When Seeds Change the Evidence: Seed-Conditional PEFT Comparisons

Anonymous authors
Paper under double-blind review

Abstract

Average parameter-efficient fine-tuning (PEFT) performance can appear stable while individual stochastic runs exhibit severe lower-tail failures. In a matched LLaMA-3-8B panel spanning 20 seeds, eight commonsense tasks, and rank-8 LoRA, DoRA, and rsLoRA, DoRA averages 84.65% but reaches 62.17% in its worst seed–task cell; rsLoRA averages 83.55% but reaches 52.47%. Lower-tail reliability is only one part of the evaluation problem: even when the directional winner is stable, the magnitude of a reported advantage can depend on the selected seeds. We separate these two objects with complementary finite-panel summaries. The *seed recovery profile* (SRP) describes method-level worst-task recovery, while the *claim survival profile* (CSP) conditions on k-seed subsets supporting a favorable practical claim and measures how often the complementary seeds retain the same margin. For DoRA over LoRA, the sign persists on all 13 complements of favorable one-seed reports at $\delta = 0$, but a 0.5-point margin persists on only 3 of 5. Exhaustive enumeration shows that small reporting budgets can estimate mean accuracy closely while missing lower-tail outcomes and overstating practical margins. A prospectively frozen five-seed, four-task Qwen3-8B panel documents severe seed-indexed differences on a second backbone. Complete paired outcomes remain the primary empirical record; SRP and CSP expose two aspects that average accuracy omits.

1 Introduction

Parameter-efficient fine-tuning (PEFT) methods are commonly compared using average downstream accuracy over one or a few random seeds. That average can look healthy while individual stochastic runs fail severely: in our 20-seed LLaMA-3-8B panel, DoRA averages 84.65% across eight tasks but reaches 62.17% in its worst seed–task cell, while rsLoRA averages 83.55% but reaches 52.47%. Such outcomes are hidden by a conventional winner table.

Lower-tail reliability is only one part of the problem. A comparison can preserve the same directional winner while the magnitude of its claimed advantage changes with the reported seeds. This distinction matters when claims use a practical effect margin rather than a zero-margin sign: in our panel, complements of favorable DoRA-over-LoRA reports nearly always retain the direction at $\delta = 0$, whereas the 0.5-point advantage is sensitive to the reporting subset. Thus seed randomness affects both run-level reliability and the strength of a comparative claim.

Random-seed instability is well established in sequence tagging, encoder fine-tuning, and parameter-efficient tuning (Reimers & Gurevych, 2017; Dodge et al., 2020; Mosbach et al., 2021; Chen et al., 2022; Bui et al., 2025). We examine these consequences in low-rank adaptation of 8B decoder LLMs. LoRA learns low-rank updates to frozen weight matrices, while DoRA and rank-stabilized LoRA (rsLoRA) modify the parameterization or scaling (Hu et al., 2022; Liu et al., 2024; Kalajdzievski, 2023). The concern is not merely variance within one method: small seed panels can provide an incomplete picture of both lower-tail reliability and the practical magnitude of paired method differences.

We summarize these complementary questions with the *seed recovery profile* (SRP), the distribution of worst-task normalized recovery, and the *claim survival profile* (CSP), the fraction of favorable reporting subsets whose complements retain the same practical margin. Both are traceable to the complete paired seed-by-task table; neither implies that lower-tail failures cause claim sensitivity.

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.

Contributions.

- We distinguish two seed-conditioned evaluation problems in PEFT: method-level lower-tail exposure and comparison-level practical-margin sensitivity.
- We introduce SRP for method-level lower-tail exposure and CSP for selection-conditioned practical-margin sensitivity, while treating complete paired seed-wise outcomes as the primary evidence.
- We demonstrate both problems in a 20-seed, eight-task LLaMA panel and examine their scope with a prospectively frozen Qwen panel, a second training corpus, and a documented-recipe boundary case. We release cell-level results, manifests, and code.

2 Related Work

Low-rank adaptation and PEFT. PEFT adapts a frozen model through a small number of trainable parameters, including adapters (Houlsby et al., 2019), prefix tuning (Li & Liang, 2021), IA^3 (Liu et al., 2022a), and low-rank updates (Hu et al., 2022). Many LoRA variants change the update parameterization, rank allocation, or scaling rule, including AdaLoRA, VeRA, rsLoRA, and DoRA (Zhang et al., 2023; Kopiczko et al., 2024; Kalajdzievski, 2023; Liu et al., 2024). Their empirical comparisons primarily report downstream accuracy, sometimes aggregated over a small number of runs. Our object of study is instead the paired directional comparison and its sensitivity to the reported seed subset. PEFT-Bench provides a complementary end-to-end benchmark over diverse autoregressive LLM tasks and methods, together with an efficiency-aware cost measure (Belanec et al., 2026). We isolate seed-conditional comparison under a matched training recipe.

Randomness and empirical reporting. The importance of repeated runs predates PEFT. Reimers & Gurevych (2017) show that single scores can misrepresent non-deterministic LSTM sequence taggers and advocate reporting score distributions. For full-parameter encoder-model fine-tuning, Dodge et al. (2020) identify random head initialization and data order as important sources of BERT variation, while Mosbach et al. (2021) analyze fine-tuning instability and distinguish optimization failures from residual generalization variation. PEFT-specific work shows that instability and problematic validation can yield unreliable conclusions (Chen et al., 2022), while recent LLM fine-tuning work measures score variation and prediction-level consistency across seeds (Bui et al., 2025). These studies characterize variation within a method. We study whether a directional comparison supported by one seed subset retains its margin on the complementary seeds. Recent studies also show that LoRA conclusions depend on method-specific learning-rate tuning (Lee et al., 2026) and that initialization changes learning-rate robustness (Li et al., 2025). Optimization interventions such as DENI (Pecher et al., 2024) and Stable-LoRA (Wu et al., 2026) address training or feature-learning stability; our analysis concerns the resulting comparative evidence under a fixed, matched recipe.

Instability measures and selection-conditioned claims. Du & Nguyen (2023) compare standard deviation with six prediction- and representation-level instability measures for fine-tuning. These characterize variation within a model; our target is different: whether a directional *comparison claim* made from a reported seed subset persists on its complement. This target borrows the idea of conditioning on a selection event from post-selection inference (Berk et al., 2013; Lee et al., 2016), but not its inferential guarantees. CSP is a descriptive finite-panel audit of how a comparison varies across reporting subsets, not a confidence interval or population-level test. Appendix G gives an object-level comparison with all seven of their measures.

PEFT reporting practice. To assess whether these objects are normally visible in PEFT papers, we conduct a descriptive audit of 37 papers. The appendix reports its frozen sampling frame, coding rules, and evidence locations. The audit documents the reporting gap that the proposed protocol addresses.

3 Two Seed-Conditioned Evidence Problems

Let $a_{m,s,t} \in [0, 1]$ denote accuracy of method m at seed s on task t; we express accuracy gaps and comparison margins in percentage points. A conventional comparison reports $a_{m,s,t}$ for one s, or a

mean over a few seeds. This can omit two different properties. First, method-level lower-tail events may be task-specific or co-occur across tasks for one stochastic run. Second, paired advantages may change in sign or practical magnitude across reporting subsets. The second property does not require the first to cause it.

We use two distinct terms. A *lower-tail outlier* is a seed–task value that is unusually low relative to the other seeds for the same model, method, rank, and task. For cell (m, s, t), we report its severity as the leave-one-seed-out median gap $g_{m,s,t} = 100(\mathrm{median}_{s' \neq s}\, a_{m,s',t} - a_{m,s,t})$. Positive values measure the deficit relative to the other paired runs; their magnitude, rather than positivity alone, distinguishes large failures from routine dispersion. We retain the continuous gap instead of imposing a universal cross-task cutoff. *Instability* denotes broader variation over seed–task outcomes.

4 Seed-Conditional Comparison

The framework begins with a complete paired seed-by-task table and produces two complementary summaries. SRP compresses each seed to its weakest task recovery and describes method-level lower-tail exposure. CSP operates on paired method differences and describes whether favorable practical margins persist on the unreported complement. Neither summary replaces the underlying table or assumes that the two objects form a causal chain.

4.1 Seed Recovery Profile

For each method, seed, and task, let $a_{m,s,t}$ be accuracy, F_t a task floor, and $H_t > F_t$ a healthy reference, on the same scale as $a_{m,s,t}$. For profile summaries, define $r_{m,s,t} = \mathrm{clip}((a_{m,s,t} - F_t)/(H_t - F_t), 0, 1)$. F_t is the empirical majority-label accuracy for binary tasks and the mean uniform-choice accuracy for variable-choice tasks. Before observing Stage-2 outcomes, we freeze H_t as the 75th percentile of pooled canonical-method accuracies in a frozen development registry (development seeds $7, 42, 123, 2024, 3407$), rather than recomputing it from the retained 20-seed cell table. We summarize a seed by its weakest task recovery,

$$R_{m,s} = \min_{t \in \mathcal{T}} r_{m,s,t}. \tag{1}$$

Headroom scaling gives a fixed drop more leverage when $H_t - F_t$ is small (Appendix D), and clipping maps all at- or below-floor outcomes to zero; raw accuracy therefore remains necessary for sub-floor severity. The empirical survival curve $\Pr_s[R_{m,s} \geq \tau]$ over thresholds $\tau \in [0, 1]$ is the SRP. Its area, lower-CVaR_{10}, and worst-seed value are continuous lower-tail summaries. We report the full curve rather than selecting a single cutoff, pointwise Wilson intervals in the appendix, and seed-bootstrap intervals for AUC and CVaR.

4.2 Claim Survival Profile

Let $U(m, s) = 100|\mathcal{T}|^{-1} \sum_{t \in \mathcal{T}} a_{m,s,t}$ be the pre-specified, unweighted mean percentage accuracy for method m at seed s. Let Ω be the retained seed panel, $S \subset \Omega$ a reporting subset of k seeds, and $\bar{\Delta}_{m,b}(S)$ the mean paired utility advantage $|S|^{-1} \sum_{s \in S}[U(m, s) - U(b, s)]$. For a practical margin δ, the claim survival profile of the directional statement "m exceeds b" is

$$\mathrm{CSP}_{m,b}(k, \delta) = \Pr_{S:|S|=k} \left[\bar{\Delta}_{m,b}(\Omega \setminus S) \geq \delta \;\middle|\; \bar{\Delta}_{m,b}(S) \geq \delta \right]. \tag{2}$$

The probability is uniform over all size-k subsets of the retained panel. We report exact counts at common budgets $k \in \{1, 3, 5\}$ and margins $\delta \in \{0, 0.25, 0.5, 1\}$ percentage points, and trace $k = 1, \ldots, 10$ at the pre-specified 0.5-point margin. Unlike an unpaired mean difference, CSP conditions on the event that a favorable claim was reportable, then checks that claim on seeds excluded from the report. We report both its numerator and denominator; CSP is undefined when no subset satisfies the conditioning event. It describes reporting-subset sensitivity within the retained finite panel, not a population replication probability or an independent held-out test. With $N = |\Omega|$, $\mu = \bar{\Delta}_{m,b}(\Omega)$, and $d_S = \bar{\Delta}_{m,b}(S)$, the complement mean is $(N\mu - kd_S)/(N - k)$ for $1 \leq k < N$. Thus $\mu < \delta$ forces CSP to zero whenever a report is possible, while an unusually large favorable subset can exhaust the panel's margin, as seeds 2 and 3 do in Figure 1(a). CSP depends on the subset-mean distribution, beyond its mean and variance: Appendix C gives two panels with identical moments and win/tie/loss counts but different CSP. Appendix E derives the full-panel relationship. Figure 1 illustrates the conditioning event and its complement using the retained LLaMA panel.

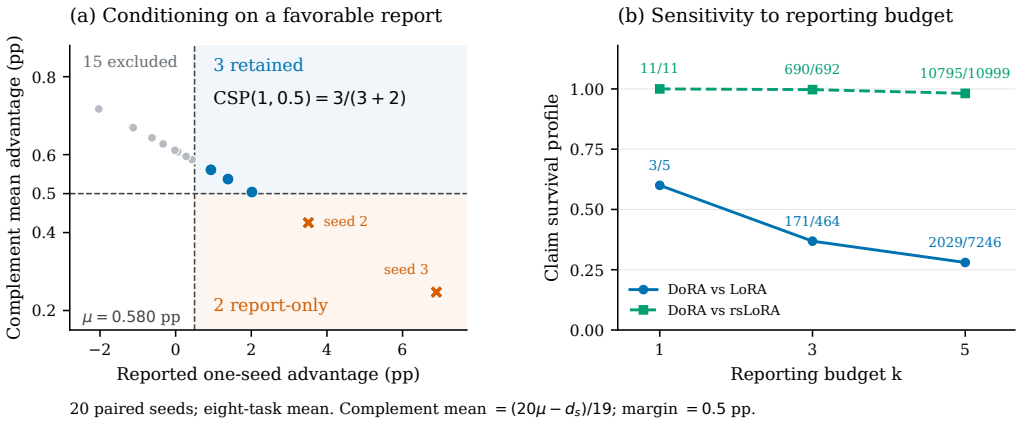


Figure 1: **How CSP summarizes favorable reports.** (a) DoRA versus LoRA, $k = 1$: each point pairs one seed's mean advantage with the mean on the other 19 seeds. Five reports reach the vertical 0.5-point threshold; three also reach the horizontal complement threshold (blue circles), while two do not (orange crosses). Thus CSP is $3/5$ at $\delta = 0.5$; at $\delta = 0$, all 13 favorable one-seed reports retain their direction. (b) Exact retained/reportable counts across reporting budgets for two comparisons. Complements overlap; the curves describe the observed panel, not new-seed replication probabilities.

5 Experimental Protocol

LLaMA panel. We fine-tune LLaMA-3-8B (Grattafiori et al., 2024) on Commonsense-170K (Hu et al., 2023) and evaluate eight commonsense tasks: BoolQ, HellaSwag, PIQA, SocialIQA, WinoGrande, ARC-Challenge, ARC-Easy, and OpenBookQA (Clark et al., 2019; Zellers et al., 2019; Bisk et al., 2020; Sap et al., 2019; Sakaguchi et al., 2020; Clark et al., 2018; Mihaylov et al., 2018). The canonical matched comparison uses rank 8, $\alpha = 16$, and the same training recipe for LoRA, DoRA, and rsLoRA. It spans 20 seeds in two retained ten-seed stages. Lower-learning-rate LoRA and rank-16 LoRA are controls, not canonical method winners.

Held-out Qwen panel. Qwen3-8B uses the same Commonsense-170K corpus, prompt format, and matched rank-8 recipe. After exploration, we freeze unused seeds 6–10, four tasks (BoolQ, HellaSwag, WinoGrande, PIQA), methods, hyperparameters, floors, and references before matching result directories exist. The three methods yield $3 \times 5 \times 4 = 60$ cells. Appendix A documents the preceding exploration panels, which are excluded from confirmation, and the complete protocol.

Documented-recipe audit. To distinguish the matched comparison from a published training recipe, we also run the documented DoRA and LoRA recipes on ten paired seeds and the same eight-task suite. This audit preserves each recipe's published rank-32, $\alpha = 64$, method-specific learning rates, and best-validation checkpoint selection. It tests the recipe pair, not adapter parameterization alone.

6 Results

6.1 Every canonical PEFT method has a vulnerable lower tail

Figure 2 displays all 20 seeds for BoolQ and HellaSwag; the appendix profiles cover the remaining six tasks. Figure 3 summarizes weakest-task recovery across all eight. Higher curves indicate that more seeds reach a given weakest-task recovery threshold. The failure patterns are heterogeneous. BoolQ contains repeated near-majority failures for LoRA, DoRA, and rsLoRA; HellaSwag contains large degradations for all three canonical methods; and rsLoRA seed 3 also degrades on PIQA and SocialIQA. Thus the issue is neither unique to binary tasks nor explained by one adapter rule.

Table 1 shows why mean accuracy alone is inadequate. DoRA has the highest rank-8 mean among the canonical methods (84.65%), yet its global minimum over all 20 seeds and eight tasks is 62.17%

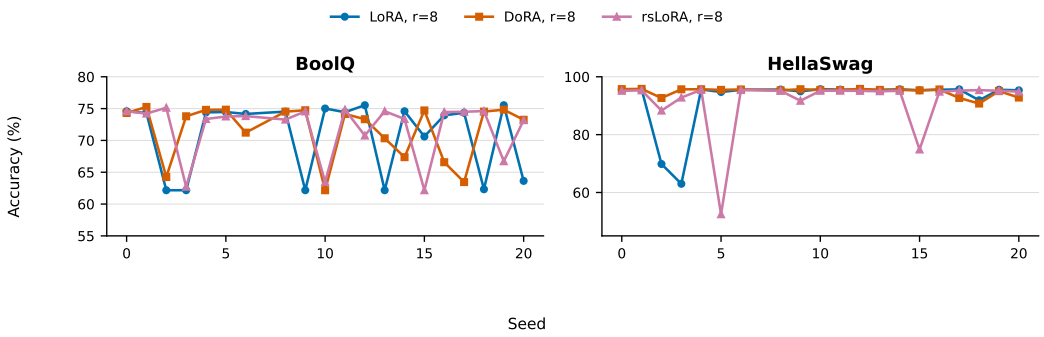


Figure 2: Twenty-seed LLaMA-3-8B outcomes on BoolQ and HellaSwag. Appendix A.6, Figure 8, provides all eight task-wise profiles.

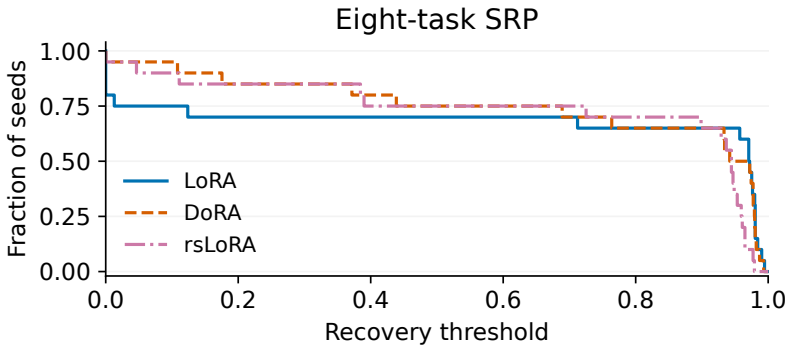


Figure 3: SRP across all eight LLaMA tasks: the fraction of seeds whose weakest-task recovery reaches each threshold. Appendix A.6, Figure 7, provides separate curves with pointwise intervals.

(Stage-1 seed 10 on BoolQ). RSLORA has a 52.47% global worst cell despite a comparable mean. LoRA and DoRA's identical minima are separate BoolQ runs at the 62.2% majority floor. LoRA has four zero-recovery seeds, giving lower-CVaR$_{10} = 0$; rsLoRA has a lower raw minimum but fewer floor-reaching seeds, giving lower-CVaR$_{10} = 0.023$. The wide seed-bootstrap intervals show that twenty seeds identify heterogeneous lower-tail exposure without resolving a population tail rate or method ordering. This conclusion is unchanged when the development-reference percentile varies from 50 to 90: min-based SRP still contains zero-recovery seeds for every method. It does depend, by design, on the task aggregation. A bottom-two mean retains substantial lower-tail exposure, whereas a 25th task quantile is less sensitive to failures confined to one or two of eight tasks; the SRP-AUC ordering also changes across aggregations. Appendix D reports the full grid. SRP is therefore a descriptive lower-tail profile, not a winner-ranking metric.

6.2 SMALL REPORTING BUDGETS MISS OBSERVED LOWER TAILS

The full profiles make lower tails visible. To quantify how much a small reporting budget can understate them, we enumerate every k-seed subset of the retained 20-seed LLaMA panel for every $k \in \{1, \ldots, 20\}$. Let $W_m = \min_s R_{m,s}$ denote method m's full-panel minimum recovery, and let $O_m(S) = \min_{s \in S} R_{m,s} - W_m \geq 0$ be a subset's *tail optimism*. This finite-panel enumeration describes what smaller reporting budgets would reveal within the retained panel.

Table 2 contrasts central mean-accuracy error with tail optimism. At one seed, the median absolute mean-accuracy error is only 0.43–0.90 points, yet 80–95% of subsets miss each method's full-panel minimum recovery (the zero-optimism rate is 0.05–0.20). At three seeds, 49% of LoRA subsets and 85% of DoRA and rsLoRA subsets still miss their observed worst recovery. Even at five seeds, 75% of DoRA and rsLoRA subsets miss it; their 90th-percentile optimism remains 0.764 and 0.899. In this PEFT panel, a small seed budget can make average accuracy look stable while materially overstating

Table 1: Twenty-seed, eight-task canonical reliability. Brackets are percentile seed-bootstrap 95% intervals (20,000 resamples); their width reflects only 20 seed observations. Higher SRP AUC and CVaR_{10} are better.

Method	Mean	Worst cell	SRP AUC [95%]	CVaR_{10} [95%]
LoRA (r=8)	84.07	62.17	0.678 [0.484, 0.860]	0.000 [0.000, 0.013]
DoRA (r=8)	84.65	62.17	0.758 [0.603, 0.892]	0.054 [0.000, 0.405]
rsLoRA (r=8)	83.55	52.47	0.748 [0.588, 0.884]	0.023 [0.000, 0.387]

Table 2: All-subset reporting-budget analysis in the retained 20-seed LLaMA panel. Med. error is median absolute error in eight-task mean accuracy across subsets. Tail optimism is sampled minimum recovery minus full-panel minimum recovery; zero means that the subset contains an equally bad observed seed. "Zero" is its probability over all subsets.

	$k = 1$			$k = 3$			$k = 5$		
Method	Med. error	P90 opt.	Zero	Med. error	P90 opt.	Zero	Med. error	P90 opt.	Zero
LoRA	0.90	0.985	0.200	0.77	0.972	0.509	0.58	0.712	0.718
DoRA	0.43	0.982	0.050	0.25	0.972	0.150	0.17	0.764	0.250
rsLoRA	0.87	0.966	0.050	0.84	0.945	0.150	0.56	0.899	0.250

the observed lower tail. The artifact reports the complete enumeration, including 90th-percentile mean-accuracy errors. For an isolated minimum, the miss probability is mechanically $1 - k/20$. The budget required to characterize this lower tail is method-dependent. At an illustrative P90 tail-optimism tolerance of 0.10, the first qualifying budgets are 7 seeds for LoRA, 18 for DoRA, and 14 for rsLoRA. These values summarize the retained panel rather than prescribe a universal seed budget. A smaller budget here means that the observed low tail is easier to encounter. Repeated poor outcomes can therefore reduce this budget; the budget itself does not rank methods by robustness.

6.3 A favorable seed claim can be sensitive to its reporting subset

The paired comparison is central. In the complete 20-seed panel, LoRA exceeds DoRA by at least 0.5 points on four seeds, while DoRA exceeds LoRA on five; eleven seeds are practical ties. LoRA likewise exceeds rsLoRA by at least 0.5 points on 13 seeds and trails by at least 0.5 points on six, with one practical tie. These seed-wise differences show that a reported practical comparison can support either direction when the full-panel average difference is small. At the strict sign level, 7 of 20 one-seed LoRA–DoRA comparisons favor LoRA, whereas DoRA has the higher 20-seed suite mean by 0.58 points. We report $\delta = 0$, which asks whether the direction is retained, alongside the pre-specified $\delta = 0.5$ practical-effect operating point. CSP makes the reporting risk concrete: it conditions on a reportable directional claim and records whether the same margin also holds in aggregate on the omitted seeds.

At $\delta = 0$, DoRA-over-LoRA retains its direction for 13/13 favorable one-seed reports, 788/790 three-seed reports, and 11,623/11,834 five-seed reports. At $\delta = 0.5$, the corresponding counts are 3/5, 171/464, and 2,029/7,246. This decline reflects a tightening finite-sum bound when the full-panel mean (0.58 points) lies close to δ, not evidence that larger independent samples are intrinsically less reliable. This example therefore exposes sensitivity of the *effect magnitude*, rather than frequent reversal of the complement mean. Table 3 reports every directed pair at $\delta = 0.5$ points, while Figure 5 traces the three full-panel-favored directions over reporting budgets $k = 1, \ldots, 10$. DoRA-over-rsLoRA retains it for 11/11, 690/692, and 10,795/10,999 reports, respectively. Although both comparisons favor DoRA in the full-panel mean, their advantages differ substantially in sensitivity to the reporting subset. In the opposite direction, LoRA-over-DoRA is reportable on four individual seeds but on none of their complements. This zero follows from the negative full-panel gap, as derived in Appendix E.

CSP thus describes reporting-subset sensitivity at a stated effect margin. It measures a different object from a bootstrap interval or win fraction; it is not proposed as a universally superior declaration rule.

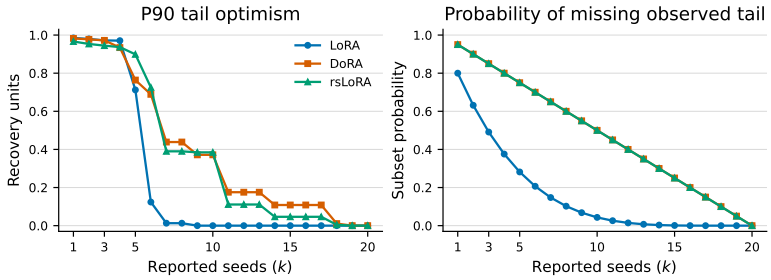


Figure 4: Exact within-panel reporting-budget enumeration over all $\binom{20}{k}$ subsets of the LLaMA panel. A small seed budget can leave the average accurate while missing the observed lower tail. Curves describe this retained panel, not an unobserved population distribution.

Table 3: Exact claim survival in the retained 20-seed LLaMA panel at a 0.5-point margin. Each cell is surviving report subsets / reportable subsets; all six directed method comparisons are shown.

Reported directional claim	$k = 1$	$k = 3$	$k = 5$
LoRA > DoRA	0/4	0/111	0/510
DoRA > LoRA	3/5	171/464	2,029/7,246
LoRA > rsLoRA	9/13	121/548	993/7,616
rsLoRA > LoRA	0/6	0/173	0/1,401
DoRA > rsLoRA	11/11	690/692	10,795/10,999
rsLoRA > DoRA	0/1	0/16	0/2

Appendix F separately examines a possible use as a declaration rule, including comparisons with paired means, win fractions, and bootstrap bounds.

6.4 A PROSPECTIVELY FROZEN QWEN PANEL CONTAINS LOWER-TAIL EVENTS

For the Qwen seed holdout, seeds 6–10 and the four-task evaluation panel were frozen before training; Appendix A records the preceding exploratory panels and the scope of this holdout. Figure 6 reports every resulting cell. LoRA seed 6 averages 29.85% across the four tasks: 65.26% on BoolQ, 20.79% on HellaSwag, 6.63% on WinoGrande, and 26.71% on PIQA. Under the same seed and matched rank-8 recipe, DoRA and rsLoRA average 82.48% and 85.71%, respectively. The panel also contains a method-local lower tail: DoRA seed 9 falls to 57.19% on BoolQ while its other three tasks remain near their peers. rsLoRA has no comparable event in these five held-out seeds. This small panel does not estimate method-specific failure frequencies. It documents recurrence of lower-tail events and seed-conditional rankings under a pre-specified protocol on a second 8B backbone.

An aggregate output-validity check confirms that the largest drops coincide with malformed or concentrated outputs; Appendix A.3 reports the diagnostics and shows that final validation loss does not flag the largest event.

6.5 A SECOND TRAINING CORPUS SHOWS LOWER-TAIL VARIATION AND RANKING CHANGES

We next hold the LLaMA-3-8B family and rank fixed, fine-tune on the official GSM8K training split (Cobbe et al., 2021), and evaluate five prospectively fixed seeds on GSM8K, MultiArith (Roy & Roth, 2015), SVAMP (Patel et al., 2021), SingleEq (Koncel-Kedziorski et al., 2015), AddSub (Hosseini et al., 2014), and AQuA (Ling et al., 2017). This 90-cell breadth panel extends the evidence to continuous lower-tail variation and seed-wise ordering under a second corpus. The three methods are close in mean accuracy (LoRA 58.78%, DoRA 58.57%, rsLoRA 58.83%), but their paired orderings change across the five seeds. The largest leave-one-out gaps are 7.87 points for LoRA (SingleEq, seed 3), 6.00 for DoRA (SingleEq, seed 5), and 5.61 for rsLoRA (SingleEq, seed 3). These are material lower-tail observations in this five-seed panel; estimating their frequency requires a larger panel.

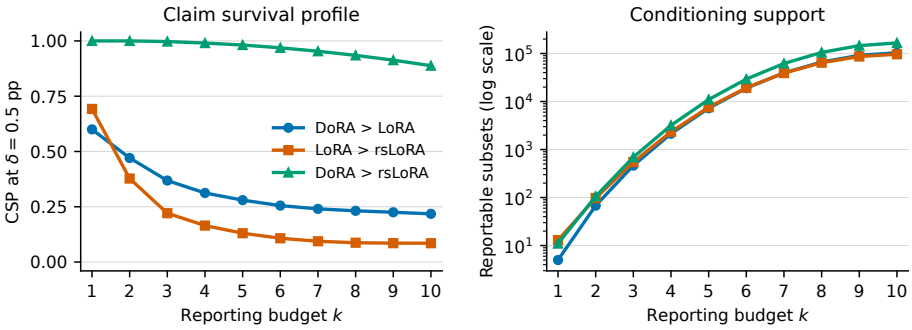


Figure 5: Exact CSP profiles and conditioning support in the retained 20-seed LLaMA panel at $\delta = 0.5$ points. We orient each unordered pair according to its full-panel mean. Left: fraction of reportable k-seed subsets whose complement retains the margin. Right: number of reportable subsets entering each fraction (log scale). Counts are exact finite-panel enumerations, not independent trials.

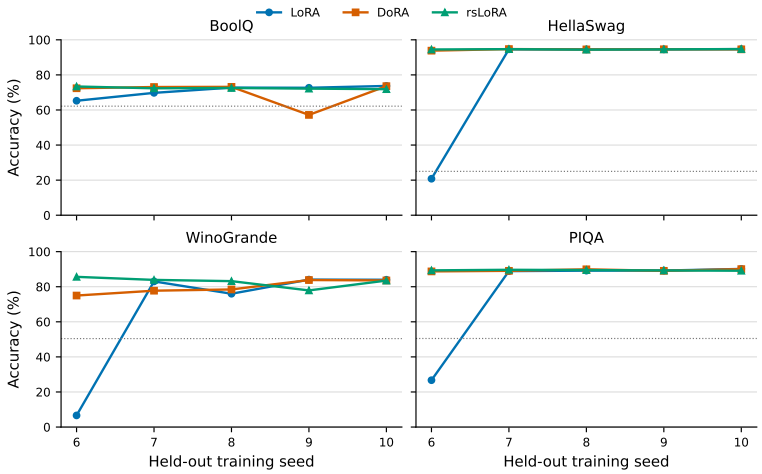


Figure 6: Prospectively frozen Qwen3-8B panel. Seeds 6–10, rank-8 methods, and four tasks were frozen before outcomes were observed. Dotted lines mark the pre-specified task floors. The LoRA seed-6 failure is cross-task; DoRA seed 9 has a BoolQ-local lower tail.

For AQuA, every cell uses the same fixed 249-example numeric-option scoring scope; the five excluded items have non-numeric or ambiguous numeric option formatting. Appendix A.4 lists their indices.

6.6 A DOCUMENTED DORA RECIPE PROVIDES A STABLE BOUNDARY CASE

The separate ten-seed documented-recipe audit shows that a directional comparison can remain stable under paired multi-seed evaluation. DoRA reaches 85.44% mean accuracy versus 80.81% for the accompanying LoRA recipe, and neither method reaches a severe lower-tail endpoint. The paired mean advantage is 4.62 points (95% paired bootstrap interval: 4.32–5.07), and the seed-42 ordering agrees with seeds 43–51. This recipe-pair result provides a stable boundary case for the matched comparison.

6.7 AGGREGATE MULTI-SEED REPORTING STILL OMITS TAIL BEHAVIOR

Our descriptive audit covers 37 PEFT papers from 2019–2025. The frozen stratified frame, coding procedure, exclusions, and evidence locations appear in Appendix H.

While 24/37 papers report results from more than one independent training RNG somewhere in the paper and 22/37 report some uncertainty, only 1/37 exposes downstream per-seed outcomes or a worst-seed/tail statistic. Only 5/37 state a seed-selection rule or rationale. Among the 31 papers with accessible official training code, 25 use a fixed or default seed without a stated selection rationale. These observations show that aggregate reporting in this audit often cannot reveal the seed-task failures measured by SRP or test a claim with CSP.

7 Discussion

The resulting reporting hierarchy starts with seed identifiers, complete seed-by-task outcomes, the aggregation rule, and paired differences. SRP and CSP compactly expose worst-task recovery and margin sensitivity, but paired intervals do not resolve a tail-summary winner (Appendix B), and retrospective tests do not establish either profile as a superior decision rule (Appendix F). Under compute limits, authors should release every paired cell and avoid unsupported margin-stability claims.

8 Limitations

The 20-seed canonical panel covers one fine-tuning corpus and eight commonsense benchmarks. The mathematics panel adds a second corpus, while the Qwen3-8B panel adds a second backbone, but each contains only five seeds. These panels therefore support recurrence and scope checks rather than universal failure frequencies or adapter rankings. The profiles also require multi-seed evaluation and do not identify the causal source of a lower-tail event; the Qwen panel further shows that ordinary final validation loss need not flag one. Because H_t is development-relative, raw task-wise accuracy remains necessary when development methods share a task-level failure. Finally, lower-tail behavior may change with the training recipe, and the Qwen confirmation holds out seeds rather than all possible task subsets.

9 Conclusion

Average accuracy alone can obscure two aspects of PEFT evaluation: method-level lower-tail behavior and the seed sensitivity of practical method comparisons. SRP summarizes the former, while CSP summarizes the reporting-subset sensitivity of the latter. A complete 20-seed LLaMA panel shows that close means can coexist with heterogeneous lower tails and margin-sensitive claims; a prospectively frozen Qwen panel documents severe seed-indexed differences on a second backbone. The resulting recommendation is concrete: report complete paired seed-wise outcomes first, then use lower-tail and claim-sensitivity profiles to summarize the aspects of that record that average accuracy omits.

AI use disclosure. Generative AI assisted with research ideation and execution, literature retrieval, code and plotting, and drafting and editing. The authors verified the sources, code, results, and claims and take responsibility for the manuscript and artifacts.

REFERENCES

Robert Belanec, Branislav Pecher, Ivan Srba, and Maria Bielikova. PEFT-Bench: A parameter-efficient fine-tuning methods benchmark. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, 2026. doi: 10.18653/v1/2026.eacl-long.140. URL https://aclanthology.org/2026.eacl-long.140/.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077. URL https://arxiv.org/abs/1306.1059.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of AAAI*, 2020. URL https://arxiv.org/abs/1911.11641.

Nghia Tuan Bui, Guergana K. Savova, and Lijing Wang. Assessing the macro and micro effects of random seeds on fine-tuning large language models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 41–46, 2025. doi: 10.18653/v1/2025.ijcnlp-short.3. URL https://aclanthology.org/2025.ijcnlp-short.3/.

Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. Revisiting parameter-efficient tuning: Are we really there yet? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2612–2626, 2022. doi: 10.18653/v1/2022.emnlp-main.168. URL https://aclanthology.org/2022.emnlp-main.168/.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL*, 2019. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

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.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020. URL https://arxiv.org/abs/2002.06305.

Yupei Du and Dong Nguyen. Measuring the instability of fine-tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6209–6230, 2023. doi: 10.18653/v1/2023.acl-long.342. URL https://aclanthology.org/2023.acl-long.342/.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 523–533, 2014. doi: 10.3115/v1/D14-1058. URL https://aclanthology.org/D14-1058/.

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 *International Conference on Machine Learning*, 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.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. LLM-Adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.319. URL https://aclanthology.org/2023.emnlp-main.319/.

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.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597, 2015. doi: 10.1162/tacl_a_00160. URL https://aclanthology.org/Q15-1042/.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NjNfLdxr3A.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. doi: 10.1214/15-AOS1371. URL https://arxiv.org/abs/1311.6238.

Yu-Ang Lee, Ching-Yun Ko, Pin-Yu Chen, and Mi-Yen Yeh. Learning rate matters: Vanilla LoRA may suffice for LLM fine-tuning. *arXiv preprint arXiv:2602.04998*, 2026. URL https://arxiv.org/abs/2602.04998.

Shiwei Li, Xiandi Luo, Xing Tang, Haozhao Wang, Hao Chen, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. Beyond zero initialization: Investigating the impact of non-zero initialization on LoRA fine-tuning dynamics. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 35519–35535, 2025. URL https://proceedings.mlr.press/v267/li25bm.html.

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*, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 158–167, 2017. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015/.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0cde695b83bd186c1fd456302888454-Abstract-Conference.html.

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 *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 61–68, 2022b. doi: 10.18653/v1/2022.acl-short.8. URL https://aclanthology.org/2022.acl-short.8/.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.04647.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of EMNLP*, 2018. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nzpLWnVAyah.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 2080–2094, 2021. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

Branislav Pecher, Jan Cegin, Robert Belanec, Jakub Simko, Ivan Srba, and Maria Bielikova. Fighting randomness with randomness: Mitigating optimisation instability of fine-tuning using delayed ensemble and noisy interpolation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11005–11044, 2024. doi: 10.18653/v1/2024.findings-emnlp.644. URL https://aclanthology.org/2024.findings-emnlp.644/.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348, 2017. doi: 10.18653/v1/D17-1035. URL https://aclanthology.org/D17-1035/.

Subhro Roy and Dan Roth. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1743–1752, 2015. doi: 10.18653/v1/D15-1202. URL https://aclanthology.org/D15-1202/.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of AAAI*, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQA: Commonsense reasoning about social interactions. In *Proceedings of EMNLP*, 2019. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Yize Wu, Ke Gao, Ling Li, and Yanjun Wu. Stable-LoRA: Stabilizing feature learning of low-rank adaptation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xSa19DAieH.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of ACL*, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

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://arxiv.org/abs/2303.10512.

A Protocol and Reproducibility Details

A.1 Canonical LLaMA Panel and Randomness

All LLaMA canonical adapters use rank 8, $\alpha = 16$, learning rate 10^{-4}, and the same Commonsense-170K training and generative evaluation recipe. The 20 seeds are retained as Stage 1 $\{0, 1, 2, 3, 4, 5, 6, 8, 9, 10\}$ and Stage 2 $\{11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. A separate frozen development registry (seeds $7, 42, 123, 2024, 3407$; pooled canonical methods) supplies the 75th-percentile task references H_t; its run directories, checkpoints, and prediction records are disjoint from the retained panel. In particular, seed 7 is not a retained seed. Neither retained stage recomputes H_t. The protocol records task SHA-256 hashes, task floors, reference construction, training and evaluation commands, saved prediction hashes, and per-cell runtimes. For the canonical LLaMA panel, the entry-point seed controls Python, NumPy, and Torch state before model and adapter construction, the train–validation split, and a pre-Trainer dataset shuffle. The frozen DoRA-derived entry point retains the Transformers Trainer default seed 42; GPU kernels are not forced to be deterministic. Thus "seed" denotes this documented bundle of stochastic choices rather than an independent intervention on every random source. Evaluation generation is deterministic. Machine-readable cell-level results, protocol manifests, prediction hashes where retained, and analysis scripts are included with the submission artifact.

Table 4: Frozen LLaMA normalization inputs. H_t is the pooled development-registry 75th percentile over canonical methods and seeds $7, 42, 123, 2024, 3407$; it is not recomputed from the retained 20-seed cell table.

Task	F_t	H_t
BoolQ	0.622	0.740
HellaSwag	0.250	0.954
PIQA	0.505	0.885
SocialIQA	0.333	0.811
WinoGrande	0.504	0.854
ARC-Challenge	0.250	0.799
ARC-Easy	0.250	0.905
OpenBookQA	0.250	0.864

A.2 Qwen Holdout Protocol

For Qwen, historical targeted seeds $42, 123, 2024$, a frozen seeds 1–5 two-task panel, and a completed seeds 11–15 breadth panel precede the seed-6–10 holdout. The holdout uses the previously unrun consecutive integers 6–10; its four-task subset, all method and training choices, task floors, and references were frozen before matching seed-6–10 result directories existed. Qwen uses the 170,420-record Commonsense-170K file, with 120 records held out for validation and 170,300 used for training. The released `generate_prompt` function uses an instruction-following preamble and `### Instruction:`, optional `### Input:`, and `### Response:` sections; training appends the target output to the response section. The frozen Qwen references for BoolQ, HellaSwag, WinoGrande, and PIQA are 0.7324, 0.9467, 0.8398, and 0.8958, respectively. SRP values from this four-task panel and the eight-task LLaMA panel are not directly comparable: the task set changes the minimum's distribution, and each backbone uses its own frozen references.

A.3 Qwen output-validity check

For LoRA seed 6, only 23.20%, 8.68%, and 31.50% of generations parse to an allowed label on HellaSwag, WinoGrande, and PIQA, respectively; on BoolQ, 89.24% of all predictions concentrate on `true`. For the task-local DoRA seed-9 BoolQ event, parse validity falls to 77.77%, compared with 100% at its other four seeds. The artifact provides all 60 cells with accuracy, parse validity, dominant-label share, normalized label entropy, and prediction-file hashes. These diagnostics validate the reported scores and characterize their output-level failure patterns.

Final validation loss also fails to distinguish the largest event. LoRA seed 6 has final validation loss 1.0301 despite its 29.85% four-task mean, while the paired DoRA run has loss 1.0275 and

82.48% mean accuracy. Across the 15 held-out runs, the descriptive Pearson correlation between final validation loss and four-task mean accuracy is -0.21. Each run contributes 399 validation-loss measurements, and the artifact includes the complete run-level table.

A.4 Mathematics Panel and AQuA Scoring

The mathematics panel likewise freezes seeds 1–5 before evaluation. Its training file is a deterministic instruction-format conversion of all 7,473 official GSM8K training records: questions and final answers match the upstream split exactly, inline calculator annotations are removed, and the final-answer marker is verbalized. The protocol records the upstream and derived SHA-256 hashes and the upstream MIT license.

For AQuA, all methods and seeds are scored on the same 249 examples whose options admit an unambiguous numeric mapping. Five examples (indices 43, 136, 150, 155, and 177) have non-numeric or ambiguous numeric option formatting and are excluded identically from every comparison. This common scoring scope changes no training run, seed, generation, or method choice.

A.5 Compute and Software

Each final LLaMA and Qwen training run is allocated one NVIDIA A800 GPU. The held-out Qwen confirmation comprises 15 completed training-and-evaluation runs and consumed 284.73 A800 GPU-hours, measured from scheduler elapsed times; queue time is excluded. The protocol manifests record the training and evaluation options, data hashes, and a package-environment hash. The anonymous artifact includes the analysis dependencies, training and evaluation entry points, and portable rerun instructions; it does not redistribute model weights or benchmark data.

A.6 Artifact and Complete Reporting Objects

The artifact includes the full $3 \times 20 \times 8 = 480$ canonical LLaMA cell table, all eight task-wise seed curves, Qwen's 60-cell held-out confirmation results and aggregate output-validity diagnostics, and the 90-cell mathematics panel. The canonical table contains accuracy, normalized recovery, stage and seed identifiers, and prediction hashes; the Qwen and mathematics tables contain the fields documented in their headers. The full curves are the primary reporting object; Table 1 and Table 3 are compact summaries rather than replacements for those records.

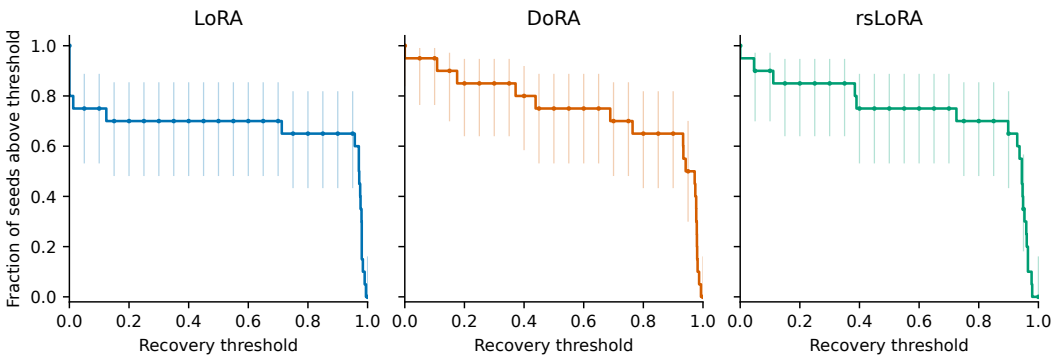


Figure 7: Complete canonical LLaMA seed recovery profiles. Dots show 21 thresholds with pointwise 95% Wilson intervals; steps use every observed recovery value. The intervals overlap substantially across methods.

A.7 Panel Scope and Generalization

Our primary evidence concerns rank-8 LoRA, DoRA, and rsLoRA adapters on two 8B decoder backbones. The LLaMA Commonsense-170K panel is the only 20-seed, eight-task canonical comparison; the Qwen confirmation has five held-out seeds and four tasks, and the mathematics panel

Table 5: Panel scope across frozen evaluations. The supported-conclusion column matches each empirical claim to its protocol.

Panel	Backbone / corpus	Tasks and seeds	Frozen before outcomes?	Supported conclusion
Canonical comparison	LLaMA-3-8B / Commonsense-170K	8 tasks, 20 seeds, 3 matched methods	Stage-2 outcomes; references fixed from a disjoint development registry	Complete paired panel exhibits lower-tail events and reporting-subset-sensitive practical margins.
Qwen recurrence	Qwen3-8B / Commonsense-170K	4 tasks, 5 consecutive seeds, 3 matched methods	Yes: seeds, tasks, methods, and references	Lower-tail events recur under this pre-specified second-backbone protocol.
Mathematics breadth	LLaMA-3-8B / mathematical reasoning corpus	6 tasks, 5 seeds, 3 matched methods	Seeds fixed prospectively	Lower-tail variation and seed-wise ordering changes also occur under a second corpus; no frequency claim.
Documented recipe audit	LLaMA-3-8B / published recipe pair	8 tasks, 10 paired seeds, LoRA and DoRA recipes	Seed 42 plus nine consecutive seeds, 43–51	This particular, non-controlled recipe pair retains its directional advantage over the audited seeds.

has five seeds. Accordingly, neither the reported lower-tail frequency nor a tail-summary ordering should be generalized beyond these panels without additional matched multi-seed studies.

B Configuration Boundaries and Paired Intervals

Table 6 reports two completed LoRA controls from the same 20-seed, eight-task LLaMA panel. They change one configuration from canonical rank-8 LoRA and are evidence that lower-tail behavior is recipe-sensitive, not matched competitors in the canonical three-method comparison. Rank-16 LoRA improves the observed tail; lower learning rate improves mean and SRP AUC but retains a zero-recovery tail.

Table 6: Completed LLaMA LoRA controls. Recovery values are clipped to $[0, 1]$ for profile summaries. All rows use final-step checkpoints.

Configuration	Mean acc.	SRP AUC	CVaR_{10}	Worst recovery
LoRA $r = 8$, lr 10^{-4}	84.07	0.678	0.000	0.000
LoRA $r = 8$, lr 5×10^{-5}	84.46	0.748	0.000	0.000
LoRA $r = 16$, lr 10^{-4}	84.70	0.885	0.236	0.124

The documented-recipe audit in Section 6 changes several factors: both methods use rank 32 and $\alpha = 64$, restore the minimum-validation-loss checkpoint, and follow method-specific learning rates (3×10^{-4} for LoRA and 10^{-4} for DoRA). It is therefore a comparison of documented recipes, not a controlled attribution to adapter parameterization.

Table 7 replaces comparison of overlapping method-wise intervals with paired seed-bootstrap intervals for metric differences. All SRP-AUC and CVaR_{10} difference intervals include zero. Only the DoRA–rsLoRA mean-accuracy difference excludes zero in this descriptive bootstrap analysis.

C Interpreting the Reporting Summaries

CSP as a reporting summary. The complete paired seed-wise table remains the primary record and is sufficient to recompute every CSP value. CSP makes explicit a reporting decision that a mean,

Table 7: Canonical LLaMA paired differences with percentile 95% seed-bootstrap intervals (20,000 paired resamples). Positive values favor the row method.

Method pair	Metric	Difference	Low	High
LoRA–DoRA	Mean accuracy (pp)	-0.580	-1.479	0.130
LoRA–DoRA	SRP AUC	-0.080	-0.326	0.170
LoRA–DoRA	CVaR_{10}	-0.054	-0.405	0.006
LoRA–rsLoRA	Mean accuracy (pp)	0.526	-0.292	1.387
LoRA–rsLoRA	SRP AUC	-0.070	-0.317	0.166
LoRA–rsLoRA	CVaR_{10}	-0.023	-0.387	0.006
DoRA–rsLoRA	Mean accuracy (pp)	1.106	0.310	2.050
DoRA–rsLoRA	SRP AUC	0.010	-0.195	0.222
DoRA–rsLoRA	CVaR_{10}	0.031	-0.276	0.372

standard deviation, or a single worst-cell statistic leaves implicit. Specifically, it conditions on size-k reporting subsets whose mean paired advantage reaches a margin δ, and measures the fraction for which the complementary seeds retain that margin. The table provides the underlying evidence, while CSP gives a compact, selection-conditioned summary of that record. Prospectively frozen seed panels add evidence about recurrence beyond the observed panel.

What the summary distinguishes. Consider two constructed six-seed vectors of paired accuracy differences, in percentage points:

$$d^A = (-2, -2, 0, 2, 5, 5), \qquad d^B = (-2, -2, 0, 3, 3, 6).$$

Both have mean $4/3$, sum of squares 62, and sample variance $154/15$. At margin $\delta = 1$, both have three wins, one tie, and two losses, with wins defined by $d_s \geq 1$ and losses by $d_s \leq -1$. For $k = 1$, the three favorable reports in panel A have complement means $(6/5, 3/5, 3/5)$, whereas those in panel B have complement means $(1, 1, 2/5)$. Thus CSP is $1/3$ for A and $2/3$ for B. The same average advantage, dispersion, and win count can therefore accompany different agreement between favorable reports and their remaining seeds. This constructed example isolates the quantity being summarized; it is not an experimental result or evidence of a superior decision rule. A complete paired-difference distribution also preserves this distinction.

Recommended reporting minimum. For a paired PEFT comparison, we recommend reporting the seed identifiers, complete seed-by-task outcomes, paired seed-wise differences, and the aggregation rule. Where lower-tail risk or a directional superiority claim is central, authors can additionally report a worst-task empirical profile (such as SRP) and a reporting-subset sensitivity summary (such as CSP), including its numerator and denominator. Raw cell-level results and prospectively frozen panels remain the basis for transparent reporting and evidence of recurrence, respectively.

D SRP Definition Sensitivity

SRP deliberately emphasizes the weakest tasks, so we test both components of its definition. We recompute the complete canonical panel using development-reference percentiles $50, 60, 70, 75, 80,$ and 90, crossed with the minimum, the mean of the bottom two task recoveries, and the 25th task quantile. For eight tasks, task-CVaR_{25} is exactly the bottom-two mean and is therefore not shown twice. Table 8 summarizes the grid; the artifact contains every method–definition value.

Changing the reference percentile rescales the profiles but does not change the AUC ordering within any aggregation. Under the minimum, every method has at least one zero-recovery seed throughout the reference grid. The bottom-two mean still assigns 5–25% of method–seed summaries below 0.5. In contrast, the 25th task quantile places none below 0.5 in this panel. With linear interpolation over eight sorted recoveries, it equals $0.25r_{(2)} + 0.75r_{(3)}$: the second-lowest task still contributes, while the minimum does not. Aggregation also changes the AUC ordering from DoRA–rsLoRA–LoRA to DoRA–LoRA–rsLoRA. The empirical phenomenon is therefore robust to plausible reference choices but conditional on the lower-tail object being summarized; SRP should not be used as a standalone adapter ranking.

Table 8: SRP sensitivity over development-reference percentiles 50–90. AUC ranges cover all six reference choices. "Below 0.5" is the range across methods and references. Higher AUC is better, but its ordering is descriptive.

Task aggregation	AUC ordering	Method AUC ranges	Fraction below 0.5
Minimum	D>R>L	L [0.666,0.699]; D [0.738,0.822]; R [0.729,0.786]	0.15–0.30
Bottom-two mean	D>R>L	L [0.800,0.824]; D [0.856,0.906]; R [0.831,0.869]	0.05–0.25
25th quantile	D>L>R	L [0.962,0.977]; D [0.983,0.995]; R [0.951,0.969]	0.00

E Metric Estimation and Sensitivity

Dependence on the full-panel mean. Write $N = |\Omega|$, $\mu = \bar{\Delta}_{m,b}(\Omega)$, and $d_S = \bar{\Delta}_{m,b}(S)$, with $1 \le k < N$. Partitioning the paired differences gives

$$\bar{\Delta}_{m,b}(\Omega \setminus S) = \frac{N\mu - k d_S}{N - k}. \tag{3}$$

Consequently, whenever a reportable subset exists,

$$\mathrm{CSP}_{m,b}(k,\delta) = \frac{\#\{S : |S| = k,\ \delta \le d_S \le \delta + \frac{N}{k}(\mu - \delta)\}}{\#\{S : |S| = k,\ d_S \ge \delta\}}. \tag{4}$$

If $\mu < \delta$, the interval in the numerator is empty and CSP is zero. If $\mu = \delta$, only subsets with $d_S = \delta$ can contribute. If $\mu > \delta$, CSP is the fraction of reportable subsets that do not exceed this mean-dependent upper bound; failures occupy the extreme upper tail of reported advantages. Thus CSP is a transformation of the subset-mean distribution and full-panel mean. We report the claim counts alongside it to distinguish how often a claim arises from how the complement behaves. In particular, the LoRA-over-DoRA zeros in Table 3 follow from the full-panel mean; they provide no separate confirmation evidence.

Uncertainty and enumeration. Lower-CVaR$_{10}$ is the mean of the lowest $\lceil 0.1N \rceil$ ordered seed recoveries, counting repeated values with multiplicity. For $N = 20$, it is $(R_{(1)} + R_{(2)})/2$; bootstrap samples use the same rule. For bounded seed recoveries, the empirical survival-curve area satisfies $\mathrm{AUC}_m = \int_0^1 N^{-1} \sum_s \mathbf{1}[R_{m,s} \ge \tau]\, d\tau = N^{-1} \sum_s R_{m,s}$; we compute it directly as this mean. For SRP AUC and lower-CVaR$_{10}$, we resample the observed seeds with replacement 20,000 times and report percentile 95% intervals. These empirical bootstrap intervals describe sensitivity to resampling the retained seed distribution under an exchangeability approximation; they are not uncertainty intervals for the fully observed finite-panel statistic, nor do they make the seeds a random sample from a known population. For method comparisons, the same sampled seed indices are applied to both methods and all tasks. For CSP, we enumerate every reporting subset at $k \in \{1, 3, 5\}$ and report the surviving/reportable count for every directed pair and every margin. The six directional results at $\delta = 0.5$ are in Table 3; the complete margin sweep is included with the artifact. Because report subsets overlap, these counts are exact finite-panel descriptions, not independent Bernoulli observations.

For the reporting-budget analysis, we enumerate all $\binom{20}{k}$ subsets for every $k \in \{1, \ldots, 20\}$ without replacement. For each method and subset, we compute the absolute error in suite mean accuracy relative to the retained 20-seed panel and the tail-optimism quantity $O_m(S)$ from Section 6. This calculation describes what these smaller budgets would have shown within the observed panel; it does not extrapolate to unseen seeds.

E.1 Cross-Panel Direction and Practical-Margin Sensitivity

We retrospectively apply the same paired comparison to all four panels, using equal-weight task means within each panel and retaining both directions for every available method pair. Table 9 contrasts one-seed CSP at zero and 0.5-point margins alongside paired mean intervals and seed-level win/tie/loss counts. Changing the margin also changes the conditioning subsets. The accompanying analysis files provide a 0–1-point margin sweep in 0.025-point increments for $1 \le k \le \min(5, \lfloor N/2 \rfloor)$, keeping the complement at least as large as the report. Panels are not pooled; the shared margin is a sensitivity probe rather than a claim of equal practical value across task suites.

Table 9: Cross-panel reporting sensitivity ($k = 1$). Gaps and margins are in percentage points. Intervals are paired-seed percentile 95% bootstrap intervals (20,000 resamples), with limited resolution in five-seed panels. W/T/L counts use a 0.5-point margin. CSP entries are retaining/reportable subset counts, not independent trials; a dash denotes no reportable subset. L, D, and R denote LoRA, DoRA, and rsLoRA. The last panel compares complete published recipes.

Comparison	Mean gap [95% interval]	W/T/L	CSP, $\delta = 0$	CSP, $\delta = 0.5$
LLaMA canonical: 20 seeds, eight tasks				
L > D	-0.58 [-1.47, 0.13]	4/11/5	0/7	0/4
L > R	0.53 [-0.29, 1.41]	13/1/6	14/14	9/13
D > L	0.58 [-0.13, 1.47]	5/11/4	13/13	3/5
D > R	1.11 [0.31, 2.04]	11/8/1	15/15	11/11
R > L	-0.53 [-1.41, 0.29]	6/1/13	0/6	0/6
R > D	-1.11 [-2.04, -0.31]	1/8/11	0/5	0/1
Qwen confirmation: five seeds, four tasks				
L > D	-9.79 [-31.50, 2.27]	1/2/2	0/3	0/1
L > R	-11.25 [-33.71, 0.80]	2/0/3	0/2	0/2
D > L	9.79 [-2.27, 31.50]	2/2/1	1/2	1/2
D > R	-1.46 [-2.58, -0.25]	1/0/4	0/1	0/1
R > L	11.25 [-0.80, 33.71]	3/0/2	3/3	2/3
R > D	1.46 [0.25, 2.58]	4/0/1	4/4	4/4
Mathematics: five seeds, six tasks				
L > D	0.20 [-0.69, 1.07]	2/1/2	2/3	0/2
L > R	-0.05 [-1.17, 1.07]	2/1/2	0/2	0/2
D > L	-0.20 [-1.07, 0.69]	2/1/2	0/2	0/2
D > R	-0.25 [-1.79, 0.98]	2/1/2	0/3	0/2
R > L	0.05 [-1.07, 1.17]	2/1/2	1/3	0/2
R > D	0.25 [-0.98, 1.79]	2/1/2	1/2	0/2
Documented recipes: ten seeds, eight tasks				
L > D	-4.62 [-5.07, -4.32]	0/0/10	–	–
D > L	4.62 [4.32, 5.07]	10/0/0	10/10	10/10

The panels distinguish directional sensitivity from practical-margin sensitivity. For rsLoRA over LoRA on Qwen, direction holds on 3/3 favorable complements, but a 0.5-point advantage holds on 2/3; rsLoRA over DoRA retains both criteria on 4/4. The mathematics panel instead has weak directional persistence, and all pairwise full-panel gaps are below 0.5 points in magnitude. Its zero CSP values at that margin therefore follow from sum conservation, not additional replication evidence. The documented-recipe comparison retains DoRA's advantage throughout the evaluated margin and budget sweep. These small-panel contrasts illustrate different reporting patterns without establishing their population frequency or CSP's superiority as a decision rule.

F Retrospective Evaluation of Reporting Rules

We examine whether the profiles also help select methods or decide when to declare a practical advantage. This analysis uses the existing ten-seed Stage-1 panel to form decisions and the disjoint ten-seed Stage-2 panel to assess them across all eight tasks. Both panels had already been inspected when these rules were designed. Tables 1 and 3 describe the full panel; this retrospective analysis evaluates a decision use of the same observations.

Rules and comparison. We enumerate Stage-1 reporting subsets of size $k \in \{1, 3, 5, 10\}$ and use practical margins $\delta \in \{0, 0.25, 0.5, 1\}$ points. A mean-based rule declares an advantage when the paired subset mean is at least δ. A win-fraction rule requires at least 80% of its differences to reach δ. A bootstrap rule requires the fifth percentile of 4,000 paired seed-bootstrap means to reach δ. For the CSP-based rule, we compute one-seed-report CSP *inside the available Stage-1 subset*, using its other $k - 1$ seeds as each complement, and declare when this value reaches 0.5, 0.8, or 1. This internal

Table 10: Retrospective declaration rules at $k = 5$, $\delta = 0.5$ points. There are $252 \times 6 = 1,512$ subset–direction opportunities. The rules have different coverage; these operating points do not establish an equal-coverage ordering.

Rule	Declarations	Coverage	Stage-2 retention
Paired mean	584	38.62%	58.90%
CSP ≥ 0.5	478	31.61%	67.15%
CSP ≥ 0.8	274	18.12%	65.69%
CSP $= 1$	274	18.12%	65.69%
Paired bootstrap bound	129	8.53%	88.37%
Win fraction ≥ 0.8	192	12.70%	100.00%

profile differs from the 20-seed profile in Table 3. CSP and bootstrap rules are unavailable at $k = 1$. Every rule has access to the same reporting subset.

Declaration coverage is the fraction of subset–direction opportunities on which a rule declares. Margin retention is the fraction of declarations whose Stage-2 suite-mean difference reaches δ. Table 10 shows an example. The bootstrap bound is nominally one-sided 95%; small seed samples can impair its calibration. Our released sweep includes bootstrap quantiles $\{0.01, 0.025, 0.05, 0.1, 0.2\}$ and every distinct score cutoff for each rule.

Findings. CSP filtering increases retention relative to declaring every subset-mean advantage at these operating points, while reducing coverage. Stronger CSP filtering does not monotonically increase retention. CSP ≥ 0.8 and the bootstrap rule agree on 90.41% of decisions, largely through 1,238 shared abstentions. They share 129 declarations; CSP makes 145 additional declarations, of which 66 retain the margin. Accordingly, overall agreement is insufficient to establish equivalence between rules. Each directed comparison shares one Stage-2 mean across all reporting subsets, so these counts are correlated descriptions of this validation panel. At this operating point, CSP $\geq .8$ and CSP $= 1$ coincide because 1 is the only attainable internal one-seed CSP value in $[0.8, 1]$.

Matched coverage. Figure 9 compares score thresholds at the same expected declaration coverage. At a boundary score shared by several opportunities, we include each tied opportunity with a common probability independent of its Stage-2 outcome. For adjacent threshold counts (D_0, R_0) and (D_1, R_1) and target declaration count D, set $p = (D - D_0)/(D_1 - D_0)$ and report $[R_0 + p(R_1 - R_0)]/D$. This is a ratio of expected retained and declared counts, not a realized randomized experiment. Dots show attainable whole-tie thresholds; the intervening curves use this boundary randomization. The sweep varies each score's cutoff rather than fixing the declaration rules in Table 10.

At $k = 5$, $\delta = 0.5$, and 20% coverage, retention is 68.92% for CSP, 68.58% for the paired mean, 78.51% for the bootstrap score, and 94.26% for the win fraction. At 10% coverage, the corresponding values are 65.69%, 55.03%, 90.08%, and 100%; at 30%, they are 69.13%, 61.07%, 68.34%, and 90.93%. Thus the relative performance depends on coverage, and this retrospective panel does not establish CSP as a superior declaration score. The artifact includes all four margins and $k \in \{3, 5, 10\}$; $k = 10$ has only six directed opportunities. At $\delta = 1$, every retention curve is zero because the largest Stage-2 mean gap is 0.612 points; this margin cannot distinguish the scores on this panel.

In method selection, maximizing Stage-1 SRP AUC at $k = 5$ yields an average Stage-2 SRP AUC of 0.7340, compared with 0.7330 when selecting by mean accuracy. Its mean-accuracy regret relative to the observed Stage-2 winner is 0.1983 points, compared with 0.0706 for mean selection. These small tail gains do not establish superior selection performance. The profiles' demonstrated role in the main paper is to summarize observed tails and reporting sensitivity; the effectiveness of a decision rule derived from them depends on its operating point and evaluation panel.

G Relation to Fine-Tuning Instability Measures

Table 11 makes the distinction from Du & Nguyen (2023) explicit. Their seven quantities summarize dispersion or disagreement among repeated runs of a *single* fine-tuned method; none conditions on a directional comparison claim made on one seed subset and tests that claim on a disjoint seed

Table 12: Illustrative audit coding examples. Evidence locations are retained for every row in the released coded table.

Paper	Coding state	Recorded evidence
Compacter (2021) (Mahabadi et al., 2021)	Partial	Table 1 restarts flagged unstable settings with six seeds and reports the best result; Figure 4 reports five-seed mean and standard deviation only for a low-resource study.
P-Tuning v2 (2022) (Liu et al., 2022b)	Not reported	Its experimental sections and appendix state neither an independent training-seed count nor a multi-run aggregation for headline comparisons; released task scripts fix a task-specific seed.
LoRA (2022)	Partial	Task-dependent coverage: Appendix D reports five-seed medians for RoBERTa/DeBERTa and three-run GPT-2 means, while the GPT-3 discussion gives only typical deviations; the released RoBERTa launcher fixes training seed 0.

subset. CSP therefore complements rather than replaces prediction- or representation-level instability measures.

Table 11: Object-level comparison with the seven fine-tuning instability measures of Du & Nguyen (2023). Here i, j index runs, x examples, p_i class probabilities, $\hat{y}_i$ labels, and H_i^l layer-l representations.

Measure	Summary form	Comparison object
Score SD	$\mathrm{sd}_i(\mathrm{score}(M_i))$	Score dispersion
Pairwise disagreement	$\mathrm{mean}_{i<j,x}\ \mathbf{1}[\hat{y}_i(x) \neq \hat{y}_j(x)]$	Predicted labels
Fleiss' Kappa	$1 - \kappa(\{\hat{y}_i(x)\}_{i,x})$	Label consistency
Pairwise JSD	$\mathrm{mean}_{i<j,x}\ \mathrm{JSD}(p_i(x) \Vert p_j(x))$	Predictive distributions
SVCCA distance	$\mathrm{mean}_{i<j,l}[1 - \mathrm{CCA}(H_i^l, H_j^l)]$	Layer representations
Orthogonal Procrustes	$\mathrm{mean}_{i<j,l} \min_{R^\top R=I} \Vert H_j^l - H_i^l R \Vert_F^2$	Layer representations
Linear CKA distance	$\mathrm{mean}_{i<j,l}[1 - \mathrm{CKA}(H_i^l, H_j^l)]$	Layer representations
CSP (ours)	$\Pr_S[\bar{\Delta}_{m,b}(\Omega \backslash S) \geq \delta \mid \bar{\Delta}_{m,b}(S) \geq \delta]$	Favorable cross-method claims and their complements

The first seven measures compare runs within one method. CSP conditions on a favorable paired comparison and evaluates the same margin on complementary seeds.

H Descriptive PEFT Reporting Audit

The frozen audit protocol includes every candidate in three purposive strata meeting the eligibility rules: foundational adapter/prompt methods (2019–2022), low-rank or quantized methods (2022–2024), and recent accepted method papers (2025). It contains 37 included papers and records excluded candidates with one reason each. Every included row supplies its paper and code evidence locations. The coding states were assigned and rechecked by the authors in two passes; independent coder agreement was not measured. The states are *yes*, *no*, *partial*, and *not reported*; the latter does not assert that an experiment was not performed. The complete coded table and protocol are released with the artifact, together with the frozen candidate and exclusion registries, pass-two recheck log, and conservative derived-summary coding. The reported $24/37$ multi-seed count uses that derived coding: a paper counts only when it reports more than one independently trained run in at least one downstream experiment. We use this audit only to characterize its frozen frame, not to estimate field-wide prevalence or author intent.

Table 12 gives representative, directly checkable coding examples. These examples illustrate the coding rule rather than quantify the audit.

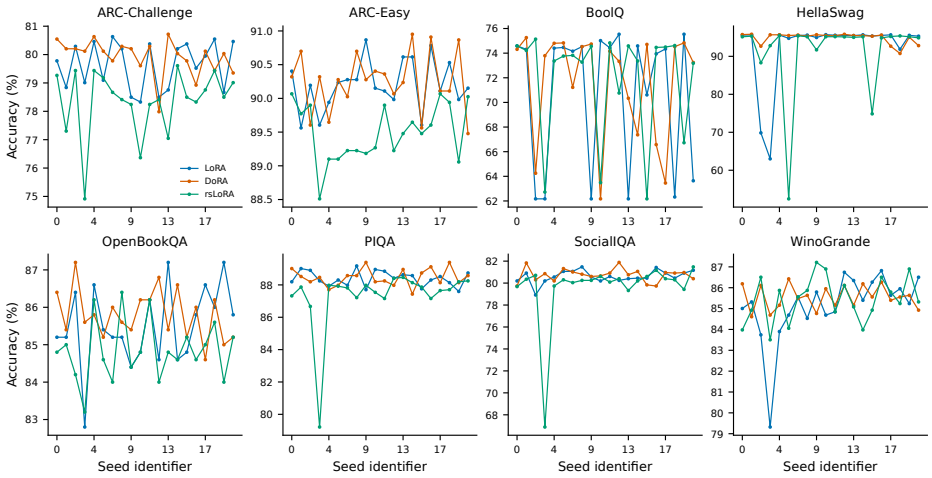


Figure 8: All eight canonical LLaMA task-wise seed curves for rank-8 LoRA, DoRA, and rsLoRA. Seed identifiers are paired across methods.

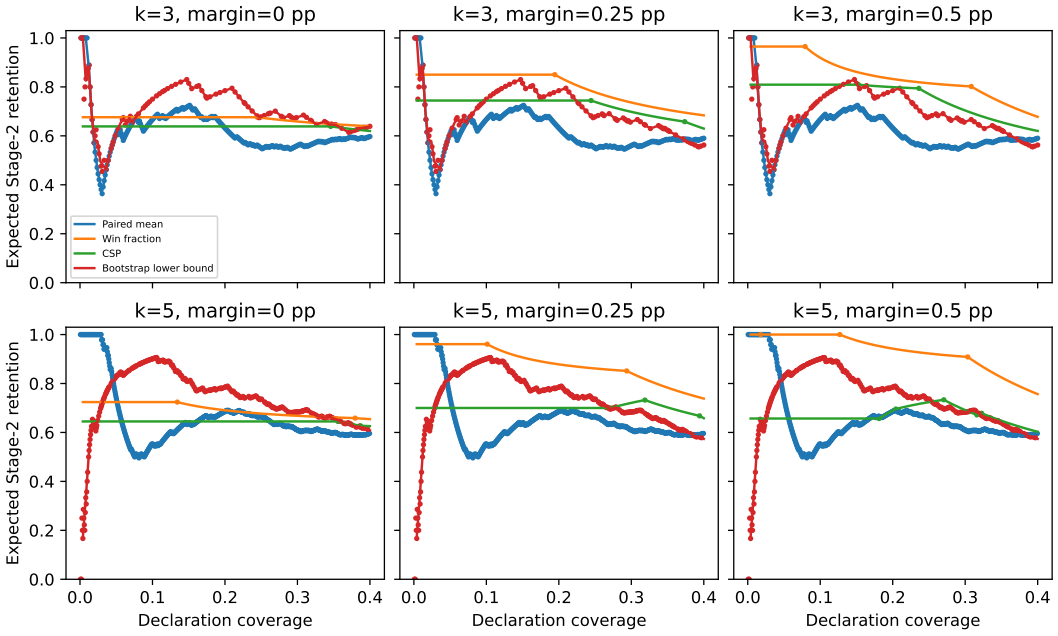


Figure 9: Retrospective matched-coverage comparison across reporting budgets and advantage margins. Curves use outcome-blind boundary randomization; dots retain whole score ties. All opportunities for a directed pair share its Stage-2 outcome, so the curves summarize this panel rather than independent confirmation trials.