BEYOND SCALAR BUDGETS: AUDITING RESOURCE SUFFICIENCY DECISIONS IN PEFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter-efficient fine-tuning (PEFT) seeks lower-resource configurations that meet a specified quality tolerance. We study whether simple resource rules preserve the sufficiency decisions of directly trained configurations. Specifically, we audit rank thresholds, combining separately tested parameter and data reductions, and equal-count interchangeability. Across tested settings, these rules can disagree with measured decisions: passing ranks need not admit a threshold, separately passing reductions can fail jointly, and equal-count configurations can receive different all-seed labels. We formulate a finite decision-validity audit that compares rule-implied decisions with same-seed paired measurements and quantifies departures from monotonicity on complete rank–data grids. Offline replays translate these violations into selection errors: missed lower-resource passing configurations and false-sufficient predictions. On prospectively frozen GSM8K grids, Qwen 1.5B and 7B configurations incur worst-seed losses below 1 pp versus their rank-eight/full-pool references, using one eighth of the adapter parameters and 3.85-fold or 1.93-fold fewer distinct examples visited. These results support validating resource rules against measured configurations rather than assuming scalar budgets preserve sufficiency.

1 INTRODUCTION

Given a target quality loss, practitioners need to decide how far they can reduce PEFT resources while retaining an acceptable trained model. A wrong shortcut has a direct cost: it can rule out a smaller configuration that works, or recommend a configuration whose measured loss exceeds the target. This decision is difficult because rank, data, parameterization, and the training recipe jointly determine an endpoint.

Three convenient rules simplify that decision. A rank threshold treats larger rank as sufficient once some minimum rank passes. A marginal rule combines separately measured parameter and data reductions to infer their intersection. A parameter-count rule treats configurations with the same trainable count as interchangeable. Each rule permits a different inference, and each can be checked against directly measured labels rather than assumed.

Prior studies have quantified how fine-tuning data, trainable parameters, rank, parameterization, and optimization affect performance (Zhang et al., 2024; Biderman et al., 2024). We ask whether a declared resource rule preserves the pass/fail decisions induced by a specified quality tolerance. This shifts the question from modeling how performance changes with resources to testing whether a resource rule supports the discrete selection decision it is used to make. The three properties are logically distinct and need not fail together. On the sealed ARC-Challenge test, validation-selected ranks yield pass–fail–pass–pass at 1 pp, while all four pass at 3 pp. Directly trained BERT/QNLI intersections show that separately passing parameter and data reductions can fail jointly; matched-count T5/QNLI configurations receive different all-seed labels at the same adapter count (Figure 2). Llama further separates order from composition: its rank–data map is an upper set at both predeclared tolerances, yet rank one/full-pool and rank eight/half-pool pass while their rank-one/half-pool intersection fails.

We compare rule-implied labels with same-seed measurements, quantify departures from monotonicity on complete rank–data grids, and leave untested cells unlabeled. Learning-rate and toler-

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.

ance analyses expose decision sensitivity. Direct GSM8K evaluations identify low-resource configurations within 1 pp of their references, with fewer distinct examples visited.

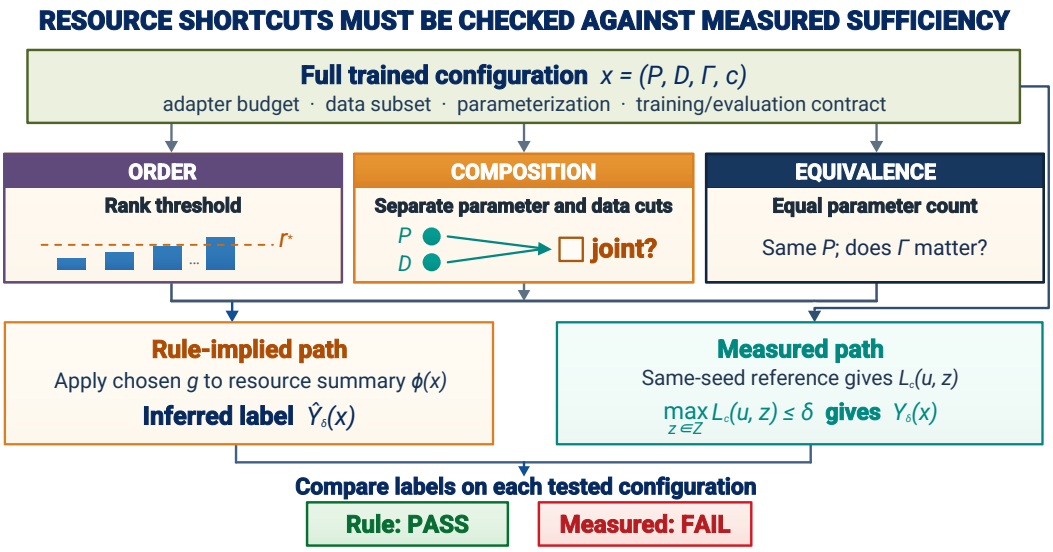


Figure 1: Three resource-selection rules and their audit. A rank threshold, separate parameter and data cuts, and equal parameter counts are read from the same trained configuration and checked against paired measurements. The arrows show the configuration entering both paths and the two labels meeting at the comparison. The shown PASS/FAIL mismatch is illustrative, not a label assigned to a particular experimental cell.

Our contributions are:

1. **Decision-validity formulation and audit.** We define a finite test of whether a declared resource summary and rule recover observed sufficiency labels, with criteria for thresholds, fixed-count groups, and upper sets.
2. **Distinct rule failures with measured selection consequences.** We quantify missed passing configurations and false-sufficient decisions, including sensitivity to learning-rate selection and quality tolerance.
3. **Directly validated low-resource operating points.** Qwen GSM8K grids at two scales identify configurations with worst-seed paired loss below 1 pp relative to same-seed rank-eight/full-pool references, using eightfold fewer adapter parameters and 3.85-fold or 1.93-fold fewer distinct examples visited; Llama and MBPP bound the scope.

This is an empirical measurement study of resource claims; its conclusions are finite, contract-conditioned, and limited to the tested configurations and seeds.

2 When Is a Resource Summary Decision-Valid?

2.1 Defining Measured Sufficiency

We first define the decision a resource rule is meant to support: whether a trained configuration meets a specified quality tolerance. Write $u = (P, D, \Gamma)$, where P is the adapter parameter count, D identifies a frozen training subset and its available pool size, and Γ specifies parameterization and initialization. A data fraction refers to the available pool, not necessarily the number of distinct examples visited under a fixed exposure budget; we report actual visits separately when they support a resource claim. The training/evaluation contract c records the pretrained checkpoint, task, preprocessing, adapter targets and scaling, optimizer, learning-rate rule, schedule, training horizon,

checkpoint selection, data splits, and seed set. The full tested configuration is $x = (u, c)$; omitting c is itself a choice about what the summary retains.

For seed z, let $S(u, c, z) \in [0, 1]$ denote the task evaluation score: Accuracy for classification, exact-match accuracy for GSM8K, or Pass@1 for MBPP. Let $S_{\mathrm{ref}}(z)$ be the score of the fixed same-seed reference configuration and recipe for that audit. Each audit fixes its reference; in the local-learning-rate audits, it remains the default-recipe rank-eight/full-pool checkpoint. The paired loss is one-sided and expressed in percentage points:

$$L_c(u, z) = 100\,[S_{\mathrm{ref}}(z) - S(u, c, z)]. \tag{1}$$

A seed passes when this loss is at most δ; a configuration passes only when every observed seed passes. For the finite tested set X_{test} and seed set Z, define

$$\begin{aligned} y_\delta(x, z) &= \mathbf{1}[L_c(u, z) \le \delta], \qquad \mathbf{y}_\delta(x; Z) = (y_\delta(x, z))_{z \in Z}, \\ Y_\delta(x; Z) &= \prod_{z \in Z} y_\delta(x, z), \\ \mathcal{A}_{\mathrm{test}}(\delta, Z) &= \{x \in X_{\mathrm{test}} : Y_\delta(x; Z) = 1\}. \end{aligned} \tag{2}$$

Here $\mathbf{y}_\delta$ records seed-wise outcomes, whereas Y_δ is the all-seed decision. We use $Z = \{42, 73, 101\}$, with 3 pp as the primary tolerance and 1 pp for sensitivity analysis. Negative losses indicate improvement over the reference; relative losses should be interpreted alongside its absolute accuracy.

2.2 Checking resource summaries and their rules

A summary $\phi(x)$ specifies what information is retained; a rule g specifies how that information is used. For example, $\phi(x) = r$ retains rank, while $g(r) = \mathbf{1}[r \ge r^*]$ applies a threshold. For a particular rule, the audit compares its implied labels $g(\phi(x))$ with measured labels $Y_\delta(x; Z)$.

A separate question is whether *any* rule in an allowed class $\mathcal{G}$ can reproduce all measured labels. For $\phi : X_{\mathrm{test}} \to B$, define

$$\begin{aligned} &\mathrm{Valid}_{\mathcal{G}}(\phi; X_{\mathrm{test}}, \delta, Z) \\ &= \mathbf{1}\big[\exists g \in \mathcal{G} :\ Y_\delta(x; Z) = g(\phi(x))\ \forall x \in X_{\mathrm{test}}\big]. \end{aligned} \tag{3}$$

Equation (3) tests whether the rule class contains a rule that recovers every measured label; the decision replays evaluate particular rules. A particular rule can fail even when another rule in the class succeeds.

Proposition 1 (Finite decision-validity criteria). Fix a finite tested set and its measured labels. (i) If arbitrary rules are allowed, a summary is decision-valid exactly when configurations sharing the same summary value also share the same label. (ii) On a finite ordered rank ladder, a threshold rule exists exactly when the passing ranks form an upper set: once a rank passes, every higher tested rank also passes. The empty passing set is represented by $r^* = +\infty$. (iii) A parameter-count summary preserves the seed-wise signature exactly when that signature is identical across configurations with the same P. The same criterion applies to the all-seed label when that is the prediction target.

A group sharing a summary value is often called a *fiber*. On ARC at 1 pp, ranks 1, 2, 4, and 8 yield labels $1, 0, 1, 1$, ruling out a rank threshold but not an arbitrary rank lookup. In T5/QNLI at 3 pp, equal-count configurations have signatures $010, 101, 111$ and all-seed labels $0, 0, 1$: parameter count alone cannot reproduce either target. These checks are distinct from testing whether separately passing parameter and data reductions also pass jointly (Table 1).

2.3 Measuring departures from resource monotonicity

For a complete measured rank–data grid under a fixed contract and subset manifest, write $X_{\mathrm{grid}} = \mathcal{R} \times \mathcal{Q}$, where $q \in \mathcal{Q}$ is a tested data fraction. Order cells by $(r, q) \preceq (r', q')$ when $r \le r'$ and $q \le q'$. Let S_δ be the cells passing the all-seed rule. It is an *upper set* if every tested cell above a passing cell also passes. This is a property of the pass/fail map, not monotonicity of raw accuracy.

We measure departures from this structure by the minimum fraction of labels that must change to obtain an upper set:

$$D_{\uparrow}(\delta; X_{\text{grid}}) = \min_{U \in \mathcal{U}(X_{\text{grid}})} \frac{|S_\delta \triangle U|}{|X_{\text{grid}}|}, \tag{4}$$

where $\mathcal{U}(X_{\text{grid}})$ contains all upper sets and the symmetric difference counts disagreeing labels. For example, $D_{\uparrow} = 3/15$ means that at least three labels on fifteen tested cells must change to impose monotonicity. The comparison upper set does not replace the measured labels.

Proposition 2 (Finite frontier criterion). $D_{\uparrow} = 0$ exactly when S_δ is an upper set. Its minimal passing cells then form an antichain, and their upward closure recovers S_δ: $S_\delta = {\uparrow}\,\mathrm{Min}(S_\delta)$. If $D_{\uparrow} > 0$, every upper-set representation differs from the observations on at least $D_{\uparrow}|X_{\text{grid}}|$ cells. Thus an upward-closed frontier cannot exactly reconstruct the measured map.

Monotonicity does not imply that marginal reductions compose. In the Llama-3.1-8B GSM8K grid, the safe map is an upper set at both tolerances, yet at 3 pp rank-one/full-pool and rank-eight/half-pool pass while rank-one/half-pool fails. The joint point lies below both marginal points in the coordinate order, so upper-set structure alone does not require it to pass.

These audits describe tested configurations under their stated contracts, references, tolerances, and observed seeds; they do not assign labels to untested cells. $D_{\uparrow}$ is computed only on complete tested grids and concerns all-seed labels; it is not a claim about raw-accuracy monotonicity, causal mechanisms, future-seed guarantees, or joint-reduction feasibility. The retrospective shortcut rules are diagnostic choices, while ARC uses validation-based learning-rate selection followed by sealed-test evaluation.

Table 1: Resource inferences, their required conditions, and finite evidence. Reference quality is a prerequisite for interpreting relative sufficiency, not an order property.

Inference	**Required condition**	**Finite evidence**
Rank threshold / upper-set closure	Passing labels form an upper set	ARC: rank hole at 1 pp; Qwen GSM8K: $D_{\uparrow} > 0$ at 1 pp; Llama: $D_{\uparrow} = 0$ at both tolerances
Marginal P/D $\rightarrow$ joint	$(Y_P = 1) \wedge (Y_D = 1) \implies Y_{\text{joint}} = 1$	QNLI and Llama GSM8K: passing marginals, failing directly trained intersections
Equal-count equivalence	Same P implies the same all-seed label	T5 at 3 pp: signatures $010, 101, 111$; all-seed labels $0, 0, 1$
Ignore the contract	Decisions remain unchanged across compared contracts	Learning-rate changes alter observed passing sets
Interpret relative loss	The reference has adequate absolute quality	MBPP: rank-eight/full-pool reference is only 2.07% Pass@1

3 Experimental Setup

We evaluate the three resource rules introduced in Section 2: rank thresholds, composition of parameter and data reductions, and equal-count interchangeability. Each configuration is compared with its setting's same-seed rank-eight/full-pool reference. It passes when its paired loss is at most the tolerance for all three seeds, $Z = \{42, 73, 101\}$. We use 3 percentage points (pp) as the primary tolerance and 1 pp for sensitivity analysis. Appendix A records the recoverable recipes and provenance limits; Appendix B maps each study to its question.

Historical census and controls. The rank census covers BERT/QNLI, BERT/RTE, Pythia/QNLI, Pythia/MRPC, T5/QNLI, and T5/MNLI at ranks $\{1, 2, 4, 6, 8, 16\}$, giving 108 seed-runs. BERT and Pythia use PEFT LoRA; T5 uses its native $n = 1$ LoRA-equivalent wrapper. No full-fine-tuning run enters this census. Accuracy is the common census metric; the Pythia/MRPC F1 scores reported elsewhere are not used for its pass/fail labels. We compare configurations within each model–task setting rather than pooling accuracies across different recipes.

Nine additional BERT runs test reduced-rank/half-pool intersections, and nine T5/QNLI runs compare three parameterizations at the same adapter count. The BERT full/half pools contain 4,096/2,048 examples for QNLI and 2,240/1,120 for RTE; their training budgets remain 500 and

700 steps, respectively. The historical census and controls use previously inspected development data. The local learning-rate audits also reuse development data for selection and reporting, so they measure sensitivity to the training recipe rather than independent tuning performance.

Prospective ARC evaluation. On DeepSeek-R1-Distill-Qwen-1.5B, we train ranks $\{1, 2, 4, 8\}$ with learning rates $\{5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}\}$ and three seeds: 36 runs of 420 steps. One learning rate per rank is selected by mean validation accuracy across seeds. Only the 12 selected checkpoints are then evaluated once on the sealed 1,172-example test set, using next-token choice-label scoring.

Prospective rank–data grids. For DeepSeek-R1-Distill-Qwen-1.5B (Qwen 1.5B below), we cross ranks $\{1, 2, 4, 8, 16\}$ with nested data fractions $\{25, 50, 100\%\}$: 15 configurations and 45 seed-runs per task. GSM8K uses 900 steps and exact-match scoring on 1,319 test examples; MBPP uses 200 steps and sandboxed Pass@1 on 500 tests. DeepSeek-R1-Distill-Qwen-7B (Qwen 7B) and Llama-3.1-8B each use a GSM8K grid of four ranks $\{1, 2, 4, 8\}$, two data fractions $\{50, 100\%\}$, and three seeds, giving 24 runs per model under the 900-step recipe. The Llama slice was evaluated on the sealed 1,319-example GSM8K test set after its training receipts and evaluation protocol were sealed; checkpoints and evaluator details are recorded in the preregistration and model manifest, and the model family is described by Grattafiori et al. (2024).

The prospective designs were frozen before their reported held-out evaluations; GSM8K had nevertheless been used earlier in the broader project. Here, data fractions denote available subset-pool sizes. The Qwen and Llama GSM8K samplers use 900 steps with effective batch size 8 (7,200 example exposures); their code-derived distinct-example visits are reported separately in Section 7 and Appendix O. Reduced-data runs reuse examples under the same step and exposure budgets; we make no compute-saving claim.

4 Rank Thresholds Can Fail at Strict Tolerances

4.1 A threshold fails on the sealed ARC test

After validation-based learning-rate selection, the ARC all-seed labels at 1 pp are **PASS–FAIL–PASS–PASS** for ranks 1, 2, 4, and 8. Their worst-seed paired losses are 0.853, 1.621, 0.853, and 0 pp, respectively. No single rank cutoff reproduces these labels. At 3 pp, all four ranks pass, so the violation depends on the quality target (Figure 2(a)). All ranks selected the largest candidate learning rate, 2×10^{-4}; this result describes the stated local selection rule, not an optimum beyond the candidate range. Appendix C reports all 12 selected test endpoints.

4.2 The census reveals missed lower-rank configurations

The default-contract census contains both ordered and non-ordered passing sets. At 3 pp, every tested rank passes for BERT/QNLI, BERT/RTE, and Pythia/QNLI. The passing sets are $\{1, 2, 8, 16\}$ for Pythia/MRPC, $\{1, 8\}$ for T5/QNLI, and $\{2, 8\}$ for T5/MNLI. Thus five settings admit rank one and the sixth admits rank two, but some passing ranks lie below failures. The largest worst-seed rank-order reversals are 4.4 pp for Pythia/MRPC and 35.6/42.6 pp for the two T5 settings. Appendix D and Figure A1 retain the complete endpoints and curves.

To measure the consequence for selection, we replay a rule that starts from the known passing rank-eight reference, tests ranks $6, 4, 2, 1$ in descending order, and stops at the first failure. It returns the lowest passing rank reached. This rule misses the smallest tested passing rank in **3/6 settings at 3 pp and 2/6 at 1 pp**. At 3 pp, it returns rank eight for Pythia/MRPC and both T5 settings despite a smaller passing rank. At 1 pp, only the two T5 minima are missed. These are offline diagnostics of the specified rule on the measured endpoints. Appendix P compares the alternative rank-16-start rule, whose miss counts differ.

4.3 Learning-rate selection changes which ranks pass

Local learning-rate selection removes some rank-selection errors but leaves others on ranks measured under both policies. Each retrospective audit selects one learning rate per rank by mean development accuracy and uses it for all three seeds. Paired losses remain relative to the frozen default-

Table 2: Passing ranks after the stated learning-rate selection. Each rank uses one rate shared across seeds. ARC selects on validation and reports sealed-test results; the two historical audits select and report on inspected development data.

Setting	Selection / reporting	3-pp set	1-pp set
ARC-Challenge	Validation / sealed test	$\{1, 2, 4, 8\}$	$\{1, 4, 8\}$
T5/QNLI	Inspected dev / same dev	$\{1, 2, 4, 8\}$	$\{1, 4, 8\}$
Pythia/MRPC	Inspected dev / same dev	$\{1, 2, 6, 8, 16\}$	$\{8\}$

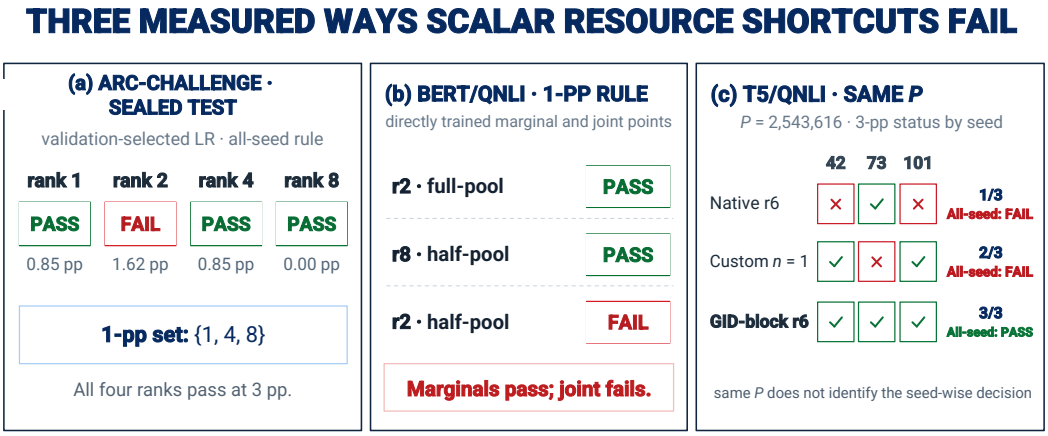


Figure 2: Three measured limits of resource rules. (a) ARC ranks 1, 2, 4, and 8 give PASS–FAIL–PASS–PASS at 1 pp after validation-based learning-rate selection; all four pass at 3 pp. Values are worst-seed paired losses. (b) BERT/QNLI rank two/full-pool and rank eight/half-pool pass at 1 pp, but their directly trained intersection fails. (c) T5/QNLI parameterizations share 2,543,616 adapter parameters but have different 3-pp seed-wise labels and all-seed decisions. Columns show seeds 42, 73, and 101; the final column gives passing-seed counts and all-seed decisions.

contract rank-eight reference. For T5/QNLI, the 3-pp passing set expands from $\{1, 8\}$ to $\{1, 2, 4, 8\}$; the selected 1-pp set is $\{1, 4, 8\}$. For Pythia/MRPC, the 3-pp set expands from $\{1, 2, 8, 16\}$ to $\{1, 2, 6, 8, 16\}$, still excluding rank four; only rank eight passes at 1 pp. Table 2 summarizes these results. Selection and reporting reuse the same inspected development data, so these comparisons establish recipe sensitivity, not independent tuning performance. An optimistic per-rank, per-seed selection on the same development set makes all four tested T5/QNLI ranks pass at both tolerances; this is a post-selection envelope rather than an independent evaluation. Appendix F retains that envelope together with the pairwise reversals.

On ranks measured under both LR policies, local selection removes the rank-eight-anchored replay's T5/QNLI miss at 3 pp, but not at 1 pp; for Pythia/MRPC the 3-pp replay still returns rank six instead of the common-grid minimum rank one. These retrospective, common-rank replays are detailed in Appendix Q. Passing labels near the tolerance boundary or against a weak reference require particular care: Pythia/MRPC rank one passes at 3 pp by only 0.059 pp, and the T5/MNLI reference includes one seed with 427/1,024 correct. A passing rank is therefore conditional on its reference quality and tolerance.

5 Parameter and Data Savings Do Not Always Compose

A parameter reduction and a data reduction can each meet a quality target while their combination fails it. In BERT/QNLI, rank two/full-pool and rank eight/half-pool pass the 1-pp rule on all three seeds. Their directly trained rank-two/half-pool intersection instead loses 1.953, 1.758, and 1.855 pp and fails on every seed (Figure 2(b)). All three points use the same reference and tolerance, and all pass at 3 pp. Table A5 provides the absolute counts and paired losses.

The same type of failure appears on Llama-3.1-8B GSM8K. At 3 pp, rank one/full-pool and rank eight/half-pool pass separately, but rank one/half-pool has a worst-seed loss of 3.184 pp. At 1 pp, rank four/full-pool and rank eight/half-pool pass separately, but rank four/half-pool loses 2.502 pp. For Qwen, the tested joint reductions inferred from passing marginals pass at both tolerances; composition failure is therefore setting-dependent.

Direct measurements also identify passing BERT joint reductions. QNLI rank one/half-pool and RTE rank one/half-pool use one eighth of the reference adapter parameters and half-size training pools (4,096 to 2,048 for QNLI; 2,240 to 1,120 for RTE). Both pools are fully visited under the fixed exposure budgets. Their worst-seed losses are 1.953 and 1.083 pp: both pass at 3 pp, but neither passes at 1 pp. The useful conclusion is the quality of each measured joint configuration, not a joint saving inferred from two marginal labels. Appendix J distinguishes tested from untested BERT intersections.

6 Equal Parameter Counts Can Yield Different Sufficiency Decisions

At the 3-pp target, Native rank-six, Custom ($n = 1$) rank-six, and GID-block rank-six configurations share 2,543,616 adapter parameters (75% of the rank-eight reference) and the same QNLI split, yet their all-seed decisions are **FAIL, FAIL, and PASS**, respectively. GID-block uses an existing parameterization (Ye et al., 2026); it is a matched-count control, not a method introduced here.

The corresponding per-configuration seed-wise signatures are 010, 101, and 111 (Figure 2(c); Table A6). A rule using parameter count alone must assign one label to all three and therefore disagrees with at least one measured decision; this comparison establishes neither method superiority nor its cause.

In the tested matched-state single-layer cases, forward outputs, gradients, and one-step updates agree within float32-scale numerical precision. Historical runs used different per-layer RNG mappings despite the same Kaiming-A/zero-B initialization, so the contrast does not isolate a cause (Appendix L).

7 Substantial Resource Headroom Persists across Decoder Settings

Qwen 1.5B at rank one/25% pool and Qwen 7B at rank one/50% pool meet the 1-pp target with one eighth of the reference adapter parameters. Their worst-seed paired losses are 0.682 and 0.455 pp, against references averaging 58.35% and 76.07% accuracy. Under the fixed training budget, sampler replay gives 3.85-fold and 1.93-fold reductions in distinct examples visited, respectively (Figure 3; Appendix O). Both configurations pass on all three seeds.

Llama-3.1-8B provides a further directly measured joint point. Rank two/half-pool loses at most 1.820 pp across observed seeds relative to its rank-eight/full-pool reference, whose mean accuracy is 54.865% (versus 3.336% for the unadapted checkpoint). This point passes at 3 pp with one quarter of the adapter parameters and half the distinct examples visited (3,363 versus 6,726). All GSM8K grids use 900 steps and 7,200 sampled example positions; detailed pool and visit counts are documented in Appendix O. The differing grids support setting-specific operating points, not a matched scaling comparison.

7.1 Resource order and shortcut errors depend on the setting

At 1 pp, Qwen 1.5B and 7B have non-monotone pass/fail maps: their upper-set defects are $3/15 = 0.20$ and $1/8 = 0.125$. Llama-3.1-8B is already an upper set, with $D_{\uparrow} = 0$. At 3 pp, every tested Qwen cell passes, while the Llama map remains an upper set with one failing cell. Thus the measured order structure changes with both the setting and the tolerance (Figure 4).

An offline upward-closure replay shows the consequence of imposing monotonicity on these maps. Taking every cell above a measured passing cell to be sufficient produces **four false-sufficient labels among 31 tested GSM8K cells at 1 pp**: three on Qwen 1.5B and one on Qwen 7B. At 3 pp there are

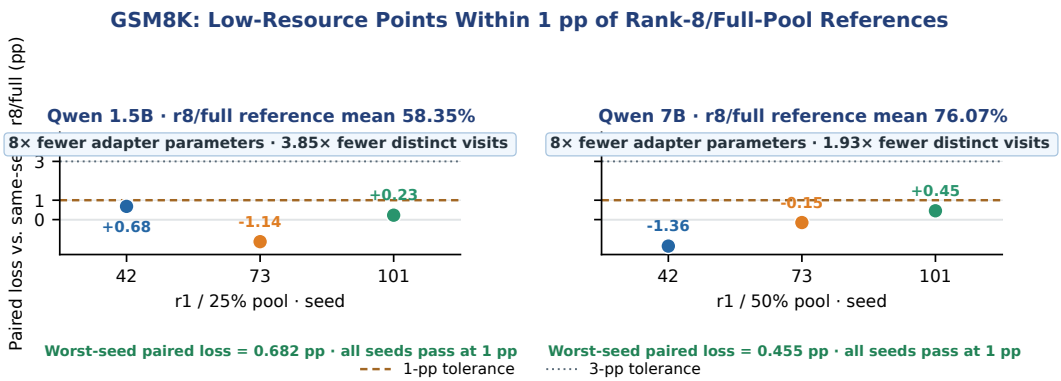


Figure 3: Joint adapter and training-pool reductions meeting the 1-pp target on GSM8K. D denotes available training-pool fraction: 25% for Qwen 1.5B and 50% for Qwen 7B. Dots show seed-wise loss relative to each rank-eight/full-pool reference; worst-seed losses are 0.682 and 0.455 pp. Both use one eighth of the adapter parameters. Under the common 900×8 exposure budget, sampler replay gives 1,868 versus 7,200 and 3,736 versus 7,200 distinct examples visited (3.85-fold and 1.93-fold reductions). Negative losses indicate improvement. Lines mark the 1- and 3-pp tolerances.

none, and neither replay produces false-insufficient labels. These counts describe disagreements on the measured grids. Appendix M.5 reports the exact upper-set comparisons. The Llama composition failures in Section 5 also show why zero order defect does not guarantee that marginal reductions compose.

Tolerance effects are not monotone: Llama's defect is zero at 1 and 3 pp but $1/8$ at 2 pp. At 1 pp, Qwen 7B's only order defect is rank-four/full-pool (1.440 pp); rank-one/full-pool passes (0.986 pp). Appendix N reports the full sweep and label transitions.

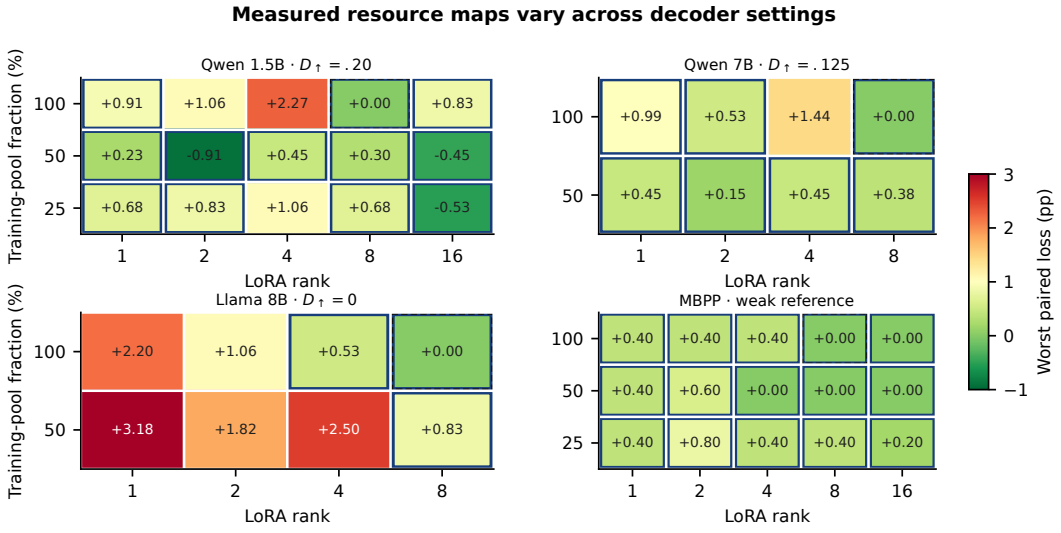


Figure 4: Measured rank–data maps depend on the setting. Each cell shows worst-seed paired loss (pp) relative to its own setting's rank-eight/full-pool reference. Navy outlines mark 1-pp passes; a black dashed frame marks the reference. The Qwen 1.5B, Qwen 7B, and Llama-3.1-8B GSM8K maps have upper-set defects of 0.20, 0.125, and 0 at 1 pp. MBPP is a weak-reference case: its reference achieves 2.07% Pass@1. Only trained configurations are shown.

Reference quality remains essential. MBPP is a weak-reference boundary case: its rank-eight/full-pool reference reaches only 2.07% Pass@1, and all cells lie within 1 pp of it. Relative closeness alone does not imply useful task quality.

8 Related Work

Rank selection and parameter allocation. LoRA freezes pretrained weights and learns low-rank updates (Hu et al., 2022). AdaLoRA allocates rank by weight-matrix importance (Zhang et al., 2023); DyLoRA trains once for multiple usable ranks (Valipour et al., 2023); and LR-LoRA learns layer-specific ranks during optimization (Garg et al., 2026). Park studies theoretical rank requirements under neural tangent kernel assumptions, including conditional guarantees for binary classification (Park, 2026). These works address adapter construction, rank selection, or optimization conditions. We instead test whether observed pass/fail labels at a declared tolerance admit a single rank cutoff. Stop-after-failure is a diagnostic replay rule here, not a procedure attributed to these methods.

Parameterization, optimization, and empirical behavior. Equal parameter counts can correspond to different update structures. RandLoRA combines fixed random low-rank matrices with learned scaling to produce full-rank updates (Albert et al., 2025); LoRA-Pro changes adapter gradients toward full-fine-tuning gradients (Wang et al., 2025); RaLoRA aligns capacity to layerwise gradient intrinsic dimensionality, while RaLoRA-Pro additionally reallocates parameters using loss sensitivity (Ye et al., 2026). Biderman et al. (2024) report a learning–forgetting trade-off and sensitivity to target modules, rank-dependent scaling, and learning rate in programming and mathematics. Our matched-count and learning-rate controls test whether such choices change which configurations meet a same-seed tolerance; they propose no parameterization and identify no cause.

Joint resource scaling and data efficiency. Zhang et al. (2024) jointly model fine-tuning data with model size, pretraining data, and trainable-parameter count across full fine-tuning, prompt tuning, and LoRA; their held-out translation and summarization fits report weak or sometimes inverse gains from larger PEFT parameter counts. MARS predicts module-specific convergence and task performance to select rank pairs in multimodal fine-tuning (Cho et al., 2026). We complement these response models by auditing finite pass/fail decisions at a fixed tolerance: fitted trends and separate parameter/data passes need not predict a directly trained intersection. Our grids vary available pool size at fixed steps and exposures, with sampler-derived distinct visits reported separately.

9 Discussion and limitations

Implications. The audit separates rank-threshold, marginal-composition, and equal-count decisions. Local learning-rate selection repairs some replay errors, but tolerance sweeps show that a larger passing set need not have a smaller order defect. Reports should retain tested configurations, exceptions, references, and selection rules. This is a diagnostic, not a cheaper search procedure.

Limitations. Labels use three observed seeds. Historical census and learning-rate audits reuse inspected development data; ARC uses a sealed test after selection, but every rank chose the largest candidate rate. Decoder evidence covers GSM8K at two Qwen scales and one Llama slice; GSM8K was used earlier, recipes were not jointly tuned, and MBPP has a weak reference. The matched-tensor check covers tested single-layer cases only and identifies no cause. Sampler replay distinguishes pool size, exposures, and distinct visits, but establishes no compute savings. We make no claims for untested configurations or seeds.

10 Conclusion

Rank thresholds, marginal resource cuts, and equal-count configurations can disagree with measured sufficiency; direct GSM8K evaluation also finds lower-resource configurations meeting the target. PEFT resource claims should report the tested configuration, reference, tolerance, training contract, and exceptions rather than infer sufficiency from a scalar budget alone.

AI USE STATEMENT

Generative AI tools assisted with refining the paper's conceptual framework and hypotheses; discussing experimental methodology and existing results; searching and summarizing literature; authoring and editing analysis code; and drafting, restructuring, and editing the manuscript.

REFERENCES

Paul Albert, Frederic Z. Zhang, Hemanth Saratchandran, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. RandLoRA: Full-rank parameter-efficient fine-tuning of large models, 2025. URL https://arxiv.org/abs/2502.00987.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=aloEru2qCG. Featured Certification.

Minkyoung Cho, Insu Jang, Shuowei Jin, Zesen Zhao, Adityan Jothi, Ethem F. Can, Min-Hung Chen, and Z. Morley Mao. MARS: Harmonizing multimodal convergence via adaptive rank search, 2026. URL https://arxiv.org/abs/2603.00720.

Arpit Garg, Simon Lucey, and Hemanth Saratchandran. Parameter-efficient fine-tuning with learnable rank, 2026. URL https://arxiv.org/abs/2606.04325.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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://arxiv.org/abs/2106.09685.

Juneyoung Park. Rethinking the rank threshold for LoRA fine-tuning, 2026. URL https://arxiv.org/abs/2605.03724.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287, 2023. URL https://aclanthology.org/2023.eacl-main.239/.

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

Jingqi Ye, Haonan He, Minglei Li, Fujun Han, Tao Chen, and Peng Ye. Gradient intrinsic dimensionality alignment: Narrowing the gap between low-rank adaptation and full fine-tuning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5b4b967d4222d87fa5b28b6ec7144058-Abstract-Conference.html.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets LLM finetuning: The effect of data, model and finetuning method, 2024. URL https://arxiv.org/abs/2402.17193.

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

A FULL EXPERIMENTAL CONTRACTS

The comparisons are paired within each model–task training contract, rather than controlled across architectures. Table A1 records the recoverable checkpoint identifiers, adapter targets and scaling, optimizer and schedule, training horizon, checkpoint rule, and evaluation protocol. Frozen subset manifests identify the available training pools. Model scores are taken from retained training and evaluation records; deterministic label, operator, sampler, and search checks are described separately below.

Some historical classification records lack package versions or immutable upstream weight revisions. These fields remain unavailable rather than being reconstructed from a current environment. The Qwen preregistration records PyTorch 2.4.0+cu121, Transformers 4.46.1, PEFT 0.13.2, and Datasets 3.1.0. The converted Llama checkpoint is identified by per-file sizes and SHA-256 hashes because its upstream revision is not recoverable. These limitations constrain exact external reproduction and should accompany the released records.

Table A1: Recoverable training and evaluation configurations for the main evidence. Exact subset identities and run-level settings remain in the frozen manifests and receipts. A field is omitted only where the retained source records do not support a more exact value.

Setting	Recoverable configuration
BERT/QNLI census	**Checkpoint/data.** Local `bert-medium-uncased`; frozen QNLI pool of 4,096 examples and 1,024-example dev slice. **Adapter.** Query/value projections in all eight encoder layers; $\alpha = 2r$; dropout 0; no bias; classifier saved/trainable; Kaiming-A/zero-B initialization. **Training/evaluation.** AdamW, LR 2×10^{-4}, weight decay 0.01, batch 32, 500 fixed steps, max length 128, no scheduler in the retained Stage-1 trainer; final-step Accuracy on the fixed dev slice.
BERT/RTE census	**Checkpoint/data.** Same local BERT-medium checkpoint; RTE training pool 2,240, search split 249, official dev 277; deterministic split seed 29. **Adapter.** Query/value projections in all eight encoder layers; $\alpha = 2r$; dropout 0; no bias; classifier saved/trainable; Kaiming-A/zero-B initialization. **Training/evaluation.** AdamW, LR 2×10^{-4}, weight decay 0.01, linear schedule with 6% warmup, batch 32, 10 epochs (700 steps), max length 128; final-checkpoint Accuracy on official dev.
Pythia/ QNLI and MRPC census	**Checkpoint/data.** Local `EleutherAI/pythia-410m`; retained task-specific train/search splits; official dev sizes 5,463 (QNLI) and 408 (MRPC). **Adapter.** Fused `query_key_value` projection in each of 24 GPT-NeoX layers; $\alpha = 2r$; dropout 0; no bias; `score` head trainable/saved; Kaiming-A/zero-B initialization. **Training/evaluation.** AdamW, weight decay 0.01, LR 10^{-4} (QNLI) and 5×10^{-4} (MRPC), linear schedule with 6% warmup, batch 16, 3 epochs, max length 128; final-epoch Accuracy on official dev.
T5/ QNLI and MNLI census	**Checkpoint/data.** Clean local `google/t5-v1_1-base` weights; each task uses its frozen 4,096-train/1,024-dev manifest. **Adapter.** Native $n = 1$ LoRA-equivalent wrapper on the frozen encoder/decoder target map; $\alpha = r$, native fan-in A, zero initial effective update, no task head. **Training/evaluation.** Adam, LR 0.002, $\epsilon = 10^{-4}$, weight decay 0, no scheduler, batch 32, 500 steps, input length 96 and target length 8; final-step Accuracy on the fixed dev slice. Full checkpoint provenance and target-count verification follow the table.

Setting	Recoverable configuration (continued)
T5/QNLI same-count controls	**Checkpoint/data.** Same clean local T5-v1.1-base checkpoint; frozen 4,096/1,024 QNLI split and same-seed rank-eight/full references. **Adapters.** Native uniform rank six, custom true-block operator with $n = 1$ on all 216 listed encoder/decoder linear targets, or a training-gradient-GID block map; all use $\alpha = 6$ and exactly 2,543,616 adapter parameters. The block map is frozen from training-only calibration with power-of-two block counts capped at 32 and constrained by matrix dimensions. **Training/evaluation.** Adam, LR 0.002, $\epsilon = 10^{-4}$, no schedule or weight decay, batch 32, 500 steps (16,000 example exposures), input length 96 and target length 8; final-step Accuracy on the fixed dev slice. Block allocation does not use dev scores.
ARC-Challenge Qwen 1.5B	**Checkpoint/data.** `deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B`, revision `ad9f0ae0864d7fbcd1cd905e3c6c5b069cc8b562`; ARC-Challenge train 1,119, validation 299, sealed test 1,172. **Adapter.** PEFT LoRA on `q_proj`/`v_proj`; ranks 1/2/4/8; $\alpha = 16$ at every rank; dropout 0, no bias, Kaiming-A/zero-B; frozen base. **Training/evaluation.** AdamW with candidates $5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}$; cosine schedule, 3% warmup, 420 steps, microbatch 2 with accumulation 4, max length 512, BF16, final step. One LR/rank selected by mean validation Accuracy across seeds; selected checkpoints evaluated by next-token choice Accuracy on sealed test.
GSM8K and MBPP Qwen 1.5B	**Checkpoint/data.** Same pinned Qwen 1.5B checkpoint as above; GSM8K and MBPP dataset revisions are pinned in the preregistration; held-out test sets sealed through training. **Adapter.** PEFT LoRA on `q_proj`/`v_proj`; ranks 1/2/4/8/16; $\alpha = 16$ with standard α/r scaling; dropout 0, no bias, Kaiming-A/zero-B; frozen base. **Training/evaluation.** AdamW, LR 10^{-4}, $\epsilon = 10^{-8}$, weight decay 0, cosine schedule with 3% warmup, effective batch 8, max length 512, BF16, final step. GSM8K uses 900 steps and exact-match final-answer scoring; MBPP uses 200 steps and sandboxed Pass@1.
GSM8K Qwen 7B	**Checkpoint/data.** Local DeepSeek-R1-Distill-Qwen-7B checkpoint; exact checkpoint files identified by retained model-manifest hashes; GSM8K test has 1,319 examples. **Adapter.** PEFT LoRA on `q_proj`/`v_proj`; ranks 1/2/4/8, data 50/100%; $\alpha = 16$ with standard α/r scaling; dropout 0, no bias, Kaiming-A/zero-B; frozen base. **Training/evaluation.** Same frozen GSM8K recipe as Qwen 1.5B: AdamW, LR 10^{-4}, cosine/3% warmup, effective batch 8, 900 steps, max length 512, BF16, final step; greedy exact-match answer evaluation.
GSM8K Llama-3.1-8B	**Checkpoint/data.** Local `meta-llama/Llama-3.1-8B` snapshot identified by per-file size/SHA-256 manifest; upstream revision is not recoverable from the converted snapshot. GSM8K revision pinned; 747 examples reserved for validation; official test 1,319. **Adapter.** PEFT LoRA on all `q_proj`/`v_proj` layers; ranks 1/2/4/8; $\alpha = 16$ with standard α/r scaling; dropout 0, no bias, Kaiming-A/zero-B; frozen base. **Training/evaluation.** Same frozen GSM8K optimization contract as Qwen 7B: AdamW, LR 10^{-4}, cosine/3% warmup, effective batch 8, 900 steps, max length 512, BF16, final step; fixed plain-text prompt and greedy exact-match scoring.

T5 checkpoint and adapter-target verification. The checkpoint conversion record identifies the public source as `google/t5-v1_1-base` (source weight-file SHA-256 `1d51351caeff6b7509689858a25b74622657f492c01cc62dbf3a61074725eb`) and the locally serialized `model.safetensors` (SHA-256 `03d8601d064997ab32c9e494221ec09399a17de07faf399f69bca4db1c3add24`). A contemporaneous cached-configuration audit records 12 encoder and 12 decoder blocks, $d_{\text{model}} = 768$, gated-GELU FFN with $d_{\text{ff}} = 2048$, and dropout 0.1; the original `config.json` was not retained. The frozen adapter target map is available as `data/t5_adapter_targets.csv`; its source rank-map SHA-256 is `1cdfc05fc9fc009f40629780964aa749e9257f60a2e186ff9e826629650a32e8`. It lists 216 linear targets (144 attention and 72 gated-FFN). Their recorded matrix dimensions sum to 423,936 trainable adapter parameters per rank and 2,543,616 at rank six.

B EVIDENCE MAP

Table A2 links each study to its question. A seed-run is one configuration trained with one seed; a grid cell is one configuration after applying the all-seed rule. Thus the six-setting, six-rank census contains 36 configurations and 108 seed-runs. Counts from different evidence layers are not counts of independent tasks, nor should records reused across analyses be treated as new experiments.

Table A2: Evidence layers and their audit questions. A seed-run is a configuration trained with one seed; grid-cell counts follow the all-seed decision. Counts are not independent-task counts and need not represent disjoint new training runs.

Evidence	**Size**	**Question**
Default rank census	108 seed-runs	Do passing ranks admit a threshold?
Learning-rate checks	2 historical audits; ARC: 36 train, 12 eval	Does local selection change labels, and does a threshold hold on sealed test?
Joint / equal-count controls	18 seed-runs	Do marginal reductions compose, and do equal counts share a decision?
Qwen 1.5B task grids	90 seed-runs	Which low-resource points pass on GSM8K and MBPP?
Scale / family grids	48 seed-runs	What is observed for Qwen 7B and Llama-3.1-8B?
Decision / order replays	Offline; 6 grids, 70 cells	Where do specified rules disagree with measured labels?

C PROSPECTIVE ARC-CHALLENGE TEST ENDPOINTS

Table A3 reports all 12 validation-selected checkpoints on the 1,172-example sealed test set. One learning rate per rank was selected by mean validation accuracy across the three seeds; all four ranks selected 2×10^{-4}, the largest candidate rate. Each endpoint was evaluated after this selection was fixed. The table gives the absolute score, same-seed rank-eight reference score, paired loss, and seed-wise labels. It distinguishes per-seed failures from the all-seed decision used in the main text.

Table A3: All 12 validation-selected ARC-Challenge checkpoints evaluated on the sealed test set. Accuracy and reference are shown as percentages; paired losses are percentage points. P/F columns are the observed per-seed labels at 1 and 3 pp. The same-seed rank-eight reference has zero paired loss.

Rank	Seed	Selected LR	Test (correct/n; Acc.)	Reference Acc.	Loss (pp)	1/3 pp
1	42	2e-04	666/1172 (56.83%)	57.08%	+0.256	P/P
1	73	2e-04	659/1172 (56.23%)	57.08%	+0.853	P/P
1	101	2e-04	658/1172 (56.14%)	55.38%	-0.768	P/P
2	42	2e-04	650/1172 (55.46%)	57.08%	+1.621	F/P
2	73	2e-04	660/1172 (56.31%)	57.08%	+0.768	P/P
2	101	2e-04	638/1172 (54.44%)	55.38%	+0.939	P/P
4	42	2e-04	661/1172 (56.40%)	57.08%	+0.683	P/P
4	73	2e-04	659/1172 (56.23%)	57.08%	+0.853	P/P
4	101	2e-04	650/1172 (55.46%)	55.38%	-0.085	P/P
8	42	2e-04	669/1172 (57.08%)	57.08%	+0.000	P/P
8	73	2e-04	669/1172 (57.08%)	57.08%	+0.000	P/P
8	101	2e-04	649/1172 (55.38%)	55.38%	+0.000	P/P

D FULL RANK ENDPOINT TABLE

Table A4 reports correct-prediction counts for seeds 42, 73, and 101, evaluation denominators, worst-seed paired losses, and the resulting all-seed decisions. These are the 108 seed-runs in the default-contract census, not the locally selected endpoints analyzed in Appendix F. P/F denotes the measured decision at the stated tolerance. The retained `data/complete_lora_grid.csv` records the run-level scores and source-receipt identifiers used for this table.

Table A4: All 108 standard-LoRA endpoints. Counts are correct predictions at seeds 42, 73, and 101. The worst paired loss is relative to same-seed rank-eight/full-pool; P/F are observed passes/fails at 3 pp / 1 pp.

Setting	Rank	42	73	101	n	Worst loss	3 / 1
BERT / QNLI	1	814	814	802	1024	2.930	P / F
BERT / QNLI	2	819	832	823	1024	0.879	P / P
BERT / QNLI	4	822	832	828	1024	0.391	P / P
BERT / QNLI	6	821	832	836	1024	0.391	P / P
BERT / QNLI	8	825	832	832	1024	0.000	P / P
BERT / QNLI	16	829	837	846	1024	-0.391	P / P
BERT / RTE	1	170	175	172	277	1.444	P / F
BERT / RTE	2	174	176	168	277	1.444	P / F
BERT / RTE	4	174	171	179	277	0.000	P / P
BERT / RTE	6	177	170	174	277	0.000	P / P
BERT / RTE	8	174	170	172	277	0.000	P / P
BERT / RTE	16	174	178	174	277	0.000	P / P
Pythia / QNLI	1	4815	4813	4820	5463	1.226	P / F
Pythia / QNLI	2	4855	4848	4834	5463	0.879	P / P
Pythia / QNLI	4	4881	4849	4882	5463	0.513	P / P
Pythia / QNLI	6	4885	4847	4864	5463	0.549	P / P
Pythia / QNLI	8	4882	4877	4882	5463	0.000	P / P
Pythia / QNLI	16	4874	4868	4875	5463	0.165	P / P
Pythia / MRPC	1	331	323	323	408	2.941	P / F
Pythia / MRPC	2	324	334	329	408	1.471	P / F
Pythia / MRPC	4	329	330	318	408	4.167	F / F
Pythia / MRPC	6	326	324	311	408	5.882	F / F
Pythia / MRPC	8	329	330	335	408	0.000	P / P
Pythia / MRPC	16	326	322	335	408	1.961	P / F
T5 / QNLI	1	887	860	857	1024	0.293	P / P
T5 / QNLI	2	518	842	531	1024	35.059	F / F
T5 / QNLI	4	882	529	870	1024	29.395	F / F
T5 / QNLI	6	512	862	824	1024	35.645	F / F
T5 / QNLI	8	877	830	860	1024	0.000	P / P
T5 / QNLI	16	512	781	517	1024	35.645	F / F
T5 / MNLI	1	342	755	752	1024	40.332	F / F
T5 / MNLI	2	785	483	766	1024	-0.098	P / P
T5 / MNLI	4	342	332	330	1024	42.480	F / F
T5 / MNLI	6	792	349	343	1024	41.211	F / F
T5 / MNLI	8	755	427	765	1024	0.000	P / P
T5 / MNLI	16	346	340	334	1024	42.090	F / F

E COMPLETE RANK-RESPONSE CURVES

Figure A1 displays the seed-wise accuracies from Table A4. Each panel uses its own model–task contract and same-seed rank-eight reference. Green ticks mark ranks satisfying the observed three-seed, 3-pp rule. Connecting lines link tested ranks for readability; they neither interpolate unmeasured ranks nor imply an ordered response.

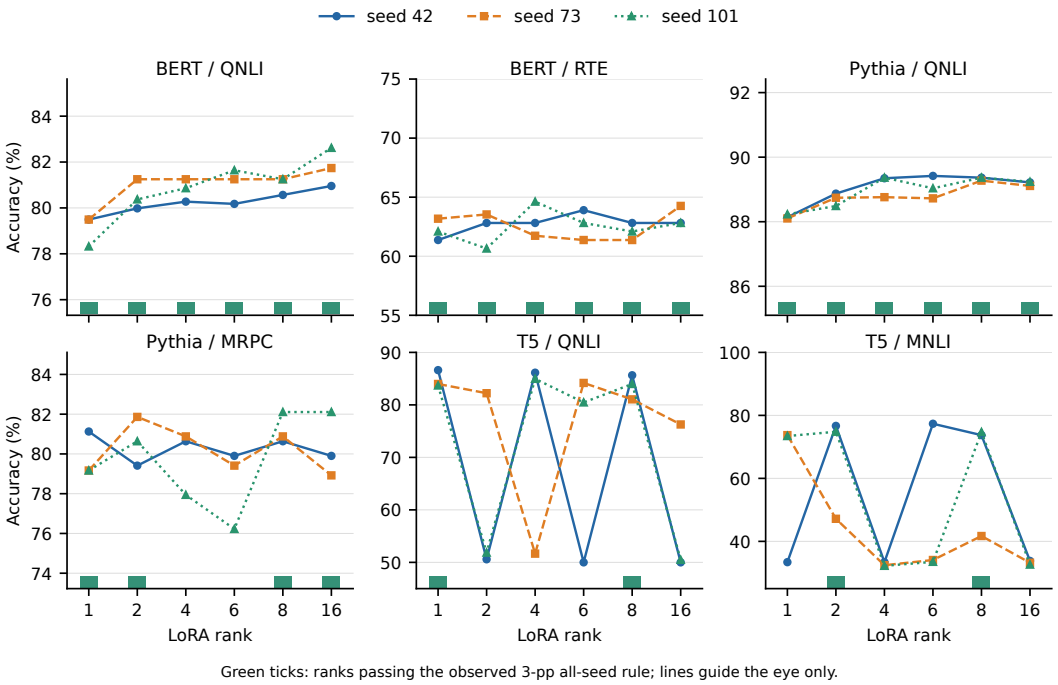


Figure A1: Complete seed-wise rank responses under the six default contracts. Each line connects one seed's measured accuracies; green ticks mark 3-pp all-seed passes. Rank eight is the same-seed reference. Connections between listed ranks are visual guides only.

F LOCAL LEARNING-RATE ROBUSTNESS AUDITS

The two retrospective audits compare the default learning rate with the neighborhood $\{0.5\eta_0, \eta_0, 2\eta_0\}$. For each rank, one rate is selected by mean development accuracy over seeds 42, 73, and 101, and that rate is used across all three seeds. Losses remain relative to the frozen default-contract rank-eight checkpoint for each seed, not a newly selected reference. Selection and reporting use the same previously inspected development data.

For T5/QNLI, shared-across-seed selection is a reanalysis of the existing three-rate endpoints. The originally specified aggregation instead selected a separate local maximum for each rank and seed; that optimistic envelope is retained below. Under the shared rule, the selected multipliers for ranks $\{1, 2, 4, 8\}$ are $\{1, 0.5, 0.5, 0.5\}$. The 3-pp passing set expands from $\{1, 8\}$ to $\{1, 2, 4, 8\}$, while the 1-pp set is $\{1, 4, 8\}$. Rank two has a worst-seed loss of 1.855 pp. Across selected endpoints, 8 of 18 within-seed rank pairs reverse the increasing-rank accuracy ordering; the largest reversal is 3.027 pp. For seed 73, rank two has 872/1,024 correct versus 841/1,024 at rank eight; rank four has 871/1,024 correct, a 2.930-pp advantage over rank eight.

For Pythia/MRPC, selected multipliers for ranks $\{1, 2, 4, 6, 8, 16\}$ are $\{1, 1, 1, 0.5, 1, 0.5\}$. The 3-pp set changes from $\{1, 2, 8, 16\}$ to $\{1, 2, 6, 8, 16\}$; rank four remains excluded. Only rank eight passes at 1 pp. There are 19 reversals among 45 within-seed rank pairs, with a maximum of 2.696 pp. For example, seed 101 scores 329/408 at rank two and 318/408 at rank four. Selection and sufficiency labels use accuracy; MRPC F1 remains a secondary recorded outcome.

The T5 per-rank, per-seed envelope chooses the rate maximizing each endpoint's accuracy on the same development set. All four tested ranks then pass at both tolerances. This removes the 1-pp hole seen under the shared rule, but is an optimistic post-selection comparison rather than an independently evaluated selector. Its seed-73 rank-four and rank-eight scores are 871/1,024 and 841/1,024. Neither analysis establishes the result of a wider search, independent selection, or new training seeds. Appendix Q compares search consequences on ranks shared by the default and local-selection analyses.

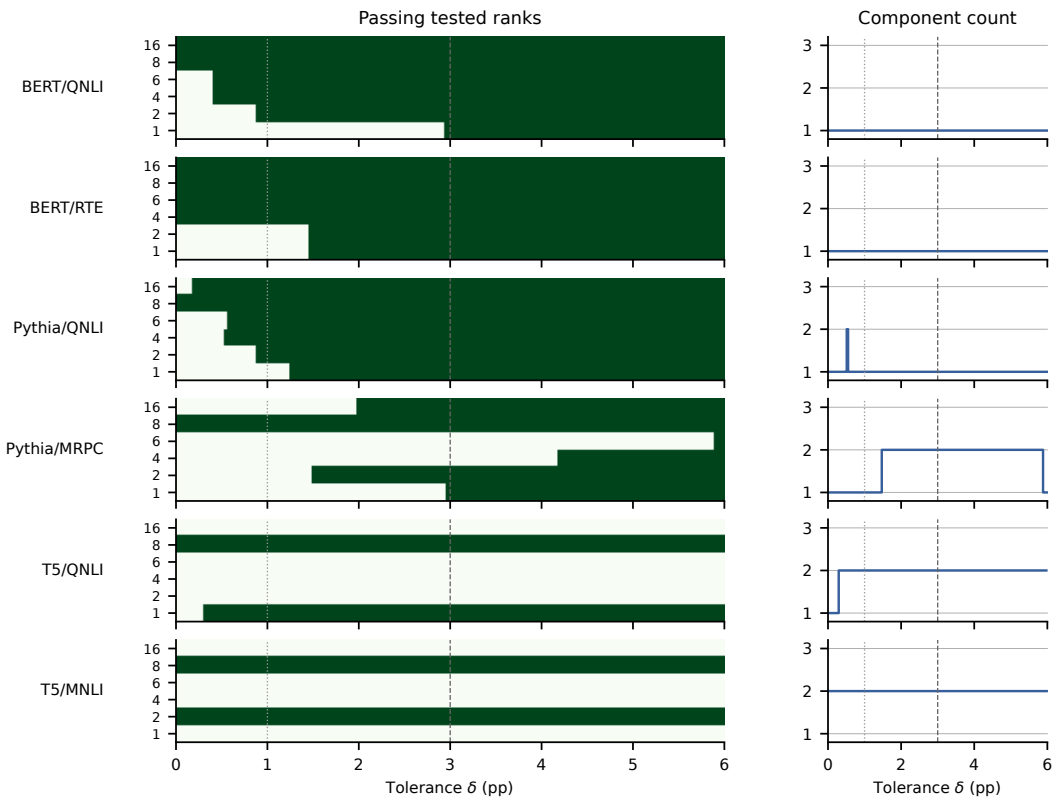


Figure A2: Passing tested ranks and their contiguous-component count as tolerance changes. Bands connect adjacent listed ranks only. This summarizes a finite ladder, not unmeasured ranks or a continuous resource boundary.

G SUFFICIENCY SENSITIVITY DESCRIPTORS

These descriptors use the 108 default-contract endpoints only; they do not incorporate the locally selected results in Appendix F. Let $W(r) = \max_{z \in Z} L(r, z)$. The empirical margin $M_\delta(r) = \delta - W(r)$ records slack relative to the target, not a confidence bound. At 3 pp, the margins are $+0.070$ pp for BERT/QNLI rank one, $+0.059$ pp for Pythia/MRPC rank one, and -2.882 pp for Pythia/MRPC rank six.

For $\delta \geq 0$, define $R_{\text{pass}}(\delta) = \{r : W(r) \leq \delta\}$. On the listed rank ladder, let $K(\delta)$ count contiguous runs of passing ranks. At 3 pp, $K = 2$ for Pythia/MRPC, T5/QNLI, and T5/MNLI. At 1 pp, both T5 settings have two passing components, while Pythia/MRPC has only its rank-eight reference. The second components first appear at 1.471 pp for Pythia/MRPC and 0.293 pp for T5/QNLI. These transitions concern adjacent tested ranks, not a continuous resource space.

For each of the 15 rank pairs $r_i < r_j$, count a reversal at margin ϵ when $W(r_j) > W(r_i) + \epsilon$. At $\epsilon = 0$, the counts for BERT/QNLI, BERT/RTE, Pythia/QNLI, Pythia/MRPC, T5/QNLI, and T5/MNLI are $0, 0, 2, 7, 9, 9$ out of 15. The maximum reversals for the two T5 settings are 35.645 and 42.578 pp. These worst-seed-loss descriptors differ from the within-seed accuracy-pair counts in Appendix F and do not alter the pass/fail rule.

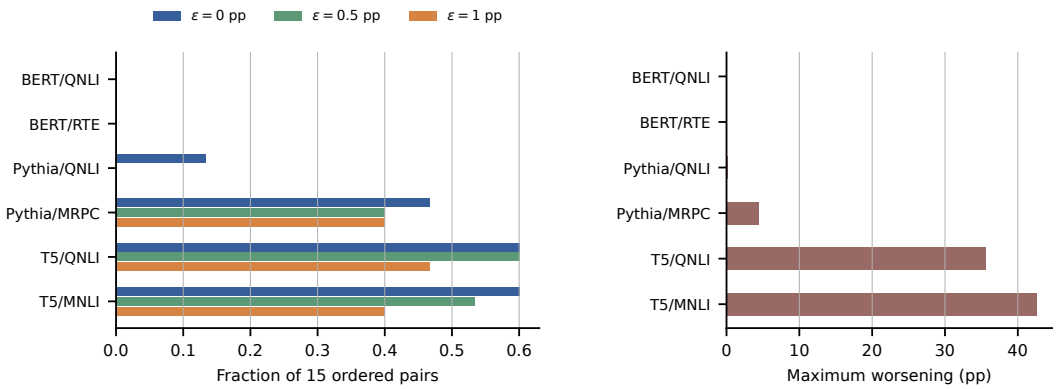


Figure A3: Pairwise reversals of worst-seed paired loss and their maximum amplitudes in the default census. The reversal margin is ϵ. These summaries differ from within-seed accuracy-pair counts and do not establish an expressivity limit.

H ORDEREDNESS AND SEED-INSTABILITY AUDIT

For each model–task setting and seed, an equally weighted isotonic fit to the six rank accuracies gives a root-mean-square residual. E_{mono} is the mean of this residual across seeds. I_{seed} is the mean, across ranks, of the sample standard deviation of accuracy over seeds. Across the six contracts, their Spearman correlation is 0.943; it is 0.800 for the four non-T5 contracts.

Both summaries are computed from the same endpoint matrix, and the large T5 failures can increase both. The small number of settings and this shared construction limit explanatory interpretation. Cross-seed Kendall ordering is less aligned with this account (reported correlation -0.486). These are descriptive associations, not an independent predictor or a causal account of trainability.

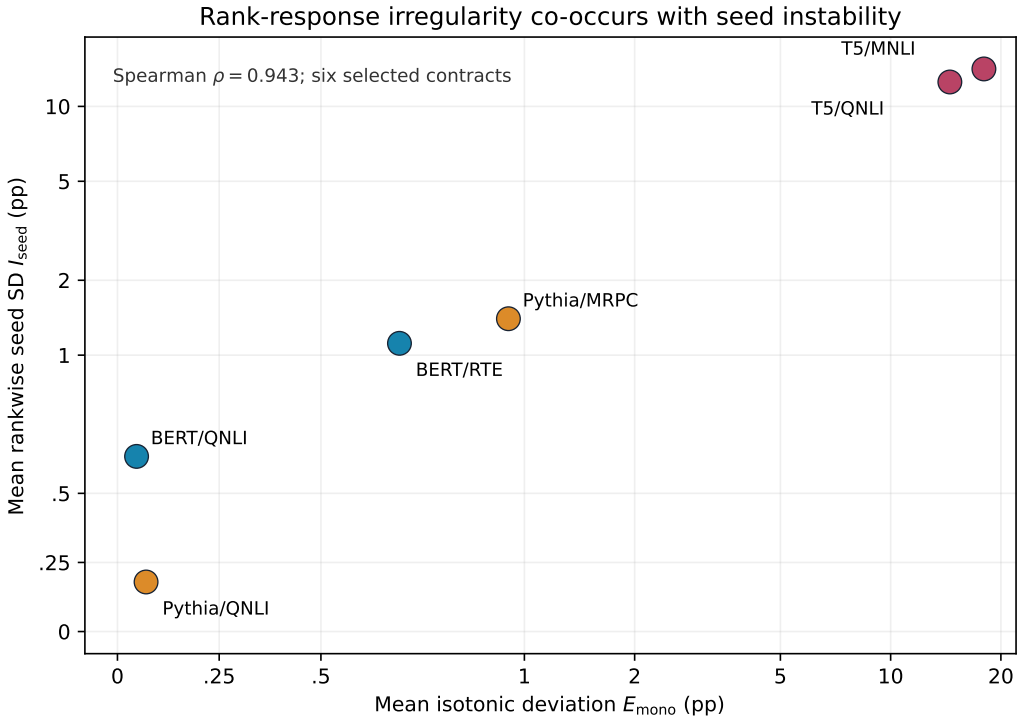


Figure A4: Isotonic rank-response residual and cross-seed spread across six frozen contracts. Both summaries use the same endpoint matrix; T5 failures and the small number of settings limit independent explanatory interpretation.

I Representation dynamics diagnostics

Selected checkpointed reruns compare known passing and failing outcomes with their same-seed, same-step rank-eight paths. In T5/QNLI, failing rank-six seeds 42 and 101 end at mean encoder CKA 0.522 and 0.723; passing rank-six seed 73 ends at 0.970. Passing rank-one seeds end at 0.973, 0.916, and 0.962. For Pythia/MRPC seed 101, passing rank one and failing rank six have endpoint CKA 0.796 and 0.733. Because these reruns were selected using known outcomes and evaluated relative to rank eight, the contrasts are descriptive rather than independent prediction tests.

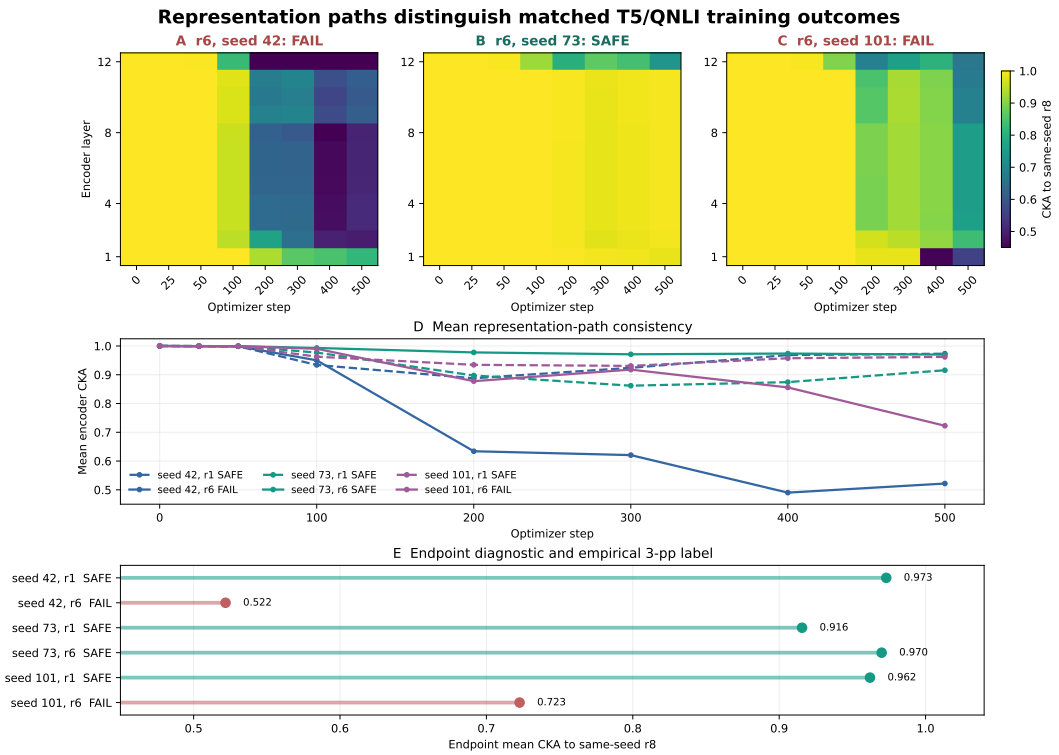


Figure A5: Selected T5/QNLI representation paths relative to same-seed, same-step rank-eight checkpoints. The panels describe known passing and failing endpoints. Selection by outcome and sparse snapshots preclude an independent predictive or causal interpretation.

Two scalar diagnostics do not consistently order these outcomes. The fixed-batch shadow-optimizer-step direction is a local probe, not the update taken on the actual training minibatch. At Pythia step 124, the local alignment is 0.033 for failing rank six and 0.015 for passing rank one. T5 rank-six endpoint effective update ranks are 2.75–2.78 for both passing and failing runs; the failing Pythia rank-six update has effective rank 5.18 versus rank one for the passing rank-one update. Neither scalar alone separates the selected outcomes, and the saved snapshots do not establish temporal causation.

J ADDITIONAL JOINT-RESOURCE EVIDENCE

Table A5 places the BERT/QNLI marginal and directly trained joint endpoints under the same reference and tolerance. At 1 pp, rank two/full-pool and rank eight/half-pool pass on every seed, while rank two/half-pool fails on every seed. At 3 pp, all three resource configurations pass. Each table entry reports the correct count and paired loss against the same-seed rank-eight/full-pool reference.

Figure A6 displays 18 measured and 18 unmeasured cells across the two BERT lattices. Seventeen measured cells pass at 3 pp. The historical BERT/QNLI rank-eight/random-quarter-pool point fails with a margin of -1.297 pp. At 1 pp, none of the three directly trained reduced-rank/half-pool arms passes the all-seed rule. Unmeasured intersections remain unlabeled. The retained `data/joint_results.csv` records the nine direct joint seed-runs and their references.

Table A5: Three-seed BERT/QNLI marginal and directly trained joint comparison on the 1,024-example evaluation set. Each entry gives correct count and paired loss against the same-seed rank-eight/full reference. At 1 pp both marginal points pass on all seeds, while every directly trained joint point fails. At 3 pp all three resource points pass.

Seed	$r2$/full (correct; loss)	$r8$/half (correct; loss)	$r2$/half (correct; loss)	1-pp labels
42	819 (+0.586 pp)	832 (-0.684 pp)	805 (+1.953 pp)	P / P / F
73	832 (+0.000 pp)	822 (+0.977 pp)	814 (+1.758 pp)	P / P / F
101	823 (+0.879 pp)	824 (+0.781 pp)	813 (+1.855 pp)	P / P / F

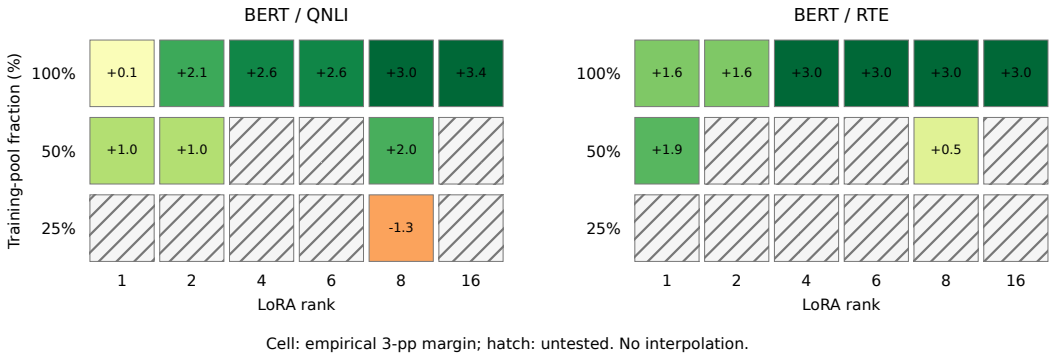


Figure A6: Sparse BERT parameter–training-pool lattices. Cell values are empirical 3-pp margins; hatching denotes unmeasured cells. Pool fractions describe available subsets, not a continuously inferred resource boundary.

K Equal-parameter-count endpoints

Table A6 provides the T5/QNLI scores used in the matched-count comparison. The native rank-six, custom $n = 1$ rank-six, and GID-block rank-six arms each contain 2,543,616 adapter parameters and use the same QNLI split. Each score is paired with the reference from the same seed, rather than a pooled mean. At 3 pp, seed-wise signatures are 010, 101, and 111, yielding all-seed labels FAIL, FAIL, and PASS. The first two arms fail on different seeds; the table does not establish that one parameterization is uniformly better. Appendix L reports the separate matched-tensor check.

L Matched-tensor audit of the T5 single-block operators

This diagnostic compares the vendored upstream `LinearWithRaLoRA` with `n_split=1` and the local true-block `RaLoRALinear` with `n_blocks=1`. Both use rank six, $\alpha = 6$, zero dropout, and copied base weights, bias, and adapter factors. The fixed float32 CPU input has shape $5 \times 3 \times 64$, and the output width is 48. The checks compare forward outputs, gradients, and one Adam update with learning rate 0.002, $\epsilon = 10^{-4}$, and zero weight decay.

No discrepancy beyond float32-scale numerical error was detected in these matched-state cases. The reported maximum differences are zero for the initial forward pass and 2.38×10^{-7} after one Adam update; Table A7 gives the factor and gradient comparisons, including the shared nonzero-B state. This is a single-layer numerical check, not a proof of equivalence across all shapes, devices, or complete training runs.

Source inspection reports Kaiming fan-in initialization for A, zero initialization for B, and scaling $\alpha/r = 1$ in both recipes. Their seed-to-tensor mappings differ: the native control uses deterministic SHA-256-derived per-layer seeds, while the custom control initializes sorted targets from a process-level random-number stream. A shared run-seed identifier therefore does not imply identical layerwise initial tensors. The diagnostic detects no matched-state operator discrepancy in the tested cases; it does not establish that the initialization mapping caused the historical endpoint differences.

Table A6: Three-seed T5/QNLI equal-parameter-count comparison on the 1,024-example evaluation set. Each arm gives its correct count and, in parentheses, its paired loss in pp against the same-seed rank-eight/full reference. The three rank-six arms each use $2,543,616$ adapter parameters (75% of the rank-eight reference). At 3 pp the seed-wise pass patterns are 010, 101, and 111.

Arm	Seed 42	Seed 73	Seed 101	Worst loss	All-seed
rank-8/full reference	877	830	860	0.000	reference
Native $r6$	512 ($+35.645$)	862 (-3.125)	824 ($+3.516$)	$+35.645$	FAIL
Custom $n = 1$ $r6$	884 (-0.684)	512 ($+31.055$)	870 (-0.977)	$+31.055$	FAIL
GID-block $r6$	864 ($+1.270$)	872 (-4.102)	859 ($+0.098$)	$+1.270$	PASS

Table A7: Maximum absolute differences in the tested matched-state single-layer operator cases. Both implementations use copied tensors, float32 CPU inputs, rank six, and the stated Adam update. These are numerical checks on the specified cases, not a full-model training-equivalence test.

Comparison	Max absolute difference
Forward at zero-B initialization	0
A and B gradients at zero-B initialization	0 and 2.98×10^{-8}
Factors after one Adam update	0 and 2.11×10^{-8}
Forward after one Adam update	2.38×10^{-7}
A and B gradients from the shared nonzero-B state	6.11×10^{-10} and 3.03×10^{-8}

Numerical values and source digests are identified in `analysis/resource_sufficiency_statistics/T5_N1_OPERATOR_AUDIT.md`.

M PROSPECTIVE BOUNDARY CHECKS AND SEARCH SCOPE

M.1 DECODER-ONLY GSM8K AND MBPP GRIDS

The retained study comprises 90 planned training seed-runs across GSM8K and MBPP, together with 92 held-out evaluation summaries. These are archive record counts, not 92 independently trained configurations. On GSM8K, rank eight/full-pool has 778, 771, and 760 correct out of 1,319; rank one/quarter-pool has 769, 786, and 757. The 15-configuration summary is identified by `data/prospective_gsm8k.csv`. At 1 pp, the passing rank sets are $\{1, 2, 8, 16\}$ for the quarter-pool row, $\{1, 2, 4, 8, 16\}$ for half-pool, and $\{1, 8, 16\}$ for full-pool. Every cell passes at 3 pp.

GSM8K generation and scoring. Qwen prompts use one user message containing `Solve the following math problem. Show your work and finish with '#### <answer>'.`, a blank line, and the question; the pinned tokenizer's chat template adds the generation prefix. Llama uses the same text prompt with a trailing blank line and no chat wrapper. Evaluation batches four prompts, left-pads, truncates inputs at 512 tokens, and uses EOS as the pad token. After Qwen chat-template rendering, no extra special tokens are added; Llama uses the tokenizer's default special tokens. One-beam greedy generation uses no sampling and a 512-new-token cap; it stops at the model's configured EOS or the cap, with no custom stop string. Only generated continuations are decoded, with special tokens skipped. The parser takes the text after the last `####`, reads its first line, removes commas and spaces, and accepts a leading integer or decimal with an optional minus sign. The same parser is applied to the reference answer; a missing delimiter, empty suffix, or unparseable answer is incorrect. The archived evaluation records contain counts and a hash of the predictions, not raw generations, so parse-failure and truncation counts or illustrative outputs cannot be recovered. In particular, the 7B base score of 4/1,319 (0.303%) cannot be decomposed from retained records into answer-quality, formatting, truncation, or parsing effects.

A separate reported diagnostic counts 13 within-seed cases in which a lower-rank endpoint exceeds a higher-rank endpoint by more than 1 pp. This accuracy-reversal count is distinct from the comparable-cell label violations in Appendix M.5; neither is a count of independent tasks.

On MBPP, rank eight/full-pool has 10, 12, and 9 correct out of 500. Every cell's worst-seed loss is at most 1 pp relative to this weak reference, and the grid has no accuracy reversal above 1 pp. Generated programs were evaluated in isolated Python subprocesses with a three-second CPU limit, a 1-GiB address-space limit, disabled user-site packages, and no inherited standard input; exceptions and timeouts count as failures. These execution restrictions are not a general security guarantee. The all-close map describes relative scores, not useful task performance.

M.2 PREREGISTERED 7B GSM8K SCALE-TRANSFER SLICE

The DeepSeek-R1-Distill-Qwen-7B slice crosses ranks $\{1, 2, 4, 8\}$, available-pool fractions $\{50, 100\%\}$, and seeds $\{42, 73, 101\}$, giving 24 training runs. The unadapted checkpoint scores 4/1,319 (0.303%). Rank-eight/full-pool reference accuracies are 76.270%, 76.194%, and 75.739%, averaging 76.068%; the recorded study reports that this clears its preregistered five-point reference-adequacy criterion. Table A8 contains the full endpoint comparison. The criterion qualifies the study's chosen reference; it does not establish an optimally tuned baseline.

All eight configurations pass at 3 pp. At 1 pp, every half-pool rank passes and the full-pool set is $\{1, 2, 8\}$; rank four/full-pool has a worst-seed loss of 1.440 pp. Four within-seed accuracy reversals exceed 1 pp: rank one over rank two at half-pool/seed 42 (1.061 pp), rank two over rank four at full-pool/seeds 73 and 101 (1.289 and 1.592 pp), and rank two over rank eight at full-pool/seed 101 (1.289 pp).

The retained evaluation protocol states that held-out evaluation began after all 24 training receipts were available and their hashes were recorded in `HELDOUT_SEAL.json`. The archive reports matching receipt digests and evaluation records for the unadapted checkpoint and 24 adapters on the same 1,319 examples. The derived table is identified by `data/prospective_gsm8k_7b.csv`. Because this grid is narrower than the 1.5B grid, the comparison is a targeted scale-transfer check rather than a matched scaling study.

Table A8: All eight 7B GSM8K tested configurations. Scores are Accuracy percentages on the same 1,319 examples. The paired loss is the worst over three seeds relative to the same-seed rank-eight/full reference; P/F denotes the observed three-seed rule at 3 pp / 1 pp. Budget is adapter parameters / unique examples.

Rank	Data	P/N	Seed 42	Seed 73	Seed 101	Worst loss	3 / 1
1	100%	315,392/7,473	76.346	75.208	75.815	0.986	P / P
2	100%	630,784/7,473	75.739	76.042	77.028	0.531	P / P
4	100%	1,261,568/7,473	76.497	74.754	75.436	1.440	P / F
8	100%	2,523,136/7,473	76.270	76.194	75.739	0.000	P / P
1	50%	315,392/3,736	77.635	76.346	75.284	0.455	P / P
2	50%	630,784/3,736	76.573	76.573	75.588	0.152	P / P
4	50%	1,261,568/3,736	77.104	75.891	75.284	0.455	P / P
8	50%	2,523,136/3,736	76.725	76.952	75.360	0.379	P / P

Table A9: All eight Llama-3.1-8B GSM8K tested configurations. Scores are Accuracy percentages on the sealed 1,319-example test set. Worst paired loss is relative to same-seed rank-eight/full-pool; P/F denotes the observed three-seed rule at 3 pp / 1 pp.

Rank	Data	Seed 42	Seed 73	Seed 101	Worst loss	3 / 1
1	100%	53.601	52.843	54.056	2.199	P / F
2	100%	53.450	54.208	54.359	1.061	P / F
4	100%	54.587	54.511	55.800	0.531	P / P
8	100%	54.511	55.042	55.042	0.000	P / P
1	50%	51.327	53.677	53.298	3.184	F / F
2	50%	52.843	53.904	53.222	1.820	P / F
4	50%	52.009	54.511	54.966	2.502	P / F
8	50%	53.753	55.951	54.208	0.834	P / P

M.3 LLAMA-3.1-8B CROSS-FAMILY GSM8K GRID

The Llama slice uses ranks $\{1, 2, 4, 8\}$, available-pool fractions $\{50, 100\%\}$, and seeds $\{42, 73, 101\}$, for 24 training runs. The sealed evaluation records the unadapted checkpoint and all adapters on the 1,319-example GSM8K test set. The unadapted score is 44/1,319 (3.336%), and same-seed rank-eight/full-pool accuracies are 54.511%, 55.042%, and 55.042% (mean 54.865%). The study reports that this clears its five-point adequacy criterion.

At 3 pp, rank two/half-pool passes with worst-seed loss 1.820 pp, while rank one/half-pool fails at 3.184 pp. Rank one/full-pool and rank eight/half-pool pass separately, so their joint reduction is a measured composition counterexample. The map is nevertheless an upper set at the two predeclared tolerances, 1 and 3 pp. This endpoint statement does not extend to every intermediate tolerance; Appendix N reports a non-upper interval. The retained manifest identifies 24 training receipts and 25 evaluation records.

M.4 EARLIER PROSPECTIVE CHECKS AND SEARCH SCOPE

The BERT-medium/BoolQ study specifies ranks $\{1, 2, 4, 8\}$, nested pool fractions $\{25, 50, 100\%\}$, and three seeds, for 36 runs of 500 steps. Evaluation uses 3,270 validation examples. All 12 configurations pass at 3 pp; only rank four/full-pool and rank eight/full-pool pass at 1 pp. The rank-eight/full-pool counts are 2,140, 2,132, and 2,135, versus 2,033 correct for the majority-class baseline. This weak reference limits the interpretation of relative compression. The summary is identified by `data/prospective_boolq.csv`.

The separately frozen 1.5B CommonsenseQA grid uses the same rank–pool shape, 300 steps, and 1,221 validation questions with constrained one-token choice scoring. Rank-eight/full-pool counts are 663, 656, and 673. At 3 pp, only rank four/full-pool and rank eight/full-pool pass; no reduced-pool point passes. At 1 pp, only rank eight/full-pool passes. The summary is identified by `data/prospective_commonsenseqa.csv`. Both earlier grid summaries retain worst-of-

Table A10: Exact nearest-upper-set distance for six complete grids at the two predeclared tolerances. $D_\uparrow$ is the minimum fraction of labels differing from an upper set. Minimal points describe observed upper sets; listed edits describe a comparison map and do not change measured outcomes. MBPP uses a weak reference.

Grid	δ	Safe	$D_\uparrow$	Upper?	Min. antichain / repair
BoolQ	1 pp	2/12	0	Yes	$r4/D100\%$
BoolQ	3 pp	12/12	0	Yes	$r1/D25\%$ (all safe)
CommonsenseQA	1 pp	1/12	0	Yes	$r8/D100\%$
CommonsenseQA	3 pp	2/12	0	Yes	$r4/D100\%$
GSM8K 1.5B	1 pp	12/15	0.200	No	repair: $r2/D100\%$, $r4/D25\%$, $r4/D100\%$ (FAIL→PASS)
GSM8K 1.5B	3 pp	15/15	0	Yes	$r1/D25\%$ (all safe)
GSM8K 7B	1 pp	7/8	0.125	No	repair: $r4/D100\%$ (FAIL→PASS)
GSM8K 7B	3 pp	8/8	0	Yes	$r1/D50\%$ (all safe)
MBPP*	1 pp	15/15	0	Yes	$r1/D25\%$ (all safe)
MBPP*	3 pp	15/15	0	Yes	$r1/D25\%$ (all safe)
Llama-3.1-8B	1 pp	3/8	0	Yes	$r4/D100\%, r8/D50\%$
Llama-3.1-8B	3 pp	7/8	0	Yes	$r1/D100\%, r2/D50\%$

three losses and labels; their complete seed-wise counts remain in the original archives. Neither grid appears in the main decoder-map figure.

On CommonsenseQA, a seed fail-fast, confirmed-safe Pareto search recovers the exhaustive 3-pp frontier $\{(r4, 100\%)\}$ using 16 of 36 seed-runs during its search phase. The remaining runs were subsequently completed to provide the exhaustive comparison, so the validation still used all 36 runs. Independent-marginal and monotone-stop heuristics also selected this point using 10 and 8 seed-runs under their respective objectives. They did not fail on this task. A guided ordering preserved exactness but saved no additional seed-runs over fixed ordering in eight retrospective grid–tolerance replays. These results delimit the search evidence; they do not establish a general speed advantage.

The manuscript source identifies rendering scripts, endpoint summaries, joint-result records, matched-count comparisons, and prospective-grid summaries. Full checkpoint access and unrecoverable historical environment fields remain separate reproducibility limitations; the existence of a summary table alone does not resolve them.

M.5 Exact finite resource-order audit

For each complete grid, the all-seed label is PASS when its worst paired loss is at most δ. The reported audit enumerates the upper sets of the measured coordinate order and minimizes label disagreement. Table A10 covers six grids at the two predeclared tolerances, giving twelve grid–tolerance cases. The nearest upper set is unique in these twelve cases; this statement is not extended to the intermediate-tolerance sweep.

For an observed upper set, its minimal passing configurations reconstruct its labels by upward closure. Llama has two minimal points at each predeclared tolerance; the other upper-set maps in this table have one. For a non-upper map, the listed edits describe its nearest comparison map, not revisions to the measured outcomes. The export paths recorded for these comparisons are `data/finite_resource_order_audit_complete.csv` and `data/finite_resource_order_cells_complete.csv`.

At 3 pp, both Qwen GSM8K grids are all-pass, so their zero defects do not reveal a nontrivial resource boundary. Llama remains ordered at 3 pp despite its marginal-to-joint failure. At 1 pp, Qwen 1.5B has 13 comparable-cell label violations but needs three label changes to reach an upper set; Qwen 7B has five such violations but needs one change. These counts use all-seed labels on the coordinate order, not the within-seed accuracy reversals reported earlier. MBPP's zero defect remains a weak-reference result because its reference averages 2.07% Pass@1.

Table A11: GSM8K label-map sensitivity to selected tolerances, with trained endpoints and references held fixed. PASS counts and $D_{\uparrow}$ are computed on each complete grid. The final column counts false-sufficient labels introduced by upward closure. Intermediate thresholds are descriptive sensitivity checks, not newly selected primary tolerances.

Grid	δ (pp)	PASS cells	$D_{\uparrow}$	Closure false-safe
Qwen 1.5B	1.0	12/15	0.200	3
Qwen 1.5B	1.1	14/15	0.067	1
Qwen 1.5B	1.5	14/15	0.067	1
Qwen 1.5B	2.0	14/15	0.067	1
Qwen 1.5B	2.5	15/15	0	0
Qwen 1.5B	3.0	15/15	0	0
Qwen 7B	1.0	7/8	0.125	1
Qwen 7B	1.1	7/8	0.125	1
Qwen 7B	1.5	8/8	0	0
Qwen 7B	2.0	8/8	0	0
Qwen 7B	2.5	8/8	0	0
Qwen 7B	3.0	8/8	0	0
Llama-3.1-8B	1.0	3/8	0	0
Llama-3.1-8B	1.1	4/8	0	0
Llama-3.1-8B	1.5	4/8	0	0
Llama-3.1-8B	2.0	5/8	0.125	1
Llama-3.1-8B	2.5	6/8	0.125	1
Llama-3.1-8B	3.0	7/8	0	0

N TOLERANCE SENSITIVITY ON MODERN DECODER GRIDS

This analysis keeps the trained endpoints and same-seed references fixed and recomputes labels as δ changes. Table A11 includes selected intermediate tolerances alongside the predeclared 1- and 3-pp values. For each map it reports the number of passing cells, the nearest-upper-set defect, and false-sufficient labels introduced by upward closure. Nearest-upper-set distance and upward closure are different operations: the former may add or remove passing labels, whereas the latter only adds them. The sweep describes threshold sensitivity, not uncertainty over new seeds or test samples.

The reported transitions explain how the maps change. For Qwen 1.5B, $(r4, 25\%)$ and $(r2, 100\%)$ enter at approximately 1.061410 pp, reducing $D_{\uparrow}$ from $3/15$ to $1/15$; $(r4, 100\%)$ enters at approximately 2.274450 pp and removes the remaining defect. For Qwen 7B, $(r1, 100\%)$ enters at approximately 0.985595 pp, reducing the defect from $2/8$ to $1/8$. At 1 pp that cell already passes: the remaining excluded cell is $(r4, 100\%)$, which enters at approximately 1.440485 pp and removes the defect.

Llama is an upper set at 1 and 3 pp but not throughout the interval between them. When $(r2, 50\%)$ enters at approximately 1.819560 pp, the higher-rank $(r4, 50\%)$ cell still fails. The map becomes ordered again when that cell enters at approximately 2.501895 pp. Thus the passing set grows with tolerance, but its distance to an upper set need not decrease monotonically. Labels should be computed from the retained full-precision losses, not the rounded transition values printed here.

O AVAILABLE TRAINING POOLS AND DISTINCT EXAMPLES VISITED

The frozen manifest identifies the available training pool. The receipt field `unique_examples` stores this pool size, not a count of distinct examples actually visited during optimization. The GSM8K samplers traverse a seeded NumPy permutation and draw a new permutation only after exhausting the current one. Replaying the specified index stream for all three seeds gives the distinct-visit counts in Table A12; these are sampler-derived counts rather than newly measured training logs.

Each run has 900 optimizer steps and 7,200 sampled example positions. Qwen's full pool has 7,473 examples, of which 7,200 are visited. The 1,868- and 3,736-example reduced pools are fully visited, giving visit-reduction factors of 3.854 and 1.927. Both Llama pools are fully visited: 6,726

Table A12: Available training pools and sampler-replayed distinct visits in GSM8K. All runs use 900 steps and 7,200 example positions. The visit-reduction factor is the full-pool reference's distinct visits divided by those of the listed pool. Visit counts are derived from the retained sampler, not read from the pool-size receipt field.

Grid	Pool fraction	Available N	Exposures	Distinct visited	Visit ratio vs full
Qwen 1.5B	25%	1,868	7,200	1,868	3.854×
Qwen 1.5B	50%	3,736	7,200	3,736	1.927×
Qwen 1.5B	100%	7,473	7,200	7,200	1.000×
Qwen 7B	50%	3,736	7,200	3,736	1.927×
Qwen 7B	100%	7,473	7,200	7,200	1.000×
Llama-3.1-8B	50%	3,363	7,200	3,363	2.000×
Llama-3.1-8B	100%	6,726	7,200	6,726	1.000×

examples for full-pool and 3,363 for half-pool, a factor of two. Pool fraction, exposure count, and distinct visits are therefore reported separately. None of these ratios establishes fewer optimizer steps or lower total compute. Per-seed replay counts are identified by `decoder_training_subset_vs_seen.csv`.

P OFFLINE RANK-SEARCH REPLAY SENSITIVITY

Table A13 compares two deterministic stop-after-first-failure rules on the same 108 default-contract endpoints. The primary rule starts from the known passing rank-eight reference, checks ranks $6, 4, 2, 1$, and returns the last passing rank reached. The alternative starts at rank 16 and checks $16, 8, 6, 4, 2, 1$; if its first evaluated point fails, it may return no passing rank. Both are retrospective diagnostics, not prospective policy evaluations or demonstrations of saved training runs.

Table A13: Offline stop-after-first-failure replays against the exhaustive smallest passing rank. The reference-start policy treats rank eight as a known pass, then checks ranks $6, 4, 2, 1$ and returns the last passing rank. The descending diagnostic instead starts at rank 16 and checks $16, 8, 6, 4, 2, 1$. These are specified counterfactual rules on existing endpoints, not prospective policy evaluations.

δ	Setting	Exhaustive min.	16-start output	Known-$r8$ output
3 pp	BERT/QNLI	1	1	1
3 pp	BERT/RTE	1	1	1
3 pp	Pythia/QNLI	1	1	1
3 pp	Pythia/MRPC	1	8	8
3 pp	T5/QNLI	1	NONE	8
3 pp	T5/MNLI	2	NONE	8
1 pp	BERT/QNLI	2	2	2
1 pp	BERT/RTE	4	4	4
1 pp	Pythia/QNLI	2	2	2
1 pp	Pythia/MRPC	8	NONE	8
1 pp	T5/QNLI	1	NONE	8
1 pp	T5/MNLI	2	NONE	8

The 16-start diagnostic misses the exhaustive minimum in 3/6 settings at 3 pp and 3/6 at 1 pp. Starting from the known passing reference reduces these counts to 3/6 and 2/6, respectively. The two T5 misses persist at 1 pp; Pythia/MRPC is correctly returned as rank eight at that tolerance because no lower tested rank passes.

Q COMMON-RANK SEARCH REPLAY BEFORE AND AFTER LOCAL LR SELECTION

This comparison uses only ranks measured under both the default and shared local-LR policies. The common sets are $\{1, 2, 4, 8\}$ for T5/QNLI and $\{1, 2, 4, 6, 8, 16\}$ for Pythia/MRPC. Both replays use the known passing default-contract rank-eight reference as their anchor, inspect lower common ranks in descending order, and stop at the first failure. All losses retain this frozen reference; the lower-rank labels come from the respective policy. Selection and reporting reuse inspected development data.

Table A14: Rank-eight-anchored stop replay before and after shared local-LR selection, restricted to common tested ranks. Common sets are $\{1, 2, 4, 8\}$ for T5/QNLI and $\{1, 2, 4, 6, 8, 16\}$ for Pythia/MRPC. P/F denotes all-seed status; Min. PASS is exhaustive on the common set.

Setting	δ	LR policy	Min. PASS	Trace below rank 8	Returned	Miss?
T5/QNLI	1 pp	Default	1	$4 : F$	8	Yes
T5/QNLI	1 pp	Shared local	1	$4 : P, 2 : F$	4	Yes
T5/QNLI	3 pp	Default	1	$4 : F$	8	Yes
T5/QNLI	3 pp	Shared local	1	$4 : P, 2 : P, 1 : P$	1	No
Pythia/MRPC	1 pp	Default	8	$6 : F$	8	No
Pythia/MRPC	1 pp	Shared local	8	$6 : F$	8	No
Pythia/MRPC	3 pp	Default	1	$6 : F$	8	Yes
Pythia/MRPC	3 pp	Shared local	1	$6 : P, 4 : F$	6	Yes

On T5/QNLI, local selection changes the 3-pp return from rank eight to the common-grid minimum rank one. At 1 pp, it changes the return from eight to four but still misses rank one. On Pythia/MRPC, the 3-pp return falls from eight to six while the minimum remains one; at 1 pp both policies return the correct minimum, eight. Local selection therefore repairs some errors without removing all missed minima in these comparisons. The evidence concerns these common measured ranks and this stopping rule, not an independently validated selection algorithm.