reward and entropy coefficient is zero.

Table 4: Core experimental configuration. Unless noted, values apply to GRPO.

Item	Configuration
Models / tasks	Qwen3-1.7B/4B/8B GRPO on GSM8K and DAPO-Math; Llama-3.2-3B GRPO on DAPO-Math; Qwen3-4B and Llama-3.2-3B teacher-forced SFT
LoRA / gauges	rank 16, alpha 32, q/k/v/o; $c \in \{1/8, 1, 8\}$; moderate $\{1/4, 1, 4\}$
Optimizer	AdamW; LR 3×10^{-6}; $(\beta_1, \beta_2) = (0.9, 0.999)$; $\epsilon = 10^{-8}$; decay 0.01
Schedule / clipping	constant LR; zero warmup; gradient norm 1.0; policy ratio $[0.8, 1.2]$
GRPO	5 responses/prompt; sample-SD advantages with 10^{-6}; token mean; KL 10^{-3}; entropy 0
Batch / lengths	train batch 32; PPO minibatch 16; GSM8K prompt/response 512/512, dynamic cap 8,192; DAPO-Math 4,096/32,768, dynamic cap 40,960; Llama model context 36,864
Reward / validation	GSM8K strict #### exact match; unparseable or limit-hit responses incorrect; $n = 1$ greedy
Online horizons / seeds	Qwen3 and Llama GRPO: 200 updates, checkpoints 0/2/5/10/20/50/100/150/200; seeds 42–44
PiSSA	PEFT `pissa_niter_16`; FP32 construction, BF16 adapter serialization, residual base, power-of-two gauges
SFT difference	7,217 reference solutions; sequence length 2,048; weight decay 0
Hardware	Qwen3 GRPO: H100 $\times 2$ per run; Llama GRPO: H100 $\times 4$ per run; rollout tensor parallelism 1

B Standard PiSSA Initialization: Audit and Complete Results

The valid PiSSA protocol serializes the parent adapter in BF16 before applying power-of-two gauges. Across $c \in \{1/8, 1, 8\}$, maximum tensor relative-L_2 error, full-vocabulary logit max/mean error, and audit response disagreement are all zero. Table 5 gives paired seed-level results.

Table 5: PiSSA paired results by training seed. Product is arm/balanced at update 2; AUC is arm−balanced over updates 0–10.

Gauge	Seed	Rule	Product ratio	AUC difference
$c = 1/8$	42	Shared / corrected	5.753 / 0.928	+13.77 / −2.11 pp
	43	Shared / corrected	5.787 / 0.937	+11.93 / +1.09 pp
	44	Shared / corrected	6.619 / 1.026	+17.70 / +1.15 pp
$c = 8$	42	Shared / corrected	6.503 / 1.021	+7.85 / +0.25 pp
	43	Shared / corrected	6.131 / 0.939	+8.30 / +0.21 pp
	44	Shared / corrected	6.517 / 1.074	+13.98 / −1.29 pp

C Full AdamW Scalar-Covariance Diagnostic

C.1 Frozen-Rollout Numerical Details

Table 6 gives the exact one-step product-displacement values underlying the extreme-gauge points in Figure 3.

Table 6: One-step adapter-product displacement on identical frozen-rollout data. All arms reuse response tokens, rewards, and advantages; optimizer moments start at zero.

Update rule	Gauge	$U_{0\to1}$	Contrast to balanced
Shared AdamW	1	0.002396	reference
	$1/8$	0.014936	$6.23\times$
	8	0.016469	$6.87\times$
Reciprocal LR	$1/8$	0.002292	-4.32%
	8	0.002290	-4.40%
LR+WD+ϵ covariance	$1/8$	0.002387	-0.35%
	8	0.002391	-0.22%

C.2 Covariance Audit

This one-step seed-42 audit reuses the exact frozen rollout. Relative to balanced $U_{0\to1} = 0.002395800$, LR-only gives errors -4.317% and -4.403% for $c = 1/8$ and $c = 8$; LR+WD+ϵ covariance gives -0.353% and -0.217%. The latter is a controlled mechanism audit and is not used in the 200-update correction.

D Functional-Equivalence Audits

D.1 Qwen3-4B Gauge Audits

Table 7 reports the recorded serialized equivalence audits for both the extreme gauges used in the primary experiment and the moderate gauges used in the ablation.

Table 7: Serialized function and fixed-input logit equivalence for Qwen3-4B. Cross-gauge errors compare each factorization with its balanced reference.

Gauge	Rel. BA err.	Abs. BA err.	Max logit err.	Mean logit err.	Pass
$c = 1/8$	0	0	0	0	Yes
$c = 1/4$	0	0	0	0	Yes
$c = 1$	reference	reference	reference	reference	–
$c = 4$	0	0	0	0	Yes
$c = 8$	0	0	0	0	Yes

The fixed-trajectory audit covers full-vocabulary logits for all response tokens on the first 32 GSM8K prompts. The BF16-serialized delta has a maximum relative error of 2.78×10^{-3} relative to the unquantized FP32 balanced delta—a quantization error, not a cross-gauge error. Across-gauge BA errors are zero.

D.2 Greedy Non-Determinism Control

Repeated greedy generation with the identical balanced ($c = 1$) adapter yields only 84.38% string agreement, while prompt-end logit errors are zero. Free-generation string comparison is therefore not used as a functional-equivalence criterion.

D.3 Qwen3-1.7B and Qwen3-8B Audits

All per-tensor BA relative and maximum-absolute errors are zero; 32-prompt final-token full-vocabulary logit max/mean errors are zero for both scales.

E TEACHER-FORCED SFT OBJECTIVE CONTROL

E.1 PROTOCOL AND ENDPOINT DEFINITIONS

We train Qwen3-4B on 7,217 GSM8K reference solutions using teacher-forced causal-LM NLL. The initial rank-16 attention adapter and three factorization arms match the transported case study. Runs use seeds 42–44, 10 updates, global batch size 32, maximum sequence length 2,048, BF16 computation, constant learning rate 3×10^{-6}, AdamW betas $(0.9, 0.999)$, default $\epsilon = 10^{-8}$, weight decay 0, and global gradient-norm clipping at 1.0. The corrected arm uses $\eta_{A,\ell} = \eta/c_\ell$ and $\eta_{B,\ell} = c_\ell \eta$ for all 144 factor pairs.

The ten-step matched-token probe uses 32 prompts and one shared set of 7,663 response positions for all arms within each seed. We teacher-force every initial and updated adapter on these tokens and compute full-vocabulary logit RMS and $D_{\mathrm{KL}}(p_0 \| p_{10})$. A separate one-step audit used 7,635 response positions; the probe sequence set was regenerated for that audit, so its token count differs from the ten-step probe.

E.2 ONE-STEP MECHANISM AUDIT

Table 8: One-step SFT audit on seed 42. All arms have identical first-batch pre-update NLL. Output probes use matched teacher-forced tokens.

Endpoint	Control	Original	Corrected	Orig./Ctrl.	Corr./Ctrl.
First-batch NLL	1.538925	1.538925	1.538925	1.000×	1.000×
Product displacement	0.002105	0.018302	0.002069	8.694×	0.983×
Logit RMS	0.081918	0.089058	0.086216	1.087×	1.052×
Fixed-probe KL	0.000476	0.000460	0.000468	0.966×	0.983×

E.3 TEN-STEP SEED-LEVEL RESULTS

Table 9: Ten-step SFT results by training seed. NLL entries are differences from Control; product-displacement, logit-RMS, and fixed-probe KL entries are ratios to Control.

Endpoint	Arm	Seed 42	Seed 43	Seed 44
NLL difference	Original	−0.055889	−0.054259	−0.053935
	Corrected	−0.000034	+0.000381	+0.000920
Product U ratio	Original	4.4362	4.4502	4.4400
	Corrected	0.9928	0.9974	0.9939
Logit RMS ratio	Original	3.0156	3.2525	2.7786
	Corrected	1.0179	1.0435	0.9126
Fixed-probe KL ratio	Original	1.2320	1.4674	1.2799
	Corrected	1.0913	1.0737	0.8405

Mean Control NLL is 1.523486 ± 0.000597. Across seeds, the absolute gap to Control is reduced descriptively by $99.18 \pm 0.83\%$ for mean NLL, $99.85 \pm 0.07\%$ for product displacement, $97.42 \pm 2.09\%$ for logit RMS, and $62.63 \pm 20.67\%$ for fixed-probe KL. The KL recovery is visibly weaker and noisier than product and raw-logit recovery; none of these summaries is a formal equivalence test. The ten recorded NLL values within a run are correlated and include the shared first-batch pre-update value; only training seeds are treated as independent repeats.

F LLAMA-3.2-3B CROSS-FAMILY REPLICATIONS

F.1 FIXED-DATA SFT REPLICATION

This confirmatory fixed-data experiment uses Llama-3.2-3B-Instruct with a standard PiSSA non-zero parent, rank-16 LoRA with alpha 32 on q/k/v/o projections, and ten teacher-forced GSM8K updates. Within each training seed, the balanced, $c = 8$ shared-rate, and $c = 8$ reciprocal-rate arms use the same examples and ordering. Optimizer moments start from zero. Post-serialization adapter-product and fixed-input function audits are performed before training.

Table 10: Cross-family SFT replication. $U_{0\to10}$ denotes aggregate LoRA-product Frobenius displacement. Ratios are computed within seed. Recovery is $1 - |r_{\text{recip}} - 1|/|r_{\text{shared}} - 1|$, reported as a percentage.

Seed	$U_{0\to10}$			Ratio to balanced		Recovery (%)
	Balanced	Shared	Reciprocal	Shared	Reciprocal	
42	0.1166	1.1101	0.1094	9.523	0.939	99.28
43	0.1181	1.1010	0.1165	9.323	0.987	99.84
44	0.1189	1.0951	0.1140	9.213	0.959	99.51
	Mean ± SD			9.353 ± 0.157	0.961 ± 0.024	99.54 ± 0.28

F.2 ONLINE GRPO THROUGH 200 UPDATES

This confirmatory experiment uses the same Llama-3.2-3B PiSSA residual parent, rank-16 attention LoRA, and $c = 8$ gauge direction. Balanced, shared-rate, and reciprocal-rate arms are trained for 200 online GRPO updates on the DAPO-Math clean-10K split and evaluated on 500 examples with seeds 42–44. Runs use five responses per prompt, train batch 32, PPO minibatch 16, prompt/response/model limits 4,096/32,768/36,864, AdamW at base learning rate 3×10^{-6} and weight decay 0.01, rollout tensor parallelism 1, and parameter/optimizer offload. Validation uses $n = 1$ greedy decoding at checkpoints 0/2/5/10/20/50/100/150/200. All nine formal runs completed; smoke tests are excluded from statistics. Post-serialization $c \in \{1/8, 1, 8\}$ adapters have zero cross-gauge product and audited full-vocabulary logit error, and identical deterministic 32-token response hashes on 32 fixed prompts.

Table 11: Llama-3.2-3B online-GRPO adapter-product displacement. Absolute values are mean ± sample SD over three independent training seeds. Ratios and recovery are computed within seed before aggregation. Recovery is $100\,(1 - |U_{\text{recip}} - U_{\text{bal}}|/|U_{\text{shared}} - U_{\text{bal}}|)$.

Update	Product displacement $U_{0\to t}$			Ratio to balanced		Recovery (%)
	Balanced	Shared	Recip.	Shared	Recip.	
2	0.0361 ± 0.0085	0.3003 ± 0.0078	0.0297 ± 0.0057	8.673 ± 2.258	0.838 ± 0.144	97.57 ± 2.21
10	0.0834 ± 0.0118	0.7762 ± 0.0105	0.0830 ± 0.0057	9.421 ± 1.240	1.004 ± 0.108	99.06 ± 0.33
200	1.1935 ± 0.1124	3.7481 ± 0.0792	1.1453 ± 0.0682	3.162 ± 0.344	0.966 ± 0.108	96.32 ± 4.30

Table 12: Llama-3.2-3B online-GRPO task-level and KL results through update 200. Entries are paired differences or ratios to balanced, reported as mean ± sample SD over three training seeds. Task metrics are descriptive because free-generation step-0 accuracy varies across iso-functional arms.

Arm vs. balanced	AUC gap (pp)	Accuracy@200 gap (pp)	KL@200 ratio
Shared ($c = 8$)	-2.051 ± 0.719	-4.933 ± 2.873	1.199 ± 0.099
Reciprocal ($c = 8$)	-0.795 ± 0.385	-0.800 ± 0.872	0.940 ± 0.045

All three seeds show shared-rate amplification of product movement, while reciprocal rates recover most of the deviation at every reported horizon. The shared-rate AUC gap is negative for every seed,

and all three shared-rate accuracy trajectories finish below their within-seed balanced references. Reciprocal rates reduce the task-level and KL discrepancies. Mean training clip rates are near zero in all arms, so response-length clipping does not explain the shared-rate separation. We make no claim of exact invariance or statistical equivalence.

G Transported Initialization and Independent Re-evaluation

G.1 Static Audit and Implementation

The case study uses one non-zero Qwen3-4B adapter obtained by transporting a rank-16 LoRA adapter from Qwen3-0.6B through a fixed linear cross-model mapping. The mapping and transported adapter were fixed before the gauge intervention and were not selected based on the case-study outcomes. The adapter contains 144 complete LoRA A/B pairs. The nearest-power-of-two gauges satisfy $c_\ell \in [1/32, 1/4]$ (median $1/16$). Serialized adapter-product and audited 32-prompt fixed-logit errors between Original and Control are zero. Because the ideal balancing scales are quantized to powers of two, the post-control factor norms are not necessarily equal; their ratio is bounded within $[1/2, 2]$.

The corrected implementation uses $\eta_{A,\ell} = \eta/c_\ell$ and $\eta_{B,\ell} = c_\ell \eta$. With base $\eta = 3 \times 10^{-6}$, the A rates span 1.2×10^{-5}–9.6×10^{-5} and the B rates span 9.375×10^{-8}–7.5×10^{-7}. All 144 factor pairs were matched by the parameter-group audit. The scalar optimizer log records only the first parameter group and is therefore not used as a complete learning-rate audit.

G.2 Complete Run-Level Results

Table 13: All nine runs for the transported-initialization case study. Validation accuracy and AUC are percentages, rounded to two decimal places. Actor KL is computed on each arm's own online rollout and is not a matched-token comparison.

Seed	Arm	Validation performance (%)					Mechanistic endpoints	
		Step 0	Step 2	Step 5	Step 10	AUC	$U_{0\to 2}$	Actor KL@10
42	Original	51.56	58.59	67.19	76.17	65.72	0.044428	0.005324
	Control	53.91	54.69	55.08	61.33	56.43	0.010337	0.001207
	Corrected	53.13	53.52	55.08	57.81	55.18	0.009457	0.000952
43	Original	52.73	60.16	66.80	73.44	65.39	0.045653	0.003458
	Control	54.69	53.91	55.47	57.42	55.49	0.009963	0.000845
	Corrected	51.95	55.47	56.25	57.81	56.02	0.009812	0.000614
44	Original	52.73	60.16	66.80	73.44	65.39	0.046203	0.003024
	Control	51.95	58.59	55.86	57.03	56.45	0.010795	0.000908
	Corrected	53.52	53.52	55.86	58.20	55.63	0.010572	0.000941

Seeds 43 and 44 have identical aggregate Original accuracies at the four reported checkpoints but different product displacement and actor KL. The six planned runs completed, and their aggregate values were cross-checked against the run summaries; sample-level validation generations were not retained. We independently re-evaluated the seed-43/44 checkpoints three times. Adapter checksums differ across seeds, and the response sets differ across evaluations, even when aggregate accuracy matches. Accuracy ranges from repeated evaluation are 0–1.56 pp across checkpoints. This rules out a shared checkpoint/cache explanation for the identical original aggregate points, but also shows that small AUC differences are not high-precision effects. We therefore treat transported AUC as secondary and base the mechanistic claim on initial product/logit audits and product displacement.

H Cross-Scale and Cross-Task Complete Results

Each reciprocal-rate contrast uses its same-model, same-seed balanced arm; the two gauges within a seed are nested comparisons. The tables below report the complete values needed to audit the main summaries.

H.1 DAPO-Math through 200 Updates

Table 14: DAPO-Math validation performance through 200 GRPO updates. Values are mean ± sample SD over three independent training seeds.

Model	Training arm	Validation performance (%)	
		$\text{AUC}_{0\text{–}200}$	Accuracy_{200}
1.7B	Balanced	18.371 ± 0.252	26.467 ± 0.462
	Reciprocal ($c = 1/8$)	17.760 ± 0.958	26.133 ± 0.462
	Reciprocal ($c = 8$)	17.860 ± 0.145	25.600 ± 0.721
4B	Balanced	34.683 ± 1.181	47.000 ± 2.107
	Reciprocal ($c = 1/8$)	32.809 ± 0.944	46.600 ± 0.346
	Reciprocal ($c = 8$)	34.155 ± 0.769	47.000 ± 1.058
8B	Balanced	46.290 ± 0.579	48.733 ± 0.503
	Reciprocal ($c = 1/8$)	45.525 ± 0.551	47.667 ± 0.945
	Reciprocal ($c = 8$)	46.055 ± 0.689	48.933 ± 0.757

Table 15: DAPO-Math reciprocal-minus-balanced paired summaries through update 200. Means and sample SDs use three within-seed contrasts per gauge.

Model	Gauge	AUC gap (pp)	Acc.@200 gap (pp)	KL@200 ratio
1.7B	$c = 1/8$	-0.611 ± 1.163	-0.333 ± 0.462	1.008 ± 0.092
	$c = 8$	-0.511 ± 0.194	-0.867 ± 0.945	1.036 ± 0.106
4B	$c = 1/8$	-1.874 ± 0.801	-0.400 ± 1.833	0.893 ± 0.021
	$c = 8$	-0.528 ± 1.639	0.000 ± 2.458	0.956 ± 0.085
8B	$c = 1/8$	-0.765 ± 0.624	-1.067 ± 1.332	0.853 ± 0.102
	$c = 8$	-0.235 ± 1.184	$+0.200 \pm 0.400$	0.786 ± 0.068

H.2 DAPO-Math Seed-Level Paired Contrasts

Table 16: All 18 DAPO-Math reciprocal-minus-balanced AUC and step-200 accuracy contrasts.

Model	Seed	Gauge	AUC gap (pp)	Accuracy gap (pp)
1.7B	42	$1/8$	$+0.316$	-0.600
	42	8	-0.300	-1.600
	43	$1/8$	-1.917	-0.600
	43	8	-0.550	$+0.200$
	44	$1/8$	-0.233	$+0.200$
	44	8	-0.683	-1.200
4B	42	$1/8$	-1.630	$+1.200$
	42	8	$+1.361$	$+2.800$
	43	$1/8$	-1.224	0.000
	43	8	-1.578	-1.000
	44	$1/8$	-2.769	-2.400
	44	8	-1.366	-1.800
8B	42	$1/8$	-0.254	-0.400
	42	8	$+0.095$	-0.200
	43	$1/8$	-0.582	-2.600
	43	8	$+0.750$	$+0.600$
	44	$1/8$	-1.461	-0.200
	44	8	-1.549	$+0.200$

H.3 GSM8K THROUGH 200 UPDATES

The GSM8K evaluation uses the same paired design. The main text reports its aggregate range because DAPO-Math supplies the second long-horizon task. Across models and gauge directions, GSM8K AUC gaps span $[-1.284, +0.808]$ pp and endpoint-accuracy gaps span $[-0.781, +2.344]$ pp.

I CONTROLLED SHARED-RATE TRAJECTORIES

The table below supplies the run-level shared-rate trajectory underlying the controlled Qwen3-4B amplification result. This experiment is complementary to the cross-scale reciprocal-rate evaluation: it establishes the confound that the reciprocal control is designed to suppress.

I.1 SHARED-RATE GAUGE TRAJECTORIES

Table 17: Shared-AdamW gauge trajectories for Qwen3-4B, seed 42. Validation accuracy and AUC are percentages. Actor KL values are reported in units of 10^{-3}.

Gauge c	Validation performance (%)					Mechanistic endpoints	
	Step 0	Step 2	Step 5	Step 10	$\text{AUC}_{0\text{–}10}$	$U_{0\to2}$	Actor KL@10 ($\times10^{-3}$)
$1/8$	55.47	57.03	62.11	69.53	62.03	0.02395	1.836
$1/4$	53.12	55.08	57.42	60.55	57.19	0.01203	1.347
1	53.91	53.52	53.91	58.59	54.98	0.00431	0.642
4	52.34	54.30	59.38	63.67	58.48	0.01385	1.229
8	51.56	57.42	62.89	69.53	62.05	0.02796	2.906

J OPTIMIZER GEOMETRY DIAGNOSTICS

Table 18: First-substep factor geometry and checkpoint adapter-product displacement under shared-rate AdamW. Gradient and update RMS values are reported in units of 10^{-6}; $U_{0\to2}$ is reported in units of 10^{-3}.

Gauge c	$\|A\|_F/\|B\|_F$	Gradient RMS ($\times10^{-6}$)		Update RMS ($\times10^{-6}$)		$U_{0\to2}$ ($\times10^{-3}$)
		A	B	A	B	
$1/8$	63.999	0.598	47.330	2.621	2.978	23.95
$1/4$	16.000	1.338	27.800	2.783	2.964	12.03
1	1.000	5.133	6.283	2.914	2.893	4.31
4	0.0625	25.680	2.047	2.975	2.754	13.85
8	0.0156	52.530	1.109	2.985	2.599	27.96

Under the scalar gauge, gradient RMS grows approximately in proportion to c for A and in proportion to $1/c$ for B, while Adam-normalized parameter update RMS remains at similar magnitudes across the two factors. This mismatch— opposite gradient scaling without a corresponding scaling of the parameter updates—causes each product-update term to be amplified by the norm of the other factor.

K LoRA-RITE Run-Level Results

Table 19 reports all 15 LoRA-RITE runs underlying the main summary. Five gauges are nested within each of three training seeds; the 15 rows are therefore not 15 independent replicates. The learning rate is fixed at 3×10^{-7} after balanced-gauge displacement matching and is not selected by accuracy.

Table 19: LoRA-RITE results on Qwen3-4B/GSM8K across three seeds and five iso-functional gauges.

Seed	Gauge c	$\text{AUC}_{0\text{–}10}$	$U_{0\to2}$	Actor KL@10
42	$1/8$	55.0195%	0.004516	0.000783
	$1/4$	56.2500%	0.004689	0.000712
	1	56.0547%	0.004614	0.000773
	4	54.3750%	0.004879	0.000628
	8	54.8242%	0.004924	0.000798
43	$1/8$	54.6875%	0.004674	0.000692
	$1/4$	56.1328%	0.004741	0.000853
	1	56.2109%	0.004601	0.000613
	4	55.9961%	0.004522	0.000673
	8	56.1523%	0.004644	0.000757
44	$1/8$	54.8633%	0.004904	0.000604
	$1/4$	56.3086%	0.004874	0.000697
	1	54.8242%	0.004967	0.000638
	4	56.1719%	0.004734	0.000792
	8	56.7383%	0.004937	0.000823

L Additional Optimizer Controls

L.1 LoRA+-Style Fixed-Ratio Control

Table 20: Gauge-independent fixed-ratio AdamW on Qwen3-4B/GSM8K, seed 42. This is a LoRA+-style structural control, not an official LoRA+ reproduction.

Gauge	$AUC_{0–10}$	Product displacement	Ratio / minimum	Actor KL@10
$c = 1/8$	83.55%	0.346864	12.16×	0.019136
$c = 1$	69.51%	0.050742	1.78×	0.006418
$c = 8$	64.45%	0.028535	1.00×	0.004653

L.2 Fixed-Probe Output-Space Measurements

Table 21: Step-0-to-2 product- and output-space movement over 32 prompts and 7,596 teacher-forced response tokens. Product displacement $U_{0\to2}$ is reported in units of 10^{-3}, and fixed-token KL in units of 10^{-4}. All output-space metrics use the same fixed probe.

Update rule	Gauge c	Product space	Fixed-probe output space				
		$U_{0\to2}$ ($\times10^{-3}$)	Logit RMS	Mean abs.	Max abs.	KL ($\times10^{-4}$)	KL / bal.
Shared AdamW	1	4.310	0.08542	0.05989	14.906	4.62	1.00×
	1/8	23.950	0.08785	0.06326	6.514	6.04	1.31×
	8	27.960	0.08819	0.06386	7.951	8.27	1.79×
Reciprocal LR	1/8	4.424	0.08602	0.06149	10.219	5.53	1.20×
	8	4.207	0.08771	0.06472	5.273	5.19	1.12×

L.3 Repeated Greedy AUC Calibration

Table 22: Repeated greedy evaluation of the Qwen3-4B seed-42 trajectory over updates 0–10. All values are validation-AUC percentages. Summary statistics use the three repeated evaluations; gaps compare each repeat mean with the balanced repeat mean.

Update rule	Gauge c	Validation AUC (%)				Repeat summary	
		Original	Repeat 1	Repeat 2	Repeat 3	Mean ± SD	Gap (pp)
Shared AdamW	1	54.98	55.18	54.28	54.57	54.67 ± 0.46	–
Reciprocal LR	1/8	53.96	56.09	55.78	55.27	55.72 ± 0.41	+1.04
	8	55.84	55.06	54.84	54.90	54.93 ± 0.11	+0.26

L.4 Optimizer Ablation

Table 23: Shared-rate optimizer ablation on moderate gauges $c \in \{1/4, 1, 4\}$ (Qwen3-4B, seed 42).

Optimizer	$U_{0\to2}$ max/min	AUC range	Conclusion
AdamW	3.21×	3.50 pp	gauge-sensitive
SGD	12.41×	17.91 pp	gauge-sensitive

M Repeated Greedy-Evaluation Diagnostics

Table 24 reports checkpoint-level variation across the three repeated evaluations. "Prediction none" counts responses for which the strict answer parser produced no prediction, and "limit hit" counts responses that reached the 512-token cap; both are counted as incorrect in accuracy. Correctness disagreement is the fraction of sample IDs whose binary correctness varies across the three evaluations.

Table 24: Checkpoint-level repeated greedy-evaluation diagnostics on Qwen3-4B, seed 42. Accuracy is reported as mean ± sample SD over three repeated evaluations. "No pred." and "Limit hit" are the mean numbers of affected responses; disagreement is the percentage of examples whose binary correctness differs across repeats.

Update rule	Gauge c	Step	Accuracy (%) mean ± SD	Generation diagnostics		
				No pred.	Limit hit	Disagree. (%)
Shared AdamW	1	0	51.30 ± 0.60	119.7	7.7	3.12
		2	52.86 ± 1.13	115.3	8.0	3.91
		5	54.69 ± 0.39	108.3	7.0	4.30
		10	57.81 ± 1.41	101.3	5.0	5.08
Reciprocal LR	$1/8$	0	52.21 ± 0.81	117.0	6.7	4.30
		2	54.30 ± 0.00	111.3	5.3	2.34
		5	56.25 ± 0.39	107.0	6.0	2.73
		10	57.68 ± 0.90	101.3	5.3	4.30
	8	0	51.56 ± 0.78	119.0	6.0	5.86
		2	54.30 ± 0.78	111.3	6.0	3.12
		5	54.95 ± 0.45	107.7	5.0	2.73
		10	56.90 ± 0.98	103.3	6.7	3.12

N Reproducibility Details

The following protocol connects the theoretical intervention to the saved adapters and optimizer state used in the experiments.

1. Pair all trainable LoRA A/B tensors and abort on missing or unexpected trainable adapter parameters.
2. Cast factors to FP32, compute reduced QR factorizations of B and $A^\top$, and take the compact SVD of $R_B R_A^\top \in \mathbb{R}^{r\times r}$.
3. Construct balanced factors, save their FP32 product, and apply the positive scalar gauge $(B_c, A_c) = (cB, A/c)$.
4. Cast to the source adapter dtype, serialize, reload, and audit relative adapter-product error and fixed-input full-vocabulary logits.
5. Create disjoint optimizer groups for all trainable A and B factors; abort if either group is empty.
6. Assign the main reciprocal rates $\eta_A = \eta/c$ and $\eta_B = c\eta$, and record group sizes and learning rates before the first optimizer step.
7. Only for the frozen one-step AdamW scalar-covariance diagnostic, additionally set $\lambda_A = c\lambda$, $\lambda_B = \lambda/c$, $\epsilon_A = c\epsilon$, and $\epsilon_B = \epsilon/c$.
8. Record model, task, seed, condition, parent-adapter identity, data order, optimizer settings, checkpoint schedule, audits, and per-run metrics in a structured record.

Table 4 is the canonical configuration table. The objective and task-specific limits appear in Appendix A; all deviations are stated in the corresponding experiment appendices.

O Coordinate Audits and Training Gates

These zero-training audits test whether proposed real-world contrasts isolate the scalar subgroup studied in the paper. They are training gates rather than performance experiments.

O.1 Native PiSSA Geometry

Across 144 rank-16 q/k/v/o LoRA modules in Qwen3-4B, native PiSSA factor-norm ratios lie in $[0.999819, 1.000181]$ and normalized Gram imbalance is at most 7.48×10^{-4}. The mean best

scalar pair error is 1.436, whereas the best orthogonal and general-R errors are both approximately 1.5×10^{-3}; the fitted general transform is nearly orthogonal. Reconstructing and re-serializing the canonical factors also introduces approximately 0.20% product difference. Native and canonical PiSSA therefore do not provide a strict same-function scalar-only contrast, so we do not train or interpret that comparison as scalar attribution evidence.

O.2 Naturally Imbalanced Parent Screening

We audit 19 available non-zero adapters using a prespecified global-scalar gate: absolute median log factor-norm ratio at least 0.5, layer-sign consistency at least 0.8, and a 95th-percentile residual after global-scalar centering at most one log-two unit. None passes all criteria. Transported adapters are strongly imbalanced but heterogeneous across layers, whereas the inspected SFT checkpoint is neither sufficiently imbalanced nor directionally consistent. This negative result motivates limiting our claim to actively imposed, controlled scalar gauges rather than asserting that realistic adapters are usually explained by one global scale.

P Robustness Checks and Scope

The auxiliary analyses sharpen the interpretation of the main result. Mixed accuracy effects show that balancing is an attribution control rather than a performance objective. Residual trajectory differences show that matching one aggregate displacement does not determine the full optimization path, and that LR-only reciprocal correction is approximate in the practical AdamW stack. Native PiSSA is nearly balanced, while the screened adapters exhibit richer layer-wise geometry rather than one coherent global scale. Accordingly, our claim concerns an actively imposed scalar gauge, not a general diagnosis of published initialization gains.

The intervention covers positive scalar gauges; general rank-space transformations require invariant optimizers. Long-horizon online evidence spans Qwen3 and one Llama-3.2 model, with the latter limited to DAPO-Math and the $c = 8$ direction. The scalar-covariance variant remains a one-step diagnostic, and product movement, fixed-token probes, and three-seed summaries are interpreted as complementary mechanistic evidence rather than a formal equivalence test. Both gauge directions share the balanced arm within a seed, so training seed remains the independent unit.