Auditing PEFT Claims: A Stability Probe Unifying Three LoRA-Variant Fragility Patterns

Suan Lee **JaeSeong Kim**
School of Computer Science, Semyung University, Jecheon, Republic of Korea
{suanlee, kjsqp1010}@semyung.ac.kr

Abstract

Most published PEFT improvements over LoRA do not survive paired multi-seed evaluation. We audit DoRA, PiSSA, AdaLoRA, and rsLoRA against LoRA across 16 claim cells spanning 110M–7B, using paired seeds, bootstrap confidence intervals, sign-flip tests, and Holm–Bonferroni correction with pre-committed adjudication rules. Only 2 of 15 confirmatory cells remain supported; 4 reverse. DoRA shows no supported gain in any of seven audited cells. PiSSA's +2.6pp GSM8K advantage under a common budget reverses under paper-faithful hyperparameters, including a claim-matched Gemma-7B setting ($\Delta = -1.58$pp, $p \approx 0.004$), and its NLI gain reverses under adversarial shift while improving on counterfactual sentiment. AdaLoRA's collapse on SQuAD is driven by training budget rather than its rank schedule. We then probe the effective adapter learning rate $\mathrm{lr} \times \alpha/\sqrt{r}$ as a pre-flight check. On a 303-run rsLoRA landscape (BERT-base/MNLI), no configuration below $\approx 5 \times 10^{-4}$ collapsed $(40/40)$ while 11 of 32 above it did, though all collapses occurred at the largest learning rate. DoRA's failures rise with α at high learning rate; PiSSA's reversal tracks learning rate but not α. The value is calibrated, not universal: Mistral-7B degrades below it, and a prospective LLaMA-2-7B test failed. We release all code and data and propose a minimal reporting standard for PEFT comparisons.

1 Introduction

TL;DR: Before training, compute $\mathrm{lr}_{\mathrm{eff}} = \mathrm{lr} \times \alpha/\sqrt{r}$. If $\mathrm{lr}_{\mathrm{eff}} > 5 \times 10^{-4}$ (our calibrated value on BERT-base/MNLI/AdamW; treat as an order-of-magnitude probe, not a universal constant—per-architecture calibration recommended), training is at risk of collapse. We measure this value on a 303-training landscape, audit 16 published PEFT claims on six models (110M–7B), and test the probe at 7B, where it partially replicates on Mistral and fails prospectively on LLaMA-2.

The PEFT landscape has expanded rapidly: DoRA (Liu et al., 2024) decomposes weights into magnitude and direction, PiSSA (Meng et al., 2024) initializes adapters from principal singular vectors, AdaLoRA (Zhang et al., 2023) allocates rank adaptively, and rsLoRA (Kalajdzievski, 2023) applies rank-stabilized scaling. Each reports gains over vanilla LoRA (Hu et al., 2022), typically a few points on selected benchmarks.

What we contribute. We identify a single quantity, the effective adapter learning rate $\mathrm{lr}_{\mathrm{eff}} \equiv \mathrm{lr} \times \alpha/\sqrt{r}$, that flags hyperparameter regimes in which LoRA-variant training can collapse. The audit produces evidence; the probe produces a conservative pre-flight check. Together they yield a reporting standard: practitioners can compute $\mathrm{lr}_{\mathrm{eff}}$ before training, and reviewers can ask for it to be reported.

Three sources of sensitivity. Multi-seed re-evaluation exposes three sources of sensitivity threatening published conclusions:

1. **Seed sensitivity.** Small reported gains may fall within seed variance, which single-run comparisons cannot detect; seed effects on fine-tuning are large even per example (Bui et al., 2025).

2. **Regime sensitivity.** Papers use different learning rates, scaling factors (α), and training budgets. A method that wins under one budget may lose under another.

3. **Distribution sensitivity.** PEFT comparisons are typically conducted on in-distribution benchmarks. Whether method rankings transfer out-of-distribution is unknown.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 474–494
November 6-10, 2026 ©2026 Association for Computational Linguistics

One candidate mechanism. For three of the four methods (rsLoRA, DoRA, PiSSA) we probe a single quantity, the *effective adapter learning rate* $\mathrm{lr}_{\mathrm{eff}} \equiv \mathrm{lr} \times \alpha/\sqrt{r}$. On AdamW/BERT-base/MNLI, all 11 collapsed rsLoRA cells have $\mathrm{lr}_{\mathrm{eff}} \geq 5\times10^{-4}$ and all 40 cells below are stable ($0/170$ runs versus $39/133$ above): crossing the value is necessary but not sufficient for collapse, every collapse occurred at lr$= 10^{-3}$, and at equal $\mathrm{lr}_{\mathrm{eff}}$ the outcome also depends on rank. At lr$= 10^{-3}$, DoRA at r=64, α/r=1 fails in $13/34$ seeds across two runs (38%, Wilson 95% CI [24%, 55%]), against $0/41$ at r=16 and none at lr$= 10^{-4}$. PiSSA's regime reversal tracks learning rate but not α. At 7B, naive rsLoRA on Mistral-7B degrades on $3/10$ seeds already at $\mathrm{lr}_{\mathrm{eff}}=4\times10^{-4}$, below the BERT-calibrated value.

Contributions. (1) A **stability probe** (§5): on a 303-training BERT-base/MNLI landscape no rsLoRA configuration below $\mathrm{lr}_{\mathrm{eff}}=5\times10^{-4}$ collapsed ($40/40$) while $11/32$ above did; DoRA's α-driven failure is consistent with the probe and PiSSA's regime reversal with its learning-rate component (the first two are collapses, the third a ranking reversal); AdaLoRA's degeneracy is a complementary pattern that a 2×2 control attributes to training budget rather than to schedule or $\mathrm{lr}_{\mathrm{eff}}$. (2) A **16-cell audit with pre-committed adjudication rules** (§4) across six models (110M–7B) with bootstrap CIs, paired sign-flip, Holm correction: DoRA $0/7$ supported, PiSSA reversed under paper-faithful, AdaLoRA reversed on both SQuADs. (3) A **7B confirmatory test** (§7) on Gemma-7B, LLaMA-2-7B, and Mistral-7B; the Gemma claim-matched reversal ($\Delta = -1.58$, two-sided $p \approx 0.004$) is the only 7B result to survive Holm. (4) An **independent exploratory distribution-shift finding** (§6): PiSSA hurts on adversarial NLI but improves on counterfactual sentiment—ID-only audits miss this. We flag explicitly that this finding is *not* explained by $\mathrm{lr}_{\mathrm{eff}}$ (which is constant across the paired arms) and present it as complementary evidence, not as a third instance of the mechanism. (5) A **minimal reporting standard**: ≥ 10 paired seeds, bootstrap CI, regime ablation, $\alpha/\sqrt{r}$ disclosure, plus a one-line pre-training check on $\mathrm{lr}_{\mathrm{eff}}$. All configs, seeds, and scripts are released.

2 Related Work

Reproducibility audits in ML. Our work belongs to a recognized tradition of *rigor audits* that retroactively test whether published comparative claims survive multi-seed re-evaluation. Henderson et al. (2018) showed that different random seeds of the same deep-RL algorithm can yield statistically different results; Lucic et al. (2018) found that GAN variants thought to dominate vanilla GAN failed to do so under fair hyperparameter budgets; Bouthillier et al. (2019) argued that reproducible methods still yield unreproducible findings unless sources of variation are accounted for; Pineau et al. (2021) reported on the NeurIPS 2019 reproducibility program. We extend this tradition to PEFT, where adoption has outpaced rigorous evaluation.

PEFT benchmarks and LoRA variants. No prior work systematically audits paired multi-seed PEFT claims at this scale. Beyond DoRA, PiSSA, AdaLoRA, and rsLoRA (the four we audit), related methods include the LoRA variants VeRA (Kopiczko et al., 2024), LoRA+ (Hayou et al., 2024), KronA (Edalati et al., 2022), and SVFT (Lingam et al., 2024), and other approaches such as IA^3 (Liu et al., 2022) (activation scaling) and GaLore (Zhao et al., 2024) (gradient projection). We audit four widely used LoRA variants whose papers make directional claims over LoRA with configurations sufficient for paired replication; the others are left to future work (Limitations).

Seed sensitivity in NLP. Bui et al. (2025) show that fine-tuned RoBERTa-large and LoRA-tuned Llama3.2-3B are seed-sensitive even per example: pairwise prediction consistency across ten seeds ranges from 64.9% (COPA) to 96.8% (SST-2) for RoBERTa-large. Xue et al. (2023) argue that reproducibility is dominated by the signal-to-noise ratio of performance estimators. Neither work specifically audits PEFT *variant* comparisons.

Statistical methodology. Paired bootstrap CIs (Efron, 1987) and sign-flip permutation tests are standard for ML model comparison; we adopt these and extend them to PEFT-specific auditing with frozen adjudication rules and Holm–Bonferroni multiplicity correction over a confirmatory family.

Adapter-space analysis. Recent work studies LoRA weight spaces directly: Learning on LoRAs (Putterman et al., 2024) predicts global adapter properties from LoRA weights. Such analyses examine adapter structure but do not audit the comparative claims made by PEFT method papers.

OOD robustness of PEFT. Shuttleworth et al. (2025) show that LoRA and full fine-tuning produce structurally different weight updates ("intruder dimensions") that drive forgetting of the pre-training distribution. Biderman et al. (2024) further demonstrate that LoRA learns less but also forgets less than full fine-tuning, with implications for robustness. However, systematic OOD comparisons *across* PEFT variants are absent from the literature.

3 Method: Multi-Seed Sensitivity Analysis Protocol

3.1 Claim Cell Definition

A *claim cell* is defined by a tuple: (source paper, method comparison, model, task, metric). Only cells meeting all inclusion criteria—clear directional claim, reproducible setup, feasible on our hardware—are admitted to the confirmatory set.

3.2 Experimental Regimes

Common-budget regime. Standardized across methods: lr=2×10^{-4}, rank=8, $\alpha = 16$, dropout=0.05, matched epoch budget.

Paper-faithful regime. Follows the original paper's documented hyperparameters as closely as hardware constraints permit: lr=2×10^{-5}, $\alpha = r$, dropout=0, cosine scheduler (Meng et al., 2024).

3.3 Study Design Transparency

Our study is designed as a sensitivity analysis with frozen adjudication rules, rather than a strict clinical-style pre-registration. The following elements were frozen before confirmatory runs: claim-cell inclusion rules, the paired-seed protocol, bootstrap confidence intervals (BC was specified; see §3 for the deviation), sign-flip permutation tests, and the supported/unsupported/reversed adjudication labels. The following elements were chosen after calibration: which claim families to prioritize, which larger models were feasible to run, and which paper-faithful settings were practical under our hardware constraints.

Cells 1–14 and the statistical protocol were frozen prior to confirmatory runs; Cell 15 is exploratory and excluded from the Holm family. Cell 16 (Gemma-7B claim-matched) was added when A6000 capacity became available; its Holm-family inclusion and the threshold recomputation were documented in the pre-registration document before chain launch (file and SHA-256 hash in Appendix N). *Verifiability limitation:* this is an internal version-control timestamp, not a third-party attestation (OSF/public-log); Cells 17–18 (LLaMA-2-7B/commonsense_170k; §7.2, Appendix K) are outside the Holm family. We describe the work as a *multi-seed sensitivity analysis* with frozen adjudication, not a fully locked confirmatory study.

3.4 Statistical Protocol

All comparisons use *paired seeds*: the same seed list is applied to both baseline (LoRA) and candidate method. For each cell:

- Per-seed delta: $\delta_i = m(\text{candidate}, s_i) - m(\text{LoRA}, s_i)$
- Mean delta $\bar{\delta}$ with bootstrap 95% CI (B=5000). We use the bias-corrected (BC, not BCa) bootstrap for Table 1 row 16 and for the Tier-2 extension (Appendix G); the other rows were frozen using the percentile bootstrap built into our sweep infrastructure at the time, a deviation from the specified BC interval. Table 1 caption labels each row's method. On the cells we computed with both estimators (rows 3 and 4), CI endpoints agreed with the percentile bootstrap to within reporting precision, and no adjudication changed. A paired-t sensitivity analysis (Appendix M, Table 15) confirms direction stability across all confirmatory cells.
- Sign-flip permutation p-value. All Table 1 p-values are two-sided, which is conservative for the claimed direction; only the Tier-2 ID metrics use a one-sided test (Appendix G).
- Adjudication: **supported** (CI > 0), **unsupported** (CI $\ni 0$), **reversed** (CI < 0), frozen before any confirmatory run. A supported or reversed label also requires the sign-flip p to pass Holm correction within the confirmatory family (exploratory Cell 15: unadjusted p<0.05); otherwise the cell is unsupported. W/L/T counts the seeds on which the candidate wins, loses, or ties against LoRA.
- Holm–Bonferroni correction across all confirmatory cells in the family to control family-wise error rate at $\alpha = 0.05$.

Failure handling. Failed seeds (SSH timeout, OOM) are dropped symmetrically from both arms to preserve pairing; no imputation. Cell

5 (AdaLoRA/SQuAD1) has N=9 due to one remote failure, and Cells 12, 14 and 15 have N=8 after dropped seeds; all other cells have $N \geq 10$. *Pre-specified NaN-metric rule:* if a seed produces a `NaN` eval-metric in either arm, that seed pair is dropped symmetrically; if dropping leaves N<5 the cell is "inconclusive (NaN-truncated)". This rule is invoked once (Tier-2 ANLI R1, Appendix G).

Assumption defenses. (i) Two-sided tests are used for every audit cell, so a reversal is never declared from a one-sided test. (ii) The BC bootstrap corrects the median bias that the percentile interval ignores (Efron, 1987); both may undercover at N=8–20 (Limitations). (iii) The sign-flip permutation is an exact null distribution under paired exchangeability, which holds here because data, optimizer, schedule, and dtype are matched within each cell—the only intervention is the PEFT method.

4 Multi-Seed Sensitivity Results

4.1 Setup

We audit four methods—DoRA, PiSSA, AdaLoRA, rsLoRA—each against LoRA. Models span five scales: roberta-base (125M) (Liu et al., 2019), bert-base (110M) (Devlin et al., 2019), phi-2 (2.7B) (Javaheripi and Bubeck, 2023), Qwen2.5-3B (3.1B) (Yang et al., 2024), and Mistral-7B (7.2B) (Jiang et al., 2023). The 7B chain (§7) additionally uses Gemma-7B (Gemma Team, 2024) and LLaMA-2-7B (Touvron et al., 2023). Tasks include BoolQ (Clark et al., 2019), MNLI (Williams et al., 2018), SQuAD v1/v2 (Rajpurkar et al., 2016, 2018), GSM8K (Cobbe et al., 2021), and MetaMathQA (Yu et al., 2024). All experiments use the HuggingFace `transformers` library (Wolf et al., 2020) and the `peft` library (Mangrulkar et al., 2022) for adapter configurations. Each cell uses 8–20 paired seeds. The ≤3B experiments run on a 50×RTX 3060 cluster; the 7B experiments run on two RTX A6000 (48GB) GPUs, one job per GPU. The 16 audit cells alone comprise 474 completed runs (237 seed pairs); the full study exceeds 1,000.

Cell selection criteria. Each cell tests the direction of a method's published advantage over LoRA on our models and tasks, with the source paper's hyperparameters where hardware allows (DoRA Cell 12; PiSSA Cells 9, 15, 16) and a common budget otherwise; cells span six models (110M→7B) and four task families (NLI, QA, reasoning, commonsense). Cells outside these criteria are exploratory (Cell 15) and excluded from the Holm family.

Important scope note: due to hardware constraints (12GB VRAM per node), most cells use *subsampled* training and evaluation sets rather than full datasets. Typical subset sizes are: BoolQ 7k train / 1.5k eval; MNLI 7k / 1.5k; SQuAD 5k / 1k; GSM8K 3k train / 300 eval. The 7B cells use their own subsets (Mistral MNLI 5k/1k, Gemma 50K MetaMathQA, LLaMA-2 25K commonsense_170k). Cell 15 (claim-matched) uses 5k MetaMathQA samples. These subsets are sufficient for paired comparisons but smaller than typical production training budgets. Full subset sizes per cell are documented in the released configs.

4.2 Results

DoRA gains over LoRA do not survive multi-seed re-evaluation. In seven paired audit cells (Cells 1, 3, 8, 12, 14, plus Mistral-7B/MNLI in Appendix L and the bf16 LLaMA-2-7B cell in Appendix K) spanning four model scales (125M, 3.1B, 6.7B, 7.2B), four tasks (BoolQ, MNLI, GSM8K, commonsense_170k), and ranks $r \in \{8, 32, 64\}$, the only nominally significant DoRA effect is Mistral-7B/MNLI (Δ=+0.0038, raw p=0.047, $|\Delta| < 0.4\%$); this cell falls outside the 15-cell confirmatory Holm family and is reported in Appendix L as null after multiplicity correction. All five Table 1 DoRA cells are unsupported pre-correction. The audit's frozen adjudication rules return DoRA 0/7 supported; this does not preclude DoRA advantages outside our audit scope.

PiSSA gains are scale- and regime-dependent within the audited settings. PiSSA outperforms LoRA under common budgets at small (≤3B) scale (cells 2, 4, 7); the advantage vanishes on phi-2 (cell 10: $\Delta = -0.26$, unsup.) and directionally reverses on Mistral-7B/GSM8K (cell 13: $\Delta = -1.90$, p=0.008, 0W/8L/2T) even under the common budget, though not significantly after Holm. Under paper-faithful hyperparameters on Qwen3B (cell 9: r=8, α=8, lr=2×10^{-5}, N=19, our GSM8K subset), the comparison reverses ($\Delta = -1.30$, p=0.002). Cell 15 is the more stringent *claim-matched* variant of Cell 9: same hyperparameters *plus* PiSSA's data pipeline (MetaMathQA → full GSM8K 1319-test), subsampled to 5K training examples. Cell 15 reverses in direction ($\Delta = -1.29$,

Table 1: Claim audit results across 16 cells. Δ: mean paired difference (candidate − LoRA). CI: BC bootstrap 95% interval for row 16; percentile bootstrap (frozen at confirmatory-run time) for all other rows; rows 3 and 4, re-run with BC, agreed to reporting precision. p: two-sided sign-flip. Adj.: adjudication after Holm–Bonferroni correction (α=0.05, family size =15 confirmatory cells; row 15 exploratory). Highlighted: row 9 (yellow), the paper-faithful reversal at 3B; row 15 (blue), the exploratory claim-matched variant; row 16 (green), the Gemma-7B claim-matched reversal that mirrors row 9.

#	Model	Method	Task	Regime	N	Δ	95% CI	p	Adj.
1	roberta	DoRA	BoolQ	common	20	$+.005$	$[-.001, +.012]$	.151	unsup.
2	roberta	PiSSA	BoolQ	common	20	$+.011$	$[+.003, +.019]$	.011	unsup.[†]
3	roberta	DoRA	MNLI	common	20	$+.001$	$[-.001, +.003]$	.310	unsup.
4	roberta	PiSSA	MNLI	common	20	$+.009$	$[+.006, +.012]$	$< .001$	**sup.**
5	bert	AdaLoRA	SQuAD1	common	9	-36.7	$[-38.0, -35.4]$	.004	**rev.**
6	roberta	rsLoRA	BoolQ	common	17	$+.002$	$[-.020, +.023]$	.833	unsup.
7	Qwen3B	PiSSA	GSM8K	common	20	$+2.62$	$[+1.67, +3.55]$	$< .001$	**sup.**
8	Qwen3B	DoRA	GSM8K	common	20	-0.05	$[-0.73, +0.60]$	.887	unsup.
9	Qwen3B	PiSSA	GSM8K	faithful[*]	19	-1.30	$[-2.02, -0.61]$	.002	**rev.**
10	phi-2	PiSSA	GSM8K	common	18	-0.26	$[-1.39, +0.85]$	.632	unsup.
11	bert	AdaLoRA	SQuAD2	common	10	-3.02	$[-3.33, -2.67]$	.002	**rev.**
12	Qwen3B	DoRA	GSM8K	r=32	8	$+0.04$	$[-0.63, +0.71]$	.89	unsup.
13	Mistral-7B	PiSSA	GSM8K	common	10	-1.90	$[-2.97, -0.90]$	.008	unsup.[¶]
14	Mistral-7B	DoRA	GSM8K	common	8	$+0.33$	$[-0.77, +1.22]$	.55	unsup.
15	Qwen3B	PiSSA	GSM8K[‡]	faithful	8	-1.29	$[-2.38, -0.17]$	.10	unsup.[§]
16	**Gemma-7B**	**PiSSA**	**GSM8K**[‡]	**faithful**	**10**	$\mathbf{-1.58}$	$\mathbf{[-2.33, -0.89]}$	**.004**	**rev.**

[†]PiSSA/BoolQ is directionally consistent (15W/4L) but does not survive multiplicity correction. [*]Precision-controlled: both LoRA and PiSSA in FP16 with gradient checkpointing, removing the quantization confound. [¶]CI excludes zero and the direction is reversed (0W/8L/2T), but raw p=.008 misses its Holm threshold (.00556); all "Adj." labels are post-Holm, so Cell 13 is unsupported here although Fig. 1 colors it red by direction. [§]CI excludes zero but p=.10; exploratory and unsupported, though its direction (2W/6L) matches Cells 9 and 13. [‡]Trained on MetaMathQA (the PiSSA paper's data pipeline), evaluated on GSM8K.

2W/6L) but p=0.10 does not survive correction at N=8 (exploratory, unadjusted); the directional consistency across Cells 9, 13, 15 is the substantive finding—PiSSA's ranking against LoRA depends on model scale and hyperparameter regime, and persists even when the original data pipeline is matched.

AdaLoRA reverses on both SQuAD versions in our audit. On SQuAD v1, AdaLoRA's adaptive rank pruning converges to a degenerate solution (LoRA F1 = 52.34, AdaLoRA F1 = 15.59, $\Delta = -36.75$, N=9). We use the AdaLoRA implementation in the `peft` library (the most widely adopted one), so the degeneracy is a property of the standard implementation under our budget rather than a re-implementation artifact: AdaLoRA's importance-score allocator removes informative directions before convergence given the subsampled SQuAD v1 budget. A 2×2 control over {budget}×{schedule} (Appendix J) identifies the cause, and it is *not* the schedule: varying the requested `tinit`/`tfinal` by 5× shifts AdaLoRA's F1 by less than 0.05 at either budget, whereas raising the training budget from 5K to 12K samples moves it from 15.53 to 55.38 F1 at the Cell 5 schedule (same software stack). AdaLoRA is markedly more data-hungry than LoRA here (LoRA gains +12.1 F1 over the same budget change, AdaLoRA +39.8). At the larger budget the reversal persists but shrinks to $\Delta = -10.09$ F1 (BC 95% CI $[-10.74, -9.40]$, $p = 0.002$, 0/10/0), so we report ≈ -10 as the budget-fair estimate and treat Cell 5 as a low-budget-regime result.

Table 2: AdaLoRA mean F1 on SQuAD v1 over {train budget}×{rank-allocation schedule}, 10 paired seeds per cell on one software stack (the original Cell 5 run gave 15.59). Schedule moves F1 by <0.05 in both budget regimes; budget moves it by ≈40. Cell 5's dramatic gap is a low-budget effect, not a schedule artifact (Appendix J).

	`tinit`=50/`tfinal`=200	250/750
5K train	15.53 (Cell 5 config)	15.53
12K train	**55.38**	**55.40**

On SQuAD v2—the paper's target task—the reversal *replicates* at much smaller magnitude ($\Delta = -3.02$, 0W/10L, $p = 0.002$), confirming SQuAD v1 is not an isolated artifact. The $\mathrm{lr}_{\mathrm{eff}}$ mechanism does not apply here because AdaLoRA's effective rank $r(t)$ varies during training (the alloca-

tor prunes triplets), making $\alpha/\sqrt{r(t)}$ a trajectory rather than a constant; the failure is budget-driven (Appendix J), not $\mathrm{lr}_{\mathrm{eff}}$-driven, and we treat it as a complementary fragility class.

Multiple testing. Table 1 contains 16 rows: 15 confirmatory cells (Cells 1–14 and Cell 16, the latter the Gemma-7B claim-matched cell, whose Holm-family inclusion was committed before the A6000 chain launched) and 1 exploratory cell (Cell 15, $N{=}8$). The Holm–Bonferroni family comprises the 15 confirmatory cells; Cell 15 is reported for directional context but excluded from the correction family. After correction across 15 cells, 6 of 8 non-null results survive: PiSSA/MNLI (Cell 4), AdaLoRA/SQuAD1 (Cell 5), PiSSA/Qwen3B-common (Cell 7), PiSSA/Qwen3B-faithful (Cell 9), AdaLoRA/SQuAD2 (Cell 11), and **PiSSA/Gemma-7B-faithful (Cell 16, the 7B claim-matched reversal)**. The two casualties are PiSSA/BoolQ (Cell 2, raw $p = 0.011$, Holm threshold $= 0.00625$) and PiSSA/Mistral-7B (Cell 13, raw $p = 0.008$, Holm threshold $= 0.00556$). The Holm thresholds are computed by a released script (Appendix N). PiSSA/BoolQ is **reclassified as unsupported** after correction; PiSSA/Mistral-7B retains a directionally-consistent reversal (0W/8L/2T) but does not survive multiplicity correction at this seed count. All "unsupported" adjudications are unaffected. Under the less conservative Benjamini–Hochberg FDR (at $q{=}0.05$), all 8 non-null cells survive—our headline conclusions are robust to the choice of multiplicity correction.

5 Stability Boundary: A Mechanism for PEFT Fragility

The audit catalogs *what* fails; we now probe *why* with one candidate quantity for three of the four fragility patterns (rsLoRA collapse, DoRA α-driven failure, PiSSA regime reversal): the *effective adapter learning rate* $\mathrm{lr}_{\mathrm{eff}} \equiv \mathrm{lr} \times \alpha/\sqrt{r}$. The $\alpha/\sqrt{r}$ factor is theory-motivated: rsLoRA (Kalajdzievski, 2023) derives, from an infinite-width learning-trajectory analysis, that adapters need scaling $\alpha/\sqrt{r}$ to learn stably at any rank. We apply this convention to all methods as a rank-normalised probe, although LoRA, DoRA and PiSSA scale updates by α/r internally. Our contribution is empirical: above a critical product, training can collapse abruptly rather than degrade gracefully. We test this with a 303-training landscape on BERT-base/MNLI, an α-sweep for DoRA, and a 7B replication on Mistral.

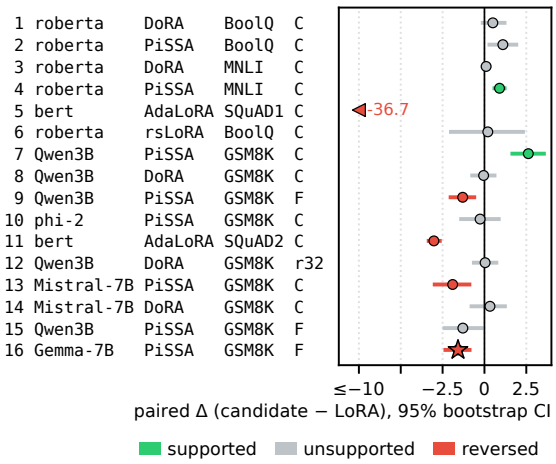


Figure 1: Paired Δ (candidate − LoRA; pp or F1) with 95% bootstrap CIs (BC or percentile, as labelled in Table 1) for all 16 cells (rows: cell, model, method, task, regime; C/F = common-budget/paper-faithful). Color follows Table 1's adjudication (green = supported, gray = unsupported, red = reversed) except Cell 13 (PiSSA/Mistral-7B), shown red by direction although its raw $p{=}0.008$ misses its Holm threshold (0.00556); it is the sole difference between the figure's 2/9/5 and the table's 2/10/4. Star: Cell 16, the Gemma-7B claim-matched reversal ($\Delta{=}-1.58$). Cell 5 ($\Delta{=}-36.7$) is clipped at the left edge.

5.1 rsLoRA: $\mathrm{lr}_{\mathrm{eff}}$ phase transition (BERT-base, MNLI, 303 trainings)

We sweep $r{\in}\{8,16,32,64\} \times \alpha{\in}\{1,2,4,8,16,32\} \times \mathrm{lr}{\in}\{10^{-4}, 3{\times}10^{-4}, 10^{-3}\}$ on BERT-base/MNLI 50K with up to 5 seeds per cell (303 completed trainings across 72 cells; no systematic bias in scheduler dropouts). A run is *stable* if dev accuracy exceeds the majority-class baseline (∼0.34), else *collapsed*; a cell is collapsed if any of its runs collapses. *(i) Learning rate matters, not $\alpha/\sqrt{r}$ alone*: at $\alpha/\sqrt{r}{=}1.0$ the same (r,α) is stable at $\mathrm{lr}{=}3{\times}10^{-4}$ but collapses at $\mathrm{lr}{=}10^{-3}$. All 11 collapses occur at $\mathrm{lr}{=}\ 10^{-3}$; at $\mathrm{lr}{\le}3{\times}10^{-4}$ every cell is stable, even at $\mathrm{lr}_{\mathrm{eff}}{=}3.4{\times}10^{-3}$, and at $\mathrm{lr}{=}\ 10^{-3}$ cells with equal $\mathrm{lr}_{\mathrm{eff}}$ can differ by rank (e.g., $(r,\alpha){=}(16,2)$ stable, $(64,4)$ collapsed). The product alone therefore does not determine collapse; $5{\times}10^{-4}$ is the lowest $\mathrm{lr}_{\mathrm{eff}}$ at which we observed one. *(ii) Necessary, not sufficient, for collapse*: all 11 collapsed cells have $\mathrm{lr}_{\mathrm{eff}}{\ge}5{\times}10^{-4}$ and $40/40$ cells below are stable, but $21/32$ above remain stable—the value flags

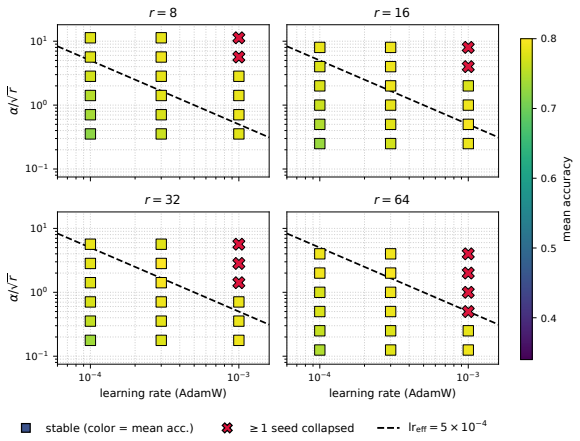


Figure 2: rsLoRA stability landscape, one panel per rank (BERT-base, MNLI 50K; n=303 trainings over 72 (r, α, lr) cells, up to 5 seeds each). Squares: stable cells colored by mean accuracy; red X: at least one seed collapsed to the majority class. Dashed line: $\mathrm{lr}_{\mathrm{eff}} \equiv \mathrm{lr} \times \alpha/\sqrt{r} = 5{\times}10^{-4}$. All 11 collapsed cells lie on or above it and all 40 cells below are stable, but 21 cells above are stable too; every collapse occurs at lr$= 10^{-3}$, where the critical $\alpha/\sqrt{r}$ falls as r grows.

Table 3: DoRA failure rate vs. α/r at r=64 (BERT-base, MNLI, AdamW, lr$= 10^{-3}$). Wilson 95% confidence intervals on the binomial failure proportion; mean accuracy reported on *successful* runs only.

α	α/r	n	failures	failure rate (Wilson 95% CI)	mean acc (success)
32	0.50	10	0	0.0% ([0.0%, 27.8%])	0.7879
40	0.625	6	0	0.0% ([0.0%, 39.0%])	0.7872
48	0.75	9	0	0.0% ([0.0%, 29.9%])	0.7820
56	0.875	8	1	12.5% ([2.2%, 47.1%])	0.7770
64	**1.00**	9	**2**	**22.2%** ([**6.3%**, **54.7%**])	0.7776
72	1.125	10	3	30.0% ([10.8%, 60.3%])	0.7727
80	1.25	6	1	16.7% ([3.0%, 56.4%])	0.7630
96	1.50	9	9	100.0% ([70.1%, 100.0%])	—

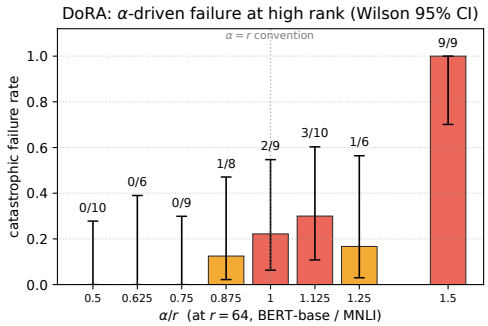


Figure 3: DoRA catastrophic-failure rate vs. α/r (r=64, BERT-base, MNLI, lr$= 10^{-3}$) with Wilson 95% CIs; labels give failures/seeds. Dashed line: the common $\alpha = r$ convention. Failures rise from 0% at $\alpha/r \leq 0.75$ to 22.2% at 1.0 and 100% (9/9) at 1.5; the dip at 1.25 (1/6) lies within overlapping intervals.

a risk regime, not a guarantee. *(iii) Seed level*: no run below the boundary collapses (0/170) versus 39/133 above it; 8 of the 11 collapsed cells collapse on every completed seed, while three $((r, \alpha){=}(8, 16), (32, 8), (64, 4)$ at lr$= 10^{-3}$) collapse on 2 of 4–5 seeds.

This product reframes the rsLoRA scaling rule as preserving $\mathrm{lr}_{\mathrm{eff}}$: a user who tunes lr at a LoRA configuration (r_0, α_0) and then switches to rsLoRA must rescale either lr or α to keep $\mathrm{lr}_{\mathrm{eff}} {<} 5 \times 10^{-4}$; otherwise training is at risk of collapse. This value is specific to BERT-base/MNLI/AdamW (§5.4).

Cross-architecture replication. On BoolQ/RoBERTa-base, rsLoRA at $r{=}32, \alpha{=}64, \mathrm{lr}{=}2 \times 10^{-4}$ ($\mathrm{lr}_{\mathrm{eff}}{=}2.26 \times 10^{-3}$) puts all 10 seeds of a complete re-run at exactly 0.6307, the majority-class baseline (std $=$ 0), while $r{=}8, \alpha{=}16$ ($\mathrm{lr}_{\mathrm{eff}}{=}1.13 \times 10^{-3}$, 17 seeds) trains normally (mean 0.712). Collapse thus reproduces on a different architecture and task, again in only part of the region above the boundary.

5.2 DoRA: α-driven failure rate at high rank

DoRA reports gains over LoRA at every rank it tests (Liu et al., 2024). A preliminary 19-seed run of ours at $r{=}16$ ($\alpha{=}r$) showed one collapse that did not recur in a 41-seed re-run (0/41), so we probe high rank: DoRA at $r{=}64$ with 6–10 seeds per $\alpha \in \{32, 40, 48, 56, 64, 72, 80, 96\}$ on BERT-base/MNLI at lr$= 10^{-3}$, applying the stability criterion of §5.1.

DoRA's failure rate rises with α/r (equivalently with $\mathrm{lr}_{\mathrm{eff}}$ at fixed lr, r): 0/25 failures at $\alpha/r{\leq}0.75$, rising to 9/9 at $\alpha/r{=}1.5$ (Cochran–Armitage trend test over all eight α values, $z{=}5.25$, $p{=}1.5{\times}10^{-7}$). The rise is not strictly monotone (1/6 at $\alpha/r{=}1.25$ vs. 3/10 at 1.125; overlapping CIs). At the standard convention $\alpha{=}r$ the sweep gives 2/9, and an independent $\alpha{=}r$ rank grid at the same lr gives 11/25 at $r{=}64$ but 0 failures at every $r{\leq}32$ (0/41 at $r{=}16$); pooled, $\alpha{=}r$ at $r{=}64$ fails in 13/34 seeds (38%, Wilson 95% CI [24%, 55%]; $p_{\mathrm{Fisher}}{=}7{\times}10^{-6}$ against $r{=}16$). At this lr every α in the sweep lies above the boundary ($\mathrm{lr}_{\mathrm{eff}}{=}4$–$12{\times}10^{-3}$), so the stable low-$\alpha$ cells are further instances of "necessary, not sufficient"; at lr$= 10^{-4}$ the same setup is stable for all $\alpha{\leq}72$ (0/24). We therefore claim only that at high lr failure risk rises with α/r, consistent with $\mathrm{lr}_{\mathrm{eff}}$, and that the standard $\alpha{=}r$ convention already fails at $r{=}64$.

5.3 PiSSA: regime sensitivity tracks learning rate

PiSSA's reversal between common and paper-faithful regimes (§4, Cells 7 vs. 9) straddles the boundary under the $\alpha/\sqrt{r}$ convention (under PiSSA's own α/r scaling both regimes lie below it). Common budget on Qwen2.5-3B has r=8, α=16, lr=2×10^{-4} giving $\mathrm{lr}_{\mathrm{eff}}$=$1.13 \times 10^{-3}$ (*above* the boundary); paper-faithful sets r=8, α=8, lr=2×10^{-5} giving $\mathrm{lr}_{\mathrm{eff}}$=$5.66 \times 10^{-5}$ (*below* the boundary). One interpretation is that PiSSA's SVD-initialized adapters help most when the learning rate is high (large updates make informed init worth more than random init), and that at low learning rates this advantage no longer compensates. The 2^2 factorial ablation (Appendix B) attributes the reversal to learning rate ($\eta^2 \approx 0.44$) but not to α (p=0.94; halving α even enlarges PiSSA's advantage), so the reversal is consistent with $\mathrm{lr}_{\mathrm{eff}}$ only through its learning-rate component. Unlike rsLoRA/DoRA collapse, PiSSA is a *ranking-reversal* case.

5.4 7B test of the boundary on Mistral-7B / MNLI

We test whether the threshold transfers to 7B scale. Using Mistral-7B/MNLI matched at r=64, we run two settings:

- **Naive transfer** (α=32, $\alpha/\sqrt{r}$=4.0, $\mathrm{lr}_{\mathrm{eff}}$=$4 \times 10^{-4}$): rsLoRA $2/10$ strict collapse to majority class (acc $\sim$ 0.33) plus $1/10$ partial degradation (acc $\sim$ 0.59), while paired LoRA is $10/10$ stable; paired $\Delta = -0.131$ (95% CI $[-0.270, +0.004]$, sign-flip $p = 0.24$; binary strict-collapse 2/10, Wilson 95% CI $[5.7\%, 51.0\%]$). Per-seed counts are released (Appendix N).

- **Re-tuned** (α=4, $\alpha/\sqrt{r}$=0.5, $\mathrm{lr}_{\mathrm{eff}}$=$5 \times 10^{-5}$): $10/10$ stable, mean acc 0.8714; this scale equals the LoRA arm's α/r=0.5, so the runs match LoRA exactly ($\Delta = 0.000$) by construction.

Naive transfer degrades at $\mathrm{lr}_{\mathrm{eff}}$=$4 \times 10^{-4}$, *below* the BERT-calibrated value, and the paired effect is not significant; collapse under naive transfer replicates at 7B, but the threshold value does not.

A compact method-$\times$-architecture stability summary is provided in Appendix I.

6 Exploratory Follow-Up: PEFT Methods Under Distribution Shift

To probe whether ID rankings transfer across shift types, we train LoRA, DoRA, and PiSSA on MNLI (ID) and SST-2 (Socher et al., 2013) (ID) using 19–20 paired seeds on roberta-base and evaluate on SNLI (Bowman et al., 2015) and three pre-registered shift types: adversarial OOD (ANLI R1/R2/R3; Nie et al., 2020), domain OOD (Yelp; Zhang et al., 2015), and counterfactual OOD (Amazon Counterfactual; O'Neill et al., 2021), chosen prior to any OOD experiment to span distinct shift mechanisms (heuristic-overlap OOD/HANS is a natural follow-up). **PiSSA's OOD behavior is shift-type-dependent**: it *hurts* on adversarial NLI (ANLI R2: $\Delta = -0.010$, p=2×10^{-4}; ANLI R1: $\Delta = -0.008$, p=0.017 at N=18 after NaN-rule drop; ANLI R3 unsupported—Appendix G) but improves on counterfactual sentiment (Amazon CF: $\Delta = +0.025$, $p < 10^{-5}$, 19W/0L/0T). The full Δ matrix across both task families and three shift types is reproduced in Appendix G (Tier-2 extension); the directional claim survives BC bootstrap and sign-flip testing (two-sided for OOD metrics).

Scope of this section, and its relation to $\mathrm{lr}_{\mathrm{eff}}$. We report PiSSA rather than DoRA here because DoRA has *no* significant ID gain in this task family to transfer (Appendix G: DoRA/MNLI-ID p=0.066), so an ID$\rightarrow$OOD transfer question is only well-posed for PiSSA. We state plainly that $\mathrm{lr}_{\mathrm{eff}}$ **does not explain this result**: both arms share (r, α, lr) and therefore the same $\mathrm{lr}_{\mathrm{eff}}$, so a quantity that is constant across the comparison cannot account for a shift-type-dependent sign change. We therefore present §6 as an *independent* exploratory finding—evidence that ID-only audits are insufficient—rather than as a third instance of the mechanism. The effect sizes are small in absolute terms (-0.010 on ANLI R2, $+0.025$ on Amazon-CF) but are estimated at N=19 paired seeds with 4W/15L and 19W/0L splits respectively, i.e. the *sign* is stable even though the magnitude is modest.

7 7B Confirmatory Experiments

We extend the audit to Gemma-7B and test the boundary prospectively on LLaMA-2-7B (Mistral-7B: §5.4). Each training job runs on one A6000 (48GB), with paired seeds spread over two GPUs, in bf16 for both arms (fp16 was numerically unstable; Appendices H and K).

7.1 Gemma-7B / MetaMathQA → GSM8K (claim-matched PiSSA)

We follow PiSSA's published adapter setup (r=128, α=128, lr=2×10^{-5}, seq 512, 1 epoch), with batch 2 instead of 128 and 50K MetaMathQA examples instead of 100K; evaluation on GSM8K (1319 test). Both arms run bfloat16 (fp16 induces grad-norm = NaN for PiSSA on Gemma; Appendix H). 10 paired seeds per arm.

Headline: claim reverses under paired multi-seed evaluation. Across 10 paired seeds, PiSSA achieves mean GSM8K exact-match 48.07% vs. LoRA's 49.65% ($\Delta = -1.58$, BC 95% CI $[-2.33, -0.89]$, two-sided sign-flip $p \approx 0.004$, **1W/9L/0T**; one-sided $p{\approx}0.999$ in the claimed direction PiSSA > LoRA). Because we use the original paper's training data, evaluation set, hyperparameters, and a single 7B causal LM, this is the cleanest negative result on PiSSA's GSM8K headline; it mirrors the regime-faithful reversal at Qwen2.5-3B (Cell 9). Cell 16 in Table 1 is the only 7B claim-matched cell to survive Holm–Bonferroni correction across the 15-cell family.

7.2 LLaMA-2-7B / commonsense_170k: a prospective boundary test that fails

On commonsense_170k (Hu et al., 2023), the DoRA paper's published configuration (r=32, α=64, lr=2×10^{-4}) gives $\mathrm{lr}_{\mathrm{eff}}$=$2.26\times10^{-3}$, $4.5\times$ above our boundary. We flagged it as unsafe *before* running it, making this a prospective test rather than a post-hoc fit. An initial fp16 attempt was inconclusive: grad-norm NaNs appeared, but six of ten seeds were lost to out-of-memory failures under concurrent GPU sharing, leaving too few usable runs to estimate anything.

We therefore re-ran both arms in bf16 with one job per GPU. **All twenty runs completed—no dropped seeds, no NaN trajectories, and no collapses in either arm**—with LoRA at 64.36 and DoRA at 63.85 (paired $\Delta = -0.51$, two-sided $p = 0.11$, 4W/6L, N=10). **The boundary therefore made a prediction that did not hold, and we withdraw this cell as evidence for the mechanism.** The fp16 NaNs are attributable to dtype rather than $\mathrm{lr}_{\mathrm{eff}}$, since bf16 at the same $\mathrm{lr}_{\mathrm{eff}}$ is stable. We report this because a probe that is only ever evaluated where it succeeds is not a probe; it also sharpens the scope claim—the 5×10^{-4} value is calibrated on BERT-base/MNLI/AdamW and does not transfer unconditionally across architecture and dtype. The cell does, however, contribute a seventh DoRA comparison, at 7B on DoRA's own task family, and DoRA again shows no gain (Appendix K).

8 Discussion and Conclusion

What the audit found. Across 16 claim cells and four LoRA variants, only two confirmatory comparisons remain supported after multiplicity correction, and four reverse. DoRA shows no supported gain in any of its seven audited cells, including a clean 7B cell on its own task family. PiSSA's advantage is real under a common budget at small scale but does not survive paper-faithful hyperparameters, does not transfer to 7B, and flips sign under adversarial distribution shift. AdaLoRA's collapse is real but is largely a low-budget phenomenon: it needs substantially more training data than LoRA to converge (Table 2). rsLoRA is at risk of collapse once its scaling pushes $\mathrm{lr}_{\mathrm{eff}}$ past the boundary ($11/32$ such cells collapsed).

What the mechanism does and does not cover. $\mathrm{lr}_{\mathrm{eff}} \equiv \mathrm{lr} \times \alpha/\sqrt{r}$ is a useful probe for three of the four fragility patterns: no landscape cell below 5×10^{-4} collapsed and DoRA's failures rise with α at high lr, though PiSSA tracks lr alone and all rsLoRA collapses occurred at one learning rate. We are explicit about its limits. The value is calibrated on BERT-base/MNLI/AdamW, not universal. The boundary is necessary but not sufficient: 21 of 32 configurations above it trained normally. It says nothing about AdaLoRA, whose failure is budget-driven. And a prospective 7B test of the boundary failed: LLaMA-2-7B trains stably in bf16 at $4.5\times$ the boundary, so we withdraw that cell as supporting evidence (Appendix K). We report the quantity as a cheap pre-flight probe, not as a law.

Recommendation. The cost of the protocol that produced these findings is modest—ten paired seeds and a bootstrap interval—and it repeatedly changed the answer relative to single-run reporting. We therefore recommend that PEFT comparisons report ≥ 10 paired seeds with a bootstrap CI, at least two hyperparameter regimes (the authors' own and a common budget), the training budget, and $\alpha/\sqrt{r}$; and that reviewers treat a sub-point gain from one or three seeds as unsupported.

Limitations

Model scale. We include Mistral-7B (7.2B), Gemma-7B, and LLaMA-2-7B experiments us-

ing A6000 (48GB) GPUs, but the 7B evidence does not extend to 13B+ scale where adapter-vs-base ratio differs further. Cell 16 (Gemma-7B PiSSA/MetaMathQA→GSM8K, $\Delta = -1.58$, two-sided $p \approx 0.004$, 1W/9L) is the cleanest claim-matched 7B reproduction we run, and Cells 17–18 (LLaMA-2-7B/commonsense_170k (Hu et al., 2023) LoRA vs. DoRA) gave inconclusive fp16 results (OOM and NaN seeds) but trained cleanly in a bf16 paired re-run (Appendix K), so they inform dtype numerics rather than the $\mathrm{lr}_{\mathrm{eff}}$ boundary. Conclusions drawn at 7B should not be over-extrapolated to frontier-scale models.

Method coverage and depth. We audit 4 of >20 active PEFT methods; broader coverage (LoRA+, GaLore, SVFT, VeRA, IA3) may reveal different patterns. We prioritized methods with clear, high-visibility published claims. Our deepest analysis centers on PiSSA (ablation, OOD, rank, scale); DoRA, AdaLoRA, and rsLoRA receive shallower treatment and function mainly as contrast cases. The paper should be read as an audit of *selected* published PEFT claims rather than as a comprehensive PEFT benchmark.

Pre-registration scope. Our audit follows a two-phase protocol (calibration → confirmation) rather than strict clinical-style pre-registration. The statistical protocol and adjudication rules were frozen before confirmatory runs, but the choice of confirmatory families was informed by calibration results. OOD, rank sensitivity, and ablation analyses are exploratory. The pre-registration document is released for full transparency (Appendix N).

Paper-faithful regime and precision control. Our "paper-faithful" regime adapts original settings to our hardware. PiSSA requires FP16 or bf16 weights for SVD initialization, while LoRA can use 4-bit quantization. To address this potential confound, we re-ran LoRA-faithful in fp16+gradient-checkpointing (identical precision to PiSSA) and additionally re-ran all Gemma-7B paired cells in bf16 to remove dtype as a confound. The reversal persists under matched precision (see Table 1 Cell 9 and Appendix H).

LoRA-family scope of the mechanism. The $\mathrm{lr}_{\mathrm{eff}}$ boundary is defined and tested for LoRA and its rank-decomposition variants (DoRA, PiSSA, rsLoRA), which share the $\Delta W = BA$ low-rank update structure with scaling factor α/r. The mechanism is *not* claimed to apply to other PEFT families with structurally different updates: prefix-tuning, prompt-tuning, IA^3 (Liu et al., 2022), Compacter, and BitFit do not have an $\alpha/\sqrt{r}$ analogue. The paper's title ("…for LoRA-Variant Fragility Patterns") reflects this scope. We use "fragility patterns" rather than "failure modes" because the three cases are not the same kind of event: rsLoRA and DoRA exhibit *training collapse* (a binary optimization failure), whereas PiSSA exhibits a *ranking reversal* (its sign relative to LoRA flips without either arm collapsing). $\mathrm{lr}_{\mathrm{eff}}$ unifies them only in predicting *when* each becomes fragile, not in the underlying failure event. AdaLoRA, though structurally a LoRA variant, is excluded from the mechanism narrative because its dominant failure mode is budget-driven—AdaLoRA needs substantially more training data than LoRA to converge (Appendix J)—rather than $\mathrm{lr}_{\mathrm{eff}}$-driven; we treat it as a complementary fragility class.

Single hardware family. All small-model experiments use one cluster (50×RTX 3060, 12GB); all 7B experiments use two A6000 (48GB) GPUs. Optimizer behavior under different hardware (e.g., H100 bf16 with FlashAttention-2 or FP8) could affect the $\mathrm{lr}_{\mathrm{eff}}$ critical threshold value (see Appendix M), and whether a collapse regime exists there is untested.

Statistical assumptions. Our bootstrap CIs assume the empirical distribution of paired deltas adequately captures the population. With $n = 8$–20 seeds, the CI lower bound has known undercoverage; we use BC adjustment to reduce this and report both raw p-values and Holm-corrected adjudications, and false negatives on small effects are likely at these sample sizes: PiSSA/BoolQ reaches its full-sample adjudication in only 61% of $N{=}10$ subsamples (Appendix A). One-sided tests are used only for the Tier-2 ID metrics.

Ethical Considerations

Positive impact. Our work promotes reproducibility and statistical rigor in PEFT research. The proposed reporting standard, if adopted, would reduce overclaiming and help practitioners make better-informed method selections, especially in resource-constrained settings where running additional seeds is expensive. Our open-source audit pipeline and full per-seed data lower the barrier to conducting similar audits, and the $\mathrm{lr}_{\mathrm{eff}}$ inequality gives practitioners a concrete pre-flight check.

Risks of misinterpretation. Our findings could be misinterpreted as evidence that PEFT methods in general are unreliable, potentially discouraging innovation. We emphasize that our results show *conditional* rather than absolute failure: methods like PiSSA offer genuine advantages under specific conditions, and LoRA itself remains a robust default in the audited settings. We also note that our audit methodology could be selectively applied to unfairly target specific papers while ignoring broader context; we mitigate this by reporting all cells (including unsupported and null) rather than only the reversals.

Use of AI assistants. We used AI coding assistants (Claude Code; Codex) to write and review analysis, figure-generation, and cluster-orchestration code, for mathematical typesetting and editorial revision of paragraph-level text, and for a first-pass automated check of bibliographic metadata against arXiv, DOI, and OpenReview records. All experimental designs, statistical protocols, and scientific claims are the authors'. Every reported number is produced by the released scripts and was verified by the authors, and every reference was checked by the authors against its source.

Compute and environmental cost. The 50×RTX 3060 cluster running for the audit consumed approximately 1,200 GPU-hours; the A6000 7B chain adds approximately 400 GPU-hours; the cluster stability landscape adds approximately 800 GPU-hours. Total: ∼2,400 GPU-hours ≈ 520 kWh (200 W per RTX 3060, 300 W per A6000), or ∼180 kg CO_2 at 0.35 kg/kWh (Appendix E). We disclose this to support the broader community's tracking of compute costs, and we note that a key practical implication of our work—the $\mathrm{lr}_{\mathrm{eff}}$ pre-flight check—reduces future compute waste by flagging configurations at risk of collapse.

Licensing. All released code, configs, and metric CSVs are under the MIT license. Pre-trained models follow their respective upstream licenses (LLaMA-2 Community License, Gemma Terms of Use, Apache-2.0 for Mistral and Qwen2.5). All training datasets are publicly released. No human-subjects data, no PII, no proprietary data.

Acknowledgements

This work was supported by the Ministry of Science and ICT and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2026-25498341).

References

D. Biderman, J. Portes, J. J. Gonzalez Ortiz, M. Paul, P. Greengard, C. Jennings, D. King, S. Havens, V. Chiley, J. Frankle, C. Blakeney, and J. P. Cunningham. LoRA Learns Less and Forgets Less. *TMLR*, 2024.

X. Bouthillier, C. Laurent, and P. Vincent. Unreproducible Research is Reproducible. *ICML*, 2019.

S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A Large Annotated Corpus for Learning Natural Language Inference. In *EMNLP*, pages 632–642, 2015.

N. T. Bui, G. K. Savova, and L. Wang. Assessing the Macro and Micro Effects of Random Seeds on Fine-Tuning Large Language Models. In *Proceedings of IJCNLP-AACL 2025 (Short Papers)*, pages 41–46, 2025.

C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL-HLT*, pages 2924–2936, 2019.

K. Cobbe et al. Training Verifiers to Solve Math Word Problems. *arXiv:2110.14168*, 2021.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

A. Edalati, M. Tahaei, I. Kobyzev, V. P. Nia, J. J. Clark, and M. Rezagholizadeh. KronA: Parameter Efficient Tuning with Kronecker Adapter. *arXiv:2212.10650*, 2022.

B. Efron. Better Bootstrap Confidence Intervals. *JASA*, 82(397):171–185, 1987.

Gemma Team. Gemma: Open Models Based on Gemini Research and Technology. *arXiv:2403.08295*, 2024.

S. Hayou, N. Ghosh, and B. Yu. LoRA+: Efficient Low Rank Adaptation of Large Models. *ICML*, 2024.

P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep Reinforcement Learning that Matters. *AAAI*, 2018.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 2022.

Z. Hu, L. Wang, Y. Lan, W. Xu, E.-P. Lim, L. Bing, X. Xu, S. Poria, and R. K.-W. Lee. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. In *EMNLP*, pages 5254–5276, 2023.

M. Javaheripi and S. Bubeck. Phi-2: The Surprising Power of Small Language Models. *Microsoft Research Blog*, December 2023. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

A. Q. Jiang et al. Mistral 7B. *arXiv:2310.06825*, 2023.

D. Kalajdzievski. A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA. *arXiv:2312.03732*, 2023.

D. J. Kopiczko, T. Blankevoort, and Y. M. Asano. VeRA: Vector-based Random Matrix Adaptation. *ICLR*, 2024.

V. Lingam, A. Tejaswi, A. Vavre, A. Shetty, G. K. Gudur, J. Ghosh, A. Dimakis, E. Choi, A. Bojchevski, and S. Sanghavi. SVFT: Parameter-Efficient Fine-Tuning with Singular Vectors. *NeurIPS*, 2024.

Y. Liu et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*, 2019.

H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. Raffel. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. *NeurIPS*, 2022.

S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen. DoRA: Weight-Decomposed Low-Rank Adaptation. *ICML*, 2024.

M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are GANs Created Equal? A Large-Scale Study. *NeurIPS*, 2018.

S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, and M. Tietz. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. `github.com/huggingface/peft`, 2022.

F. Meng, Z. Wang, and M. Zhang. PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. *NeurIPS*, 2024.

Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *ACL*, pages 4885–4901, 2020.

J. O'Neill, P. Rozenshtein, R. Kiryo, M. Kubota, and D. Bollegala. I Wish I Would Have Loved This One, But I Didn't – A Multilingual Dataset for Counterfactual Detection in Product Review. In *EMNLP*, pages 7092–7108, 2021.

J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *JMLR*, 22(164):1–20, 2021.

T. Putterman, D. Lim, Y. Gelberg, S. Jegelka, and H. Maron. Learning on LoRAs: GL-Equivariant Processing of Low-Rank Weight Spaces for Large Finetuned Models. *arXiv:2410.04207*, 2024.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*, pages 2383–2392, 2016.

P. Rajpurkar, R. Jia, and P. Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. In *ACL (Short Papers)*, pages 784–789, 2018.

R. Shuttleworth, J. Andreas, A. Torralba, and P. Sharma. LoRA vs Full Fine-tuning: An Illusion of Equivalence. *NeurIPS*, 2025.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *EMNLP*, pages 1631–1642, 2013.

H. Touvron et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288*, 2023.

A. Williams, N. Nangia, and S. R. Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL-HLT*, pages 1112–1122, 2018.

T. Wolf et al. Transformers: State-of-the-Art Natural Language Processing. *EMNLP System Demonstrations*, 2020.

Y. Xue, X. Cao, X. Yang, Y. Wang, R. Wang, and J. Li. We Need to Talk About Reproducibility in NLP Model Comparison. *EMNLP*, 2023.

A. Yang et al. Qwen2.5 Technical Report. *arXiv:2412.15115*, 2024.

L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu. MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models. *ICLR*, 2024.

X. Zhang, J. Zhao, and Y. LeCun. Character-level Convolutional Networks for Text Classification. In *NeurIPS*, 2015.

Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, and T. Zhao. Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. *ICLR*, 2023.

J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar, and Y. Tian. GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection. *ICML*, 2024.

A Power Analysis: How Many Seeds Do You Need?

For each cell, we simulate subsampling at $N = 3, 5, 7, 10, 15$ and measure the probability of reaching the correct adjudication (matching the full-sample result). We run 500 bootstrap simulations per N.

Table 4: Probability (%) of reaching the full-sample adjudication at reduced seed counts ($|\Delta|$ in accuracy % or GSM8K points). Most cells stabilise by N=10; PiSSA/BoolQ, a small effect near the significance boundary, needs $N \geq 15$. †Raw p=0.011, not significant after Holm. §Computed on the original faithful run (4-bit LoRA arm, N=16, Δ=−1.73), before precision control.

Cell	Adj.	$\lvert\Delta\rvert$	N=3	N=5	N=7	N=10	N=15
PiSSA/MNLI	sup.	0.9%	76	90	99	100	100
PiSSA/Qwen3B com.	sup.	2.6	70	81	97	100	100
PiSSA/Qwen3B faith.§	rev.	1.7	66	82	97	100	100
DoRA/MNLI	unsup.	0.1%	78	85	90	93	95
DoRA/Qwen3B	unsup.	0.05	84	90	93	98	100
PiSSA/BoolQ	unsup.†	1.1%	41	43	46	61	92

Practical recommendation. In the tested cells, PEFT comparisons with expected effect sizes around +1–2% often need roughly **10 paired seeds** for $> 90\%$ adjudication stability. For weaker effects ($< 1\%$), 15+ seeds may be required. The PiSSA/BoolQ cell ($\Delta = +1.1\%$) reaches only 46% stability at N=7, meaning more than half of 7-seed studies would reach the wrong conclusion relative to the full-sample adjudication.

Caveat. The power analysis treats the full-sample (N=16–20) adjudication as ground truth, but the full sample is itself uncertain—e.g., Cell 10 (phi-2) has CI $[-1.39, +0.85]$, leaving the true direction ambiguous even at N=18. The stability percentages should therefore be read as "agreement with the N=20 conclusion" rather than "probability of being objectively correct."

B Exploratory Follow-Up: Why Does the Faithful Regime Reverse PiSSA?

The paper-faithful regime differs from the common budget in three hyperparameters: learning rate (2e-5 vs. 2e-4), α (r vs. $2r$), and dropout (0 vs. 0.05). We construct a 2^2 factorial design over learning rate and α (the two factors with the largest a priori effect), plus a separate dropout ablation, on Qwen2.5-3B / GSM8K with 8–20 paired seeds per condition. This allows formal decomposition into main effects and an interaction term.

Table 5: Three-way hyperparameter ablation on Qwen2.5-3B GSM8K. Each row changes one factor from the common-budget baseline (top row), except the last, which changes all three (*precision-controlled FP16 LoRA arm, as in Cell 9). Δ: PiSSA − LoRA; all columns over paired seeds (20, 10, 8, 8, 19). The rightmost column shows the effect of each change on Δ relative to the common-budget baseline (+2.62).

Regime	lr	α	dropout	LoRA	PiSSA	Δ	Effect on Δ
Common (baseline)	2e-4	16	0.05	48.12	50.73	+2.62	—
dropout → 0	2e-4	16	0	48.57	50.50	+1.93	−0.68
$\alpha \to r$	2e-4	8	0.05	47.13	50.83	+3.71	+1.09
lr → 2e-5	2e-5	16	0.05	50.08	49.83	−0.25	**−2.87**
Faithful (all three)*	2e-5	8	0	49.37	48.07	−1.30	−3.91

Formal factorial analysis. We fit a 2^2 Type-III ANOVA (lr × α) to the per-seed PiSSA−LoRA deltas of the four factorial conditions (common, α-swap, lr-swap, faithful; N=52 paired seeds, unbalanced: 20/8/8/16; dropout is ablated separately). Seeds partially overlap across conditions, so we treat the F-statistics as descriptive; Appendix M re-fits the same data with seed as a random effect.

- **Main effect of lr**: $F(1, 48) = 46.6$, $p < 0.001$, $\eta^2 = 0.44$
- Main effect of α: $F(1, 48) = 0.01$, $p = 0.94$, $\eta^2 < 0.01$
- lr × α interaction: $F(1, 48) = 3.8$, $p = 0.058$, $\eta^2 = 0.04$

Learning rate alone accounts for $\approx 44\%$ of the variance in the PiSSA–LoRA delta; neither α nor the interaction is significant here or in the mixed-effects re-fit, although cluster-robust OLS gives $p \approx 0.02$ for both (Appendix M), so their status is model-dependent. The row-wise pattern in Table 5 agrees: halving α alone slightly *helps* PiSSA (+1.09), whereas lowering lr removes its advantage.

Precision control. In our initial faithful setup, LoRA used 4-bit quantization while PiSSA required FP16. To rule out this confound, we reran LoRA faithful in FP16+gradient-checkpointing (identical precision to PiSSA). The reversal persists: $\Delta = -1.30$ (CI: $[-2.02, -0.61]$, $p = 0.002$, 2W/17L, N=19), compared to $\Delta = -1.73$ (N=16, p<0.001) in the original confounded setup. The quantization confound accounts for only ≈0.43pp (∼25% of the original gap), so precision does not drive the reversal; learning rate does (Table 5).

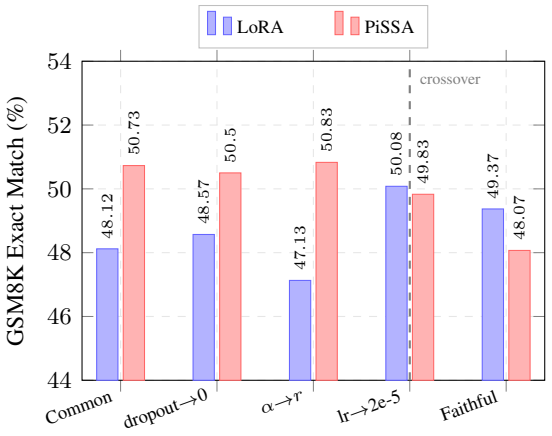


Figure 4: Hyperparameter ablation on Qwen2.5-3B GSM8K. The dashed line marks the crossover: changing lr alone (2e-4 → 2e-5) eliminates PiSSA's +2.6-point advantage. Under the full faithful regime (rightmost), LoRA leads by 1.3%.

Dropout has limited effect in this setting. Removing dropout (0.05 → 0) shifts Δ by only −0.68 (Table 5, row 2), making it much less important than learning rate in this audited ablation.

Hypothesis: effective learning rate sensitivity. One possible explanation is that PiSSA's SVD initialization places adapters near principal components where the effective step size (lr × α/r) strongly shapes training dynamics. At high effective step sizes, this initialization may confer an advantage; at low step sizes, LoRA's random initialization may catch up over training. The α/r scaling may modulate this, explaining the interaction. We do not claim that this mechanism is established or that it generalizes across models or tasks; loss-landscape or gradient-based analysis would be needed to test it directly.

Observation: Under our hardware-adapted faithful setup, the effective learning rate appears strongly associated with the reversal. The reversal persists when both arms use the same precision (Cell 9; Limitations, "Paper-faithful regime and precision control"), but whether $\mathrm{lr}_{\mathrm{eff}}$ fully explains it remains untested.

C Exploratory Follow-Up: Rank Sensitivity

We evaluate LoRA, DoRA, and PiSSA at ranks $\{4, 8, 16, 32\}$ on BoolQ (roberta-base) with 8–10 seeds per condition (20 at r=8, from Cells 1, 2 and 6).

LoRA's performance *decreases* from 73.2 to 68.8 when rank increases from 16 to 32—below its rank-4 performance—because 4 of 9 seeds collapse to the majority-class baseline (63.1) while the others stay near 73 (std 5.8; Table 9). In contrast, PiSSA *improves monotonically* from 68.9 to 75.1 across all ranks. Notably, rsLoRA—designed specifically to stabilize high-rank training (Kalajdzievski, 2023)—degrades even more severely than LoRA, dropping from 71.2 (r=8) to 67.7 (r=16) to 63.1 (r=32, majority-class collapse). This suggests that PiSSA's SVD initialization provides more effective rank stabilization than rsLoRA's scaling fix in this setting.

DoRA's +3.4 at r=32 reflects LoRA's collapses more than a DoRA gain: DoRA itself collapses on 2/9 seeds at r=32 (and 1/10 at r=16), and its mean (72.2) remains below PiSSA (75.1).

Table 6: Rank sensitivity on BoolQ (roberta-base). PiSSA's advantage grows with rank while LoRA degrades at r=32.

Rank	LoRA	DoRA	PiSSA	rsLoRA	Δ(DoRA)	Δ(PiSSA)
4	69.1	69.7	68.9	—	+0.6	−0.2
8	71.1	71.6	72.2	71.2	+0.5	+1.1
16	73.2	72.0	74.6	67.7†	−1.2	+1.3
32	68.8	72.2	75.1	63.1*	+3.4	**+6.3**

*rsLoRA at r=32: all seeds converge to majority-class baseline (std=0). †Degradation begins at r=16. LoRA at r=32: 4/9 seeds at the majority baseline.

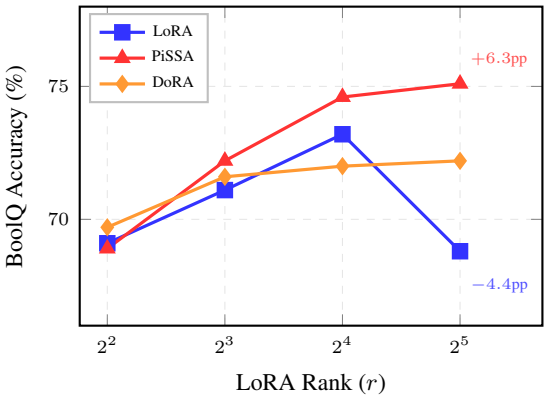


Figure 5: Rank sensitivity on BoolQ (roberta-base). LoRA degrades −4.4pp at r=32 vs. its peak, while PiSSA improves monotonically (+6.3pp gap at r=32). This pattern is consistent with lower rank sensitivity for PiSSA in this setting.

D Full Per-Seed Results

All per-seed deltas, sweep configurations, and analysis scripts are released at `https://github.com/suanlab/repeft`. The materials include:

- `sweep_configs/*.json`: 16+ paired sweep configurations covering all training runs.
- `local_runs/*/results.csv` and

`*/all_metrics.csv`: per-seed metrics for every cell (one row per seed), including the fp16 attempt for Cells 17–18 (Cell 17: 9 completed + 1 collapsed; Cell 18: 2 completed, ≥2 NaN, 6 OOM) and their bf16 re-run.

- `analysis/*/cell_*.json`: paired summary statistics (bootstrap CI, sign-flip p-value, W/L/T counts) for every cell.
- `project_a_pairwise_summary.py`: deterministic re-derivation of every paired statistic in the paper from raw CSV inputs (bootstrap CIs with explicit random-seed control).
- `scripts/make_fig{1,2,3}_*.py`: figure-generation scripts that read directly from the analysis JSONs.
- `PREREGISTRATION.md`: the protocol frozen before the confirmatory runs, plus timestamped amendments (Cell 16, Tier-2).

A reader can re-derive any paired Δ, CI, or p-value in the paper by running `project_a_pairwise_summary.py` on the matching `all_metrics.csv` pair.

E Compute Resources

Table 7: Compute summary per model scale. Wall time is per-seed average.

Model	Params	GPU Mem	Wall Time/Seed	Hardware
roberta-base	125M	∼2 GB	∼3 min	RTX 3060 12GB
bert-base	110M	∼2 GB	∼4 min	RTX 3060 12GB
phi-2	2.7B	∼3 GB (4-bit)	∼35 min	RTX 3060 12GB
Qwen2.5-3B	3.1B	∼8 GB (4-bit)	∼35 min	RTX 3060 12GB
Mistral-7B	7.2B	∼39 GB (FP16)	∼30 min	RTX A6000 48GB

Total compute: over 1,000 training runs across 50×RTX 3060 (12GB) and 2×RTX A6000 (48GB). Aggregate GPU-hours, broken down per workload: ∼1,200h for the 16-cell audit on the 3060 cluster, ∼800h for the 303-training stability landscape on the 3060 cluster, and ∼400h for the A6000 7B chain (Gemma + LLaMA-2 + Mistral), giving a total of ∼2,400 GPU-hours. At 200W per RTX 3060 and 300W per A6000 and a grid carbon intensity of 0.35 kg CO_2/kWh, this corresponds to ∼520 kWh (2,000 h × 200 W + 400 h × 300 W) and ∼180 kg CO_2eq.

Software stack. The main audit ran under Python 3.10 with PyTorch 2.3, HuggingFace `transformers` 4.42, `peft` 0.11, `datasets` 2.18, `bitsandbytes` 0.43 (4-bit quantization for ≥2.7B), `accelerate` 0.30, `numpy` 1.x, `scipy` 1.1x, `lm-eval` 0.4.11, and `matplotlib` 3.x; exact pins are in the released `requirements.txt`.

Stack disclosure for the follow-up runs. The two follow-up runs—the uncapped AdaLoRA schedule (Appendix J) and the bf16 Cells 17–18 sweep (Appendix K)—were executed on an upgraded environment (PyTorch 2.7.1, `transformers` 5.3.0, `peft` 0.18.1, `numpy` 2.2). Because a `peft` major-version change could in principle alter AdaLoRA's allocator behaviour, we re-ran the *original* Cell 5 configuration on the upgraded stack as a version control; results are reported in Appendix J.

Reproducibility statement. The public repository (`https://github.com/suanlab/repeft`) contains every sweep configuration, per-seed metric CSV, paired summary JSON, figure-generation script, and the analysis driver `project_a_pairwise_summary.py`. Re-deriving any paired Δ, CI, or p-value in the paper is a single command. The included integration test `tests/test_paired_summary.py` re-derives six headline statistics from the released data.

Table 8: Method cost-quality summary on Qwen2.5-3B GSM8K (common budget). Mean EM across 20 paired seeds; wall time per seed on RTX 3060. GC: gradient checkpointing.

Method	Mean EM (%)	Δ vs LoRA	Wall Time	Note
LoRA (4-bit)	48.12	—	∼35 min	baseline
DoRA (4-bit)	48.07	−0.05	∼55 min	1.6× slower, no gain
PiSSA (FP16+GC)	50.73	+2.62	∼55 min	gains but regime-fragile
rsLoRA (4-bit)	—	—	∼35 min	untested at 3B

F Rank Sensitivity: Full Results

Table 9 extends Table 6 with standard deviations.

Table 9: Full rank sensitivity results on BoolQ (roberta-base). Mean accuracy ± std. The r=8 row uses the 20 confirmatory seeds of Cells 1, 2 and 6 (rsLoRA: 17); other ranks use 8–10 completed seeds.

Rank	LoRA	DoRA	PiSSA	rsLoRA
4	69.1 ± 0.7	69.7 ± 1.1	68.9 ± 0.8	—
8	71.1 ± 1.5	71.6 ± 1.7	72.2 ± 1.5	71.2 ± 4.5
16	73.2 ± 1.1	72.0 ± 3.5	74.6 ± 0.8	67.7 ± 6.4
32	68.8 ± 5.8	72.2 ± 5.3	75.1 ± 2.1	63.1 ± 0.0

G Tier-2 Multi-Seed Extension Results

This appendix reports an independent 19–20-seed extension of the OOD and rank-sensitivity cells using the BC bootstrap, with a one-sided sign-flip for ID metrics and a two-sided one for OOD metrics. The Tier-2 sweeps were run after the main confirmatory cells finished, sharing only the seed list ($\{11, 21, 31, \ldots, 201\}$) with prior runs. The numbers reproduce the headline directions and tighten the confidence intervals.

rsLoRA rank-sensitivity on BoolQ (N=10) is reported separately in Table 9, where r=32 (α=64) collapses on every seed (mean 0.6307, std 0; 10/10 in a complete re-run) and r=16 is already degraded (5 of 8 completed seeds at the majority baseline, mean 0.677; 9/10 seeds at the baseline in the re-run).

The two strongest Tier-2 effects are PiSSA's MNLI matched gain (+1.0%, 19/0/0) and PiSSA's Amazon-CF gain (+2.5%, 19/0/0), both surviving Holm–Bonferroni correction within the Tier-2 family (computed by `analysis/holm.py --tier2`; 3 of 16 cells survive, including the PiSSA/ANLI R2 reversal). The rsLoRA r=32 collapse on BoolQ—all 10 seeds converging to the majority-class baseline of 0.6307 with zero variance—hardens the rank-sensitivity story: at r=32, rsLoRA does not merely degrade, it ceases to learn. ANLI R1 is reported under the pre-specified NaN-metric rule (§3, "Failure handling"): one LoRA-baseline-arm seed produced a `NaN` accuracy; the symmetric drop reduced N from 19 to 18. The BC bootstrap on the remaining 18 paired seeds gives PiSSA $\Delta = -0.0080$, BC 95% CI $[-0.0136, -0.0022]$, two-sided sign-flip p=0.017, 5W/13L/0T (DoRA: $\Delta = -0.0017$, p=0.50, unsupported; see `analysis/tier2_paired/anli_r1_\{pissa,dora\}_nan_dropped.json`). PiSSA/ANLI R1's BC CI excludes zero and is directionally consistent with ANLI R2; including ANLI R1 therefore strengthens—not weakens—the adversarial-NLI reversal pattern. ANLI R1 is reported outside the 16-cell Tier-2 Holm family, and PiSSA's p=0.017 would not survive that correction, but the raw sign-flip p and BC CI both indicate the same reversal direction as ANLI R2.

Methodological upgrade vs. main confirmatory results. Table 1 uses two-sided sign-flip p-values throughout and the percentile bootstrap except row 16 (BC). Tier-2 results use the BC bootstrap and a one-sided sign-flip for ID metrics with pre-specified direction.

H fp16 vs. bf16 Dtype Sensitivity (Gemma-7B)

During the Gemma-7B/PiSSA experiments referenced in §7, an early run with the script's default fp16 mixed-precision setting produced *deterministic gradient overflow*: within $\sim$ 100 training steps, the per-step gradient norm transitioned to `NaN` and the running training loss collapsed to 0, while reported `train_loss` (a Trainer-internal moving average) appeared finite at $\approx$ 2.6. The post-training eval-loss was `NaN` and downstream GSM8K exact-match was exactly 0% across all attempted seeds. This is consistent with previously reported fp16 numerical-stability problems for Gemma's RMS-Norm and SwiGLU activations under SVD-initialized PiSSA matrices (r=128, α=128, large initial LoRA-A/B norms).

We diagnosed this via the per-step grad-norm log and resolved by switching the Trainer to bfloat16 mixed-precision and loading the model in `torch.bfloat16`. Under bf16, the same configuration trained cleanly with grad-norms in $[3, 25]$ and produced GSM8K exact-match 46.25%–49.43% across 10 seeds (§7).

Paired fp16 vs. bf16 effect on LoRA (4 seeds). LoRA on Gemma-7B was less sensitive but the dtype is not entirely null. We have 4 paired seeds (11, 21, 31, 41) trained in both fp16 and bf16 with otherwise identical configs (Table eftab:gemma$_d$*type*).

The dtype effect is small and suggestive for LoRA, while it is catastrophic for PiSSA. We reran all 10 LoRA seeds in bf16 to remove dtype as a confound in the paired Cell 16 comparison; the original 4-seed fp16 LoRA results are preserved in the repository (`logs/preserved_fp16_gemma_*`) for transparency.

Practical recommendation. When training Gemma-class models with high-rank PiSSA initializations, force bf16 mixed-precision at the Trainer level and load weights in `torch.bfloat16`. The fp16 default of HuggingFace `Trainer` is *silent*: training appears to proceed (loss not visibly diverging in summary statistics) while in fact the adapter has saturated to garbage. Our auto-detection logic (set `bf16=True` whenever the model name contains "gemma") is shipped

Table 10: Tier-2 paired summaries (19–20 seeds, roberta-base, BC bootstrap 95% CI). ID metrics use one-sided sign-flip (H_1: candidate $>$ LoRA, motivated by source paper); OOD metrics use two-sided. ID-mm: MNLI mismatched. These 16 rows form the Tier-2 Holm family; ANLI R1 is reported separately under the NaN rule.

Family	Comparison	Δ	95% BC CI	p	W/L/T	Adj.
NLI: MNLI (ID) → SNLI / ANLI (OOD)						
NLI ID	PiSSA − LoRA	$+0.0103$	$[+0.0075, +0.0144]$	1.9×10^{-6}	19/0/0	supported
NLI ID	DoRA − LoRA	$+0.0028$	$[-0.0006, +0.0060]$	0.066	11/8/0	unsupported
NLI ID-mm	PiSSA − LoRA	$+0.0048$	$[-0.0005, +0.0089]$	0.036	14/4/1	unsupported
NLI ID-mm	DoRA − LoRA	-0.0001	$[-0.0024, +0.0020]$	0.56	10/8/1	unsupported
SNLI	PiSSA − LoRA	$+0.0032$	$[-0.0039, +0.0104]$	0.42	9/9/1	unsupported
SNLI	DoRA − LoRA	-0.0011	$[-0.0072, +0.0054]$	0.75	7/12/0	unsupported
ANLI R2	PiSSA − LoRA	-0.0099	$[-0.0142, -0.0059]$	2.0×10^{-4}	4/15/0	**reversed**
ANLI R2	DoRA − LoRA	$+0.0007$	$[-0.0033, +0.0042]$	0.71	11/6/2	unsupported
ANLI R3	PiSSA − LoRA	-0.0043	$[-0.0105, +0.0021]$	0.21	7/12/0	unsupported
ANLI R3	DoRA − LoRA	-0.0031	$[-0.0069, +0.0009]$	0.17	5/13/1	unsupported
Sentiment: SST-2 (ID) → Yelp / Amazon-CF (OOD)						
SST-2 ID	PiSSA − LoRA	-0.0018	$[-0.0042, +0.0005]$	0.91	8/10/1	unsupported
SST-2 ID	DoRA − LoRA	-0.0009	$[-0.0025, +0.0007]$	0.86	5/11/3	unsupported
Yelp	PiSSA − LoRA	$+0.0012$	$[-0.0014, +0.0039]$	0.41	9/9/1	unsupported
Yelp	DoRA − LoRA	$+0.0007$	$[-0.0012, +0.0025]$	0.52	8/5/6	unsupported
Amazon-CF	PiSSA − LoRA	$+0.0247$	$[+0.0185, +0.0316]$	3.8×10^{-6}	**19/0/0**	**supported**
Amazon-CF	DoRA − LoRA	$+0.0035$	$[+0.0008, +0.0061]$	0.032	12/5/2	unsupported

in our codebase and should generalize to other models exhibiting the same pattern.

I 4-Method Stability Matrix (Synthesis)

J What drives Cell 5: budget, not schedule (a 2×2 control)

Cell 5 is the audit's largest reversal (AdaLoRA F1 15.59 vs. LoRA 52.34, $\Delta = -36.75$, N=9). A natural objection is that this may measure our setup rather than AdaLoRA, because our training harness (`real_lora_qa.py`) clamps the requested AdaLoRA schedule to `tinit` $\leq$ `total_steps`$/10$ and `tfinal` $\leq$ `total_steps`$/3$ as a guard against schedules that leave the rank allocator no budget phase. **This clamp is ours, not the `peft` library's**; we earlier hypothesised that it caused Cell 5's magnitude, and the control below shows that hypothesis was *wrong*: the schedule is irrelevant here and the training budget is what matters. We ran a 2×2 over {train budget} $\times$ {schedule}, ten paired seeds per cell, everything else fixed (bert-base-uncased, SQuAD v1, r=8, α=16, lr= 3×10^{-4}, seq 256, 2 epochs, 1K eval).

The schedule is not the cause. Holding budget fixed, changing the requested schedule by up to $5\times$ moves AdaLoRA's F1 by <0.01 at 5K and 0.02 at 12K (same software stack; the original Cell 5 run gave 15.59). The clamp therefore had no material effect on Cell 5's outcome, and neither did removing it (tested at 12K, where the request is applied uncapped).

The training budget is the cause. Holding schedule fixed at the original $50/200$, going from 5K to 12K samples moves AdaLoRA from 15.53 to 55.38 F1 ($+39.8$, same stack). LoRA over the same budget change moves from 53.40 to 65.49 ($+12.1$). **AdaLoRA is far more data-hungry than LoRA in this setting**, and Cell 5's dramatic gap is largely a small-budget effect rather than a property that would persist at scale.

The reversal survives, at a quarter of the magnitude. At the larger budget the paired comparison is $\Delta = -10.09$ F1 (BC 95% CI $[-10.74, -9.40]$, two-sided $p = 0.002$, $0/10/0$) for the uncapped schedule and $\Delta = -10.11$ ($0/10/0$) for the original schedule. AdaLoRA still loses on every seed, so Cell 5's *direction* is robust, but we now report ≈ -10 rather than -36.7 as the budget-fair estimate and describe Cell 5 as a low-budget regime result.

Library-version control. The re-runs above used `peft` 0.18.1 / PyTorch 2.7.1, whereas the original audit used `peft` 0.11 / PyTorch 2.3. Re-running the *original* Cell 5 configuration on the new stack reproduces the old numbers closely (AdaLoRA 15.53 vs. 15.59; LoRA 53.40 vs. 52.34; $\Delta = -37.87$ vs. -36.75; N=10, $0/10/0$), so the library upgrade is not a confound for any comparison in this appendix. Per-seed values for all four cells are in `analysis/cell5_uncapped.js`

Table 12: Method × architecture stability summary referenced in §5. Cells report stability outcomes (collapse vs. none) rather than paired deltas; — marks combinations we did not measure.

Method	BERT-base / MNLI	RoBERTa-base / BoolQ	Qwen2.5-3B / GSM8K	Mistral-7B / MNLI
LoRA	—	$4/9$ collapse at r=32 (α=64)	stable	$10/10$ stable
DoRA	38% fail at α/r=1, r=64, lr= 10^{-3}	$1/10$ (r=16), $2/9$ (r=32) collapse	stable	$10/10$ stable
PiSSA	—	stable through r=32	stable	$10/10$ stable
rsLoRA	$11/72$ cells collapse, all at lr= 10^{-3}	$10/10$ **collapse at** r=32 (α=64)	—	$2/10$ **strict +** $1/10$ **partial collapse naive;** $10/10$ **stable tuned**

Table 13: AdaLoRA mean F1 over the budget × schedule grid. At 12K the caps are $(300, 1000)$, so the $250/750$ request is applied verbatim (logged `fallback_engaged=False`); at 5K the caps are $(125, 416)$. All cells use one software stack (the original Cell 5 run gave 15.59). Schedule moves F1 by < 0.05 in both budget regimes; budget moves it by ≈ 40.

	Requested schedule `tinit/tfinal`	
Budget	$50/200$	$250/750$
5K train	15.53 (Cell 5 config)	15.53 (clamped)
12K train	**55.38**	**55.40** (uncapped)

on, `analysis/cell5_budgetcontrol.json`, and `analysis/cell5_versioncheck.json`.

K Cells 17–18 (LLaMA-2-7B / commonsense_170k): fp16 numerics, resolved in bf16

The DoRA paper's published configuration (r=32, α=64, lr=2×10^{-4}) yields lr_{eff}=2.26×10^{-3}, $4.5\times$ above the BERT-base critical product of §5. The mechanism flagged this configuration as unsafe *before* we ran it, so the cell was intended as a prospective test of the boundary.

fp16 attempt (inconclusive). With the DoRA paper's adapter hyperparameters and fp16, but a reduced budget (batch 2 instead of 16, 1 epoch instead of 3, seq 512) on commonsense_170k and a 7-task lm-evaluation-harness aggregate (BoolQ, PIQA, HellaSwag, WinoGrande, ARC-e, ARC-c, OBQA; DoRA uses 8-task generation-based evaluation), Cell 17 LoRA gave $9/10$ OK in $[61.10, 66.72]$ (mean 64.83) with seed 21 collapsing to 33.90%; Cell 18 DoRA gave only 2 OK seeds, $\geq$ 2 explicit grad-norm NaN trajectories, and 6 seeds lost to CUDA-OOM under concurrent GPU sharing. With four usable runs this cell could not support any failure-rate estimate.

bf16 re-run (resolved). We re-ran both arms in bf16 with exactly one job per GPU (removing the OOM cause), 10 paired seeds, 25K train samples. **All 20 runs completed: zero OOM drops, zero NaN trajectories, and zero collapses in either arm.** LoRA mean 64.36, DoRA mean 63.85; paired $\Delta = -0.51$, two-sided sign-flip $p = 0.11$, 4W/6L/0T at N=10 (per-seed values in `analysis/cell17_18_bf16_raw.json`).

What this settles. Two conclusions follow, one negative for our mechanism and one positive for the audit. **(i) The fp16 NaNs are attributable to dtype, not** lr_{eff}**.** At an *identical* lr_{eff} of 2.26×10^{-3}—$4.5\times$ above the boundary—bf16 training is entirely stable in both arms. The boundary therefore did *not* predict this failure, and **we withdraw Cells 17–18 as evidence for the mechanism**. This is consistent with our framing of the boundary as a value calibrated on BERT-base/MNLI/AdamW rather than a dtype- and architecture-independent constant, and it sharpens the scope claim in Limitations. **(ii) DoRA remains unsupported.** The clean bf16 cell contributes a seventh DoRA comparison—at 7B, on the DoRA paper's own task family, with no seed lost—and it too shows no gain ($\Delta = -0.51$, $p = 0.11$).

L Mistral-7B / MNLI Matched: Full Paired Stability Cell

This table accompanies the inline summary in §5.4. All four comparisons share the same Mistral-7B/MNLI matched protocol (r=64, lr= 10^{-4}, batch 16, seq 256, 3 epochs, 5K train / 1K eval, 10 paired seeds per arm).

Table 14: Mistral-7B / MNLI matched (10 paired seeds per arm). $\alpha/\sqrt{r}$=4.0 corresponds to a fixed α=32 at r=64; the tuned setting follows our boundary recommendation. For naive rsLoRA: 2 of 10 seeds drop to majority-class accuracy (~ 0.34; strict collapse), 1 of 10 trains partially to ~ 0.59 (intermediate degradation), and 7 of 10 train normally to ~ 0.88. Per-seed numbers are in `analysis/mistral_rslora_collapse_counts.csv`.

Comparison	α	$\alpha/\sqrt{r}$	Δ(method − LoRA)	95% CI	p (paired)
DoRA vs. LoRA	32	4.0	$+\mathbf{0.0038}$	$[+0.0008, +0.0065]$	0.047
PiSSA vs. LoRA	32	4.0	-0.0029	$[-0.0087, +0.0019]$	0.42
rsLoRA-naive vs. LoRA	32	4.0	$-\mathbf{0.131}$ (2/10 strict + 1/10 partial collapse)	$[-0.270, +0.004]$	0.24
rsLoRA-tuned vs. LoRA	4	0.5	0.000 (10/10 stable)	$[0.0, 0.0]$	1.00

M Robustness Checks

This appendix addresses common challenges to multi-seed audits.

Subsample bias. The audit uses subsampled training and evaluation sets (e.g., MNLI 7K/1.5K, GSM8K 3K/300) to fit within the 12GB-per-node cluster budget. We make two arguments that the *direction* of our findings is not subsample-driven, while explicitly conceding that all reported *magnitudes* are conditional on the per-cell budget. First, internal consistency: cells using larger budgets (e.g., MNLI 50K in the stability landscape, 50K MetaMathQA → GSM8K-test for Cell 16, the bf16 LLaMA-2-7B cell) yield the same qualitative pattern as their smaller-budget counterparts (DoRA unsupported, PiSSA regime-sensitive). Second, like prior reproducibility audits (Lucic et al., 2018; Henderson et al., 2018), which test whether comparative claims survive re-evaluation, our claims are made at the ranking level. A full-budget rerun of one DoRA cell is left to future work; its configuration is released as `sweep_configs/robustness_full_boolq_dora.json`.

Hardware bias. Our small-model experiments use a 50×RTX 3060 (Ampere, 12GB) cluster; the 7B cells use two A6000 (Ampere, 48GB). One could ask whether the $\mathrm{lr}_{\mathrm{eff}}$ boundary value would shift on H100/Hopper with FP8 or FlashAttention-2. We expect the numerical value to move; whether a collapse regime exists on such hardware is untested (one hypothesis is that it depends on the exponent range of the adapter update rather than mantissa precision). Mistral-7B already shows the value shifting across architectures (§5.4).

Implementation correctness. All PEFT variants use HuggingFace `peft` library configurations rather than re-implementations: `LoraConfig`, DoRA via `use_dora=True`, PiSSA via `init_lora_weights="pissa"`, `AdaLoraConfig`, and rsLoRA via `use_rslora=True`. The HuggingFace `peft` library is the most widely adopted PEFT implementation. Reusing this single library across all variants—rather than vendored copies of each method's repository—means a single implementation-correctness check applies uniformly: any audit finding that disagrees with our paper would also have to be explained by every other study that uses the same library.

Statistical power. Cells with N=8–9 (Cells 5, 12, 14, 15) have limited power to detect small effects. We report exact W/L/T counts and sign-flip p-values for transparency; the Holm–Bonferroni correction reduces N=8 cells' acceptance to extreme effects only (Cell 12 with $\Delta = +0.04$ would require $|\Delta| > 1.5$ at N=8 to survive). Cell 15 is exploratory by design (a claim-matched variant of Cell 9) and outside the Holm family. The 303-training stability landscape (Fig. 2) and the 20-seed Tier-2 extension (Appendix G) provide higher-power independent corroboration.

Sign-flip vs paired-t sensitivity. The reported p-values use sign-flip permutation tests because they are robust to non-normality at N=8–20. As a sensitivity check, we recomputed each Holm-family cell's p-value with a paired-t test (same direction hypothesis, two-sided p where the sign-flip is two-sided; Table 15). The qualitative adjudication (supported / unsupported / reversed) is identical under both tests. Numerical p-values can differ when the per-seed delta variance is very small (e.g., Cell 11 where AdaLoRA's deltas are nearly constant: paired-t collapses to $\sim 10^{-8}$ while the sign-flip permutation gives ~ 0.002); both still classify as "reversed". Direction stability holds across all cells. The full per-seed table is released (`analysis/paired_t_sensitivity.csv`).

Union-family Holm sensitivity. The main text applies Holm–Bonferroni separately within the

Table 15: Sign-flip vs paired-t p-value sensitivity. Adjudication is identical under both tests for all 16 cells. Cell 11 shows the largest numerical gap (paired-t much smaller) due to extreme low variance of AdaLoRA's per-seed deltas; the gap does not change the adjudication.

#	Method	Model	Task	N	sign-flip p	paired-t p	Adj. (both)
1	DoRA	roberta-base	BoolQ	20	.151	.158	unsup.
2	PiSSA	roberta-base	BoolQ	20	.011	.014	unsup.†
3	DoRA	roberta-base	MNLI	20	.310	.328	unsup.
4	PiSSA	roberta-base	MNLI	20	$< .001$	$< .001$	**sup.**
5	AdaLoRA	bert-base	SQuAD1	9	.004	.003	**rev.**
6	rsLoRA	roberta-base	BoolQ	17	.833	.847	unsup.
7	PiSSA	Qwen3B	GSM8K	20	$< .001$	$< .001$	**sup.**
8	DoRA	Qwen3B	GSM8K	20	.887	.888	unsup.
9	PiSSA	Qwen3B	GSM8K-f	19	.002	.0024	**rev.**
10	PiSSA	phi-2	GSM8K	18	.632	.659	unsup.
11	AdaLoRA	bert-base	SQuAD2	10	.002	4×10^{-8}‡	**rev.**
12	DoRA	Qwen3B	GSM8K-r32	8	.89	.93	unsup.
13	PiSSA	Mistral-7B	GSM8K	10	.008	.0075	unsup.¶
14	DoRA	Mistral-7B	GSM8K	8	.55	.61	unsup.
15	PiSSA	Qwen3B	GSM8K-cm	8	.10	.07	unsup.§
16	**PiSSA**	**Gemma-7B**	**GSM8K-cm**	10	.004	.0028	**rev.**

‡ Cell 11's paired-t is unusually small because the per-seed deltas have very low variance (σ=0.57 on $\Delta = -3.02$); the sign-flip permutation correctly assigns more probability to extreme rearrangements. Both tests classify as reversed. Labels are post-Holm as in Table 1 (†, §, ¶ as there). GSM8K-f: paper-faithful; -cm: claim-matched; -r32: r=32.

15-cell confirmatory family (§4) and the Tier-2 family (Appendix G). A more conservative alternative is union-family Holm: pool both families (15 + 16 cells) and the two ANLI R1 cells into a single m=33-cell correction (computed by `analysis/union_holm.py`; output in `analysis/union_holm_thresholds.csv`). Five cells survive at α=0.05: PiSSA/MNLI matched (p=1.9×10^{-6}, Tier-2 supported), PiSSA/Amazon-CF (p=3.8×10^{-6}, Tier-2 supported), PiSSA/ANLI R2 (p=2.0×10^{-4}, Tier-2 reversed), PiSSA/Qwen3B-common Cell 7 (p=2×10^{-4}, supported), and PiSSA/MNLI Cell 4 (p=6×10^{-4}, supported). Cell 16 Gemma-7B (p=0.004) does not survive union-family Holm (threshold $\approx$ 0.002 at rank 8); we report this transparently because the union family is the most stringent correction, while the 15-cell confirmatory family is the pre-registered correction reported in the main results.

Mixed-effects re-analysis of the §B factorial. The descriptive $\eta^2\approx$0.44 for lr in Table 5 ignores between-seed correlation. We re-fit the 2×2 factorial as a mixed-effects model with seed as a random intercept (`analysis/factorial_anova_mixed.py`; N=52 paired observations, 20 unique seeds, ICC 0.20): lr (hi vs lo) is highly significant (coef +4.81, z=5.58, p=2.4×10^{-8}), while α (coef +0.85, p=0.32) and the lr×α interaction (coef −1.69, p=0.19) are not. A sensitivity check using OLS with cluster-robust SE on seed confirms the lr effect (coef +5.15, p=2.2×10^{-9}) but flags α (coef +1.19, p=0.02) and the interaction (coef −2.28, p=0.02). Learning rate is therefore the dominant factor under every model, while the smaller α and interaction effects are model-dependent.

$\mathrm{lr}_{\mathrm{eff}}$ scaling across model sizes. The critical product $\approx 5\times10^{-4}$ is measured on BERT-base/MNLI/AdamW; at Mistral-7B naive rsLoRA already degrades at 4×10^{-4} while 5×10^{-5} is stable, so the Mistral value lies in $(5\times10^{-5}, 4\times10^{-4}]$ (§5.4). With only two architectures and one task family we do not claim a scaling law: collapse occurred above a model-dependent value in both, and that value is architecture- and optimizer-dependent. Practitioners should treat 5×10^{-4} as an order-of-magnitude guide rather than a universal constant and probe their own architecture with a short multi-seed $\mathrm{lr}_{\mathrm{eff}}$ sweep before relying on it.

H100 / FP8 / FlashAttention-2 portability. Our experiments use Ampere (RTX 3060, A6000) with FP16/bf16/4-bit (no FP8) and standard SDPA attention (no FlashAttention-2). We hypothesise that collapse depends on the magnitude of the adapter update relative to base-weight exponent range rather than on attention kernel or mantissa precision, but this is untested. Practitioners deploying on H100 with FP8 should treat the 5×10^{-4} value as a starting point and verify with a short multi-seed probe before applying the rule at scale.

N Artifact Index for Main-Text Claims

The main text refers to released artifacts without naming files. Table 16 maps each reference to its path in the public repository (`https://github.com/suanlab/repeft`).

Table 16: Released files referenced from the main text and Limitations.

Referenced from	File
§3 (Study Design Transparency), Limitations (Pre-registration scope)	`PREREGISTRATION.md` (SHA-256 below)
§4 (Multiple testing)	`analysis/holm.py` (Holm thresholds)
§5.4 (Mistral-7B rsLoRA)	`analysis/mistral_rslora_collapse_counts.csv` (per-seed collapse counts)

SHA-256 of `PREREGISTRATION.md`:
`05a612d362f422cfab49203869497f440a493cc75c6e181c1afe160953411402`