ARE LEARNED LORA RANK ALLOCATIONS RELIABLE?
A CONTROLLED AUDIT ON LLAMA-2-7B

Anonymous authors
Paper under double-blind review

ABSTRACT

Most learning-based and heuristic LoRA rank-allocation methods rest on an uncontrolled comparison: in the coupled-seed protocol, each candidate is evaluated on its random initialization, so allocation and seed are inseparable in one number.

We propose a seed-decoupled protocol: one master seed derives three independent streams fixing scoring-head initialization, LoRA initialization, and data order/dropout; each candidate is compared pairwise with a shared equal-budget uniform_r8 reference. Within a master seed only the allocation varies, structurally removing the shared component that accounts for 94.1% of run-to-run variance. Cost: one extra reference run per master seed; gain: 4.7–6.6× tighter comparison error at equal runs.

On the same panel and budget, the two protocols disagree: the coupled one puts all four search methods ~ 1 point above uniform rank 8; under ours, the advantage disappears (the sole equal-budget search output retested: -0.63 pp, Holm $p = 0.578$). The divergence comes from the training protocol, not the search methods.

Fixing the parameter budget and rank multiset while permuting only internal positions shifts accuracy by 4.3 points—four times any method difference this field seeks to detect—and the search output we retest ranks below its own three hand permutations: reported gains thus fall within the variance the literature leaves uncontrolled.

Our protocol lets a new rank-allocation method be audited before it claims superiority over uniform rank; fair comparison needs at least three conditions: equal parameter budget, fixed rank multiset, shared reference. Without all three, a comparison may measure placement alone.

1 INTRODUCTION

LoRA (Hu et al., 2021) freezes the pretrained matrix and trains a low-rank update $\mathbf{BA}$. Layer-wise rank allocation assigns different ranks to different layers, on the assumption that uniform ranks waste budget on unimportant modules, and the family reports consistent gains over uniform baselines (Zhang et al., 2023; Liu et al., 2024). Those gains come from single-seed or few-seed evaluations without paired designs, so the allocation-specific effect is not separated from the seed variance it shares with its own baseline, and seldom with multiple-comparison correction.

Two further gaps motivate this audit: allocation effects may be regime-conditional, and no characterization exists to diagnose catastrophic failure before deployment.

Under the regime used in practice — native scaling ($s = \alpha/r$) with seeds coupled across runs — uniform rank-4 outscored uniform rank-8 on MNLI; tracing that gap exposed two protocol confounds, an implicitly reseeded classification head and a fixed $\alpha = 16$ making the effective scaling $16/r$ track capacity. We therefore audit allocation *outputs* under a protocol that removes both, treating *search identifies robust allocations* as a falsifiable hypothesis.

Contributions. (C1) A protocol that makes effects attributable. A seed-decoupled design (Section 3.1) derives head, adapter, and data seeds independently via SHA-256, so the dominant

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.

common-mode component of run-to-run variance (94.1%) cancels in paired comparisons, whose sd falls by 2.5–$8.3\times$. **(C2) Placement, not composition, dominates.** A layered candidate design (Section 3.2) fixes the rank multiset, the budget, or neither, layer by layer; five candidates sharing one multiset span 4.27 of the panel's 4.39 pp, leaving no room for a composition effect of comparable size. **(C3) A single-seed optimum does not survive reseeding.** We instantiate two paradigms—density-estimation Bayesian optimization (TPE) and surrogate-assisted evolutionary search (SAGA-RA, audited off its design axis). The equal-budget TPE optimum, selected at one seed, is the least reproducible of our eight candidates and beats uniform at no Holm-corrected level ($\Delta = -0.63$ pp, Holm $p = 0.578$); on CoLA, the only Holm-significant effect of a multiset permutation is harmful. Spectral diagnostics (Section 5) characterize how collapsed adapters redistribute update energy across depth; the study spans 301 unique configurations (Appendix A.3). We propose no new rank-allocation method: our contribution is the protocol and the audit it enables.

2 Related Work

Adaptive rank allocation. Layer-wise allocation emerged as a response to LoRA's uniform-rank baseline (Hu et al., 2021): importance-based pruning of a trainable budget (Zhang et al., 2023; Ding et al., 2023; Liu et al., 2024), search-free derivation from pretrained weight structure (Zhang et al., 2025; Liu et al., 2026), coalition-aware importance scoring (Zhao et al., 2025), and statistical rank selection (Gao & Tan, 2026). These methods refine how importance is estimated; we ask whether the resulting allocations reproduce their advantage under reseeding. We do not re-score published results — re-implementing each method under its own best configuration is itself uncontrolled — but instantiate the family's two representative philosophies in our own protocol, so every claim here is internal to it.

Adapter scaling. rsLoRA (Kalajdzievski, 2023) proposes $s = \alpha/\sqrt{r}$ in place of the PEFT default $s = \alpha/r$, which shrinks high-rank adapters' effective learning rates exponentially. We hold s fixed across modules (Section 3.3), so our constant2/constant8 comparison varies a global constant, not the functional form, and does not test rsLoRA.

Reproducibility and failure modes in PEFT. To our knowledge, audits within this family remain scarce: results are reported as single seeds or 3–5-seed averages, without variance decomposition and with the source of cross-run variance unidentified. In fine-tuning more broadly, run-to-run variance on BERT has been decomposed into weight initialisation and data order over $2{,}100$ trials (Dodge et al., 2020). We decompose three seed streams on Llama-2-7B, find one dominant at 94.1%, and measure how that component enters a *comparison* rather than a single run. Our protocol (Section 3.1) derives them independently from a master seed, so paired comparisons cancel the common-mode component. Collapse is noted only anecdotally — RSRA (Liu et al., 2026) reports AdaLoRA degenerating toward majority-class output on Mistral-7B — and we add a structural description: collapsed adapters concentrate update energy in the two shallowest layers (18–26% vs. $\approx$ 2–3% normally; Section 5).

3 Method

3.1 Seed-Decoupled Protocol

Setting. We audit allocations under the regime the field uses in practice, native scaling ($s = \alpha/r$), with head, adapter and data seeds determined jointly by a single seed for each run, and under the decoupled protocol below, which cancels that shared component in paired comparisons.

Motivation. A pilot single-source probe (Appendix A.7) attributes 92.2% of cross-seed MNLI variance to score-head initialization, against 2.2% for adapter initialization and 5.6% for data ordering; a random-effects decomposition on the main panel (Appendix A.8) puts the shared seed-block component at 94.1%, the figure we quote hereafter. Candidates sharing a master seed thus share an almost identical score head, and with it a common-mode offset that averaging dilutes but only *pairing* removes algebraically. This is why such a gap cannot be settled by re-running under fresh seeds: whatever the allocation effect is, it sits underneath a component an order of magnitude larger.

Derivation scheme. Each run is indexed by a `master_seed` $m \in [0, 2^{31}-1)$. Three independent seed streams are derived deterministically via SHA-256:

$$s_{\text{head}} = \text{derive}(\text{“score_head_init”}, m), \tag{1}$$

$$s_{\text{lora}} = \text{derive}(\text{“lora_init”}, m), \tag{2}$$

$$s_{\text{data}} = \text{derive}(\text{“data_order_dropout”}, m), \tag{3}$$

where $\text{derive}(k, m)$ hashes the label k concatenated with m under SHA-256 and reduces a fixed 8-byte prefix of the digest modulo $2^{31} - 1$; the exact byte encoding, separator and integer conversion are fixed by the released implementation, and each run's log records the three resulting seeds so that any run can be checked against it. All three labels depend on m alone, so within a seed block every candidate receives the same s_{head}, s_{lora} and s_{data}; the logged `lora_init_hash` still differs across candidates, because the rank pattern sets the shapes of the $\mathbf{A}$ and $\mathbf{B}$ factors that the shared stream fills. This is deliberate: head and data noise appear as a common offset that the paired estimator removes exactly, and adapter initialization comes from one reproducible stream, so no candidate benefits from a luckier draw. The three streams are decoupled from each other, not from the candidate.

Paired-difference estimator. For candidate a and baseline b (typically `uniform_r8`), the paired difference at master seed m is

$$\Delta(a, m) = \text{acc}(a, m) - \text{acc}(b, m). \tag{4}$$

The mean $\bar{\Delta}(a)$ aggregates over the seed set $\mathcal{M}$ ($|\mathcal{M}| = 8$ for the MNLI constant2 panel; $|\mathcal{M}| = 7$ wherever a collapsed seed block is dropped, Section 4.6). Head-initialization noise enters both terms identically and cancels, by an amount set by the cross-candidate correlation ρ that shared seeds induce; empirically $\rho \in [0.914, 0.993]$, giving sd reductions of 2.5–8.3$\times$ under constant2 (median $4.7\times$) and 3.1–15.6$\times$ under constant8 (median $6.6\times$), per candidate in Table 2. Unlike ordinary seed blocking, the construction yields candidate-specific within-block differences, which is what permits direct inference once common score-head variation is gone.

Inference. We report 95% bootstrap confidence intervals (1000 resamples of $\mathcal{M}$), the paired standard deviation $\text{sd}_{\text{paired}} = \text{sd}(\{\Delta(a, m)\}_{m \in \mathcal{M}})$ as a stability index, sign-consistency across seeds, and Holm-corrected two-sided exact sign tests within each candidate family (null: $\Pr(\Delta > 0) = 0.5$). Seed-blocked permutation tests (10,000 permutations of candidate labels within each seed block) assess the global effect span.

3.2 Layered Candidate Design

Gene-to-module mapping. Each allocation is an 8-gene *chromosome* $[r_0, \ldots, r_7]$ in which gene g sets the rank of every module in depth quartile $\lfloor g/2 \rfloor$. With 32 layers and LoRA on `q_proj`/`v_proj` only, each quartile spans 8 consecutive layers $\times$ 2 projections $=$ 16 modules: gene $2q$ controls `q_proj` in layers $8q$ to $8q + 7$ and gene $2q + 1$ controls `v_proj` over the same span, for $q \in \{0, 1, 2, 3\}$; the coarse granularity shrinks the search space from $5^{64} \approx 10^{44}$ to 5^8 while keeping the depth-segment structure the allocation literature actually resolves (Appendix A.2 justifies the choice and bounds what it costs).

Three layers separate *where* from *what* and *how much*. **Layer A** holds the rank multiset fixed (counted with multiplicity), so candidates differ only by permutation and parameter budgets are identical — e.g. `tpe_w2025` $= [2, 16, 16, 4, 1, 16, 1, 8]$ against `random64_b` $= [1, 1, 16, 2, 16, 16, 4, 8]$, both multiset $\{1, 1, 2, 4, 8, 16, 16, 16\}$. **Layer B** holds only the budget fixed ($\sum_g r_g = 64$) and varies composition, which is what tests search output against equal-budget random draws of a different multiset (`random64_a` $= [2, 16, 2, 16, 4, 4, 4, 16]$). **Layer C** has unequal budgets (`sagara_w2025`, sum 22); such comparisons confound capacity with structure and stay in the appendices. The mismatch is consequential and we flag it in both directions: SAGA-RA optimizes an accuracy–budget trade-off and `sagara_w2025` is its rank-efficient endpoint, whereas every panel in this paper scores equal-budget (sum-64) accuracy, so our panels never evaluate SAGA-RA on the axis it was built for.

Table 1: Candidate configurations and layered design (gene-to-module mapping in Section 3.2). All candidates except `sagara_w2025` (Layer C, sum 22) have equal budgets. Layer A shares `tpe_w2025`'s rank multiset and differs only in placement; Layer B keeps the budget and changes the multiset.

Candidate	Layer	Chromosome	Sum	Multiset Match	Source
`uniform_r8`	B	$[8, 8, 8, 8, 8, 8, 8, 8]$	64	No	baseline
`tpe_w2025`	A	$[2, 16, 16, 4, 1, 16, 1, 8]$	64	Yes	TPE:seed2025
`tpe_qvswap`	A	$[16, 2, 4, 16, 16, 1, 8, 1]$	64	Yes	q/v swap
`tpe_grev`	A	$[1, 8, 1, 16, 16, 4, 2, 16]$	64	Yes	layer reversal
`tpe_perm`	A	$[1, 16, 2, 16, 8, 16, 4, 1]$	64	Yes	random permutation
`random64_a`	B	$[2, 16, 2, 16, 4, 4, 4, 16]$	64	No	rejection sampling
`random64_b`	A	$[1, 1, 16, 2, 16, 16, 4, 8]$	64	Yes	rejection sampling
`sagara_w2025`	C	$[1, 4, 8, 2, 2, 2, 2, 1]$	22	No	SAGA-RA:seed2025

Candidate panel. Table 1 lists the eight candidates, and its **Source** column fixes the terminology used throughout this paper. Two of them are search outputs: `tpe_w2025` (TPE) and `sagara_w2025` (SAGA-RA). The remaining five are constructed, not searched. `tpe_qvswap`, `tpe_grev` and `tpe_perm` are *manual* permutations of `tpe_w2025`'s multiset — a q/v swap, a layer reversal, and a uniform random permutation, each specified before the panel was run and so fixed without knowledge of any outcome. `random64_a` and `random64_b` are rejection-sampled equal-budget draws, and `uniform_r8` is the baseline. This supports three comparisons: whether the searched placement beats manual permutations of its own multiset, whether it beats equal-budget random allocations of different composition, and whether any candidate reliably beats uniform.

3.3 Adapter Scaling Regimes

The LoRA forward pass is $\mathbf{y} = \mathbf{W}_0\mathbf{x} + s \cdot \mathbf{B}(\mathbf{A}\mathbf{x})$ with $\mathbf{B}$ zero-initialized. Rank-dependent choices of s such as the PEFT default α/r conflate rank with scale, leaving it unclear whether differences come from capacity or from learning-rate heterogeneity. With $\alpha = 16$ throughout we evaluate three regimes: **native** ($s = \alpha/r$, the default, where high-rank modules receive smaller effective updates), **constant2** ($s = 2.0$, weak) and **constant8** ($s = 8.0$, strong). Because the adapter is zero-initialized only at step 1, constant2 and constant8 give identical first-step contributions but different subsequent trajectories and gradients, making them distinct optimization regimes rather than global learning-rate rescalings.

Main results use constant2, with constant8 as a regime-dependence check over the same candidates; the MNLI panel runs all 8 candidates under both (Section 3.4). The native regime. We additionally catalogued what allocation search returns under the regime used in practice — native scaling, $s = \alpha/r$, coupled seeds. Each candidate in this corpus is trained exactly once, so per-candidate replication is limited to the three master seeds of its parent search; native results are therefore reported as pooled estimates, not as per-candidate inference. Full enumeration, per-search tables and regime details: Appendix A.10.

3.4 Tasks, Hyperparameters, and Computational Budget

We fine-tune `meta-llama/Llama-2-7b-hf` on MNLI, QNLI and CoLA, sampling $N = 5{,}000$ training examples uniformly from each official training split. Everything is frozen except a randomly initialized score head (tied to s_{head}) and LoRA adapters on `q_proj`/`v_proj` (64 modules over 32 layers). Training runs 2 epochs at per-device batch size 8 with gradient accumulation 2 (effective 16), learning rate 10^{-4}, linear schedule without warmup, weight decay 0, `lora_dropout` $= 0.05$, gradient checkpointing enabled, and bf16 — 624 optimizer steps for every run (Appendix A.3 gives the arithmetic). We evaluate on the official validation splits: MNLI `validation_matched` (9,815 examples), QNLI (5,463) and CoLA (1,043), reporting evaluation loss as a sample-weighted mean over batches so that a ragged final batch does not bias it. Trainable parameters obey $|\theta_{\text{train}}| = 65{,}536 \sum_g r_g + 4096\,|\mathcal{Y}|$ for label count $|\mathcal{Y}|$, so the score head contributes $12{,}288$ parameters on MNLI and $8{,}192$ on QNLI and CoLA; the identity holds exactly on

every logged run and is the gate we use to catch misconfigured adapters. The eight master seeds are $\{2025, 42, 777, 11, 22, 33, 44, 55\}$.

The study comprises 301 unique configurations; the full inventory, pilot-cell reuse and log-record accounting are in Appendix A.3. Two panels bear on the inference below. The MNLI panel is 8 candidates $\times$ 8 seeds $\times$ 2 regimes $=$ 128 cells, of which the *MNLI panel* scores the 7 sum-64 candidates: 56 cells under constant2 and 49 under constant8, after the `m777` block is excluded (Section 4.6); `sagara_w2025` is excluded throughout as a capacity confound. QNLI and CoLA each *run* 7 candidates $\times$ 8 seeds under constant2, and each enters analysis with the affected seed block removed (except CoLA), giving 56 cells for CoLA and 49 for QNLI. (Section 4.5).

4 Results

4.1 Paired-Difference Analysis: Main Effects

We first ask: *does any allocation, searched or otherwise, reliably outperform the uniform baseline across reseeded runs?* All candidate effects are paired differences against the uniform baseline at a common master seed, $\Delta(a, m) = \text{acc}(a, m) - \text{acc}(\texttt{uniform_r8}, m)$. Table 2 reports them for the six sum-64 candidates (excluding `sagara_w2025` as a capacity confound), aggregated over eight master seeds with Holm correction within the family, together with the noise reduction pairing delivers.

Table 2: Paired-difference statistics, MNLI constant2 panel (8 seeds, vs. `uniform_r8`, 6 sum-64 candidates). **Red.** is the unpaired-to-paired sd ratio and ρ the cross-candidate correlation induced by the shared score head, whose common-mode component is 94.1% of run-to-run variance (Appendix A.8). Four candidates reach Holm-corrected significance, three favourably and one adversely. Sign p is the two-sided exact binomial test at $\pi = 1/2$; at $n = 8$ the attainable values are $\{0.0078, 0.0703, 0.2891, 0.7266, 1.0000\}$. Constant8 figures and the full raw-sd breakdown are in Appendix A.4.

Candidate	Mean Δ (pp)	95% CI (pp)	Paired sd (pp)	Wins	Sign p	Holm p	SNR	Red.	ρ
`random64_a`	$+1.21$	$[+0.96, +1.43]$	0.37	8/8	0.0078	**0.0469**	3.27	$8.31\times$	0.993
`tpe_perm`	$+1.09$	$[+0.82, +1.33]$	0.40	8/8	0.0078	**0.0469**	2.73	$7.98\times$	0.993
`tpe_grev`	$+0.82$	$[+0.49, +1.19]$	0.54	8/8	0.0078	**0.0469**	1.52	$5.35\times$	0.984
`tpe_qvswap`	$+0.40$	$[+0.00, +0.90]$	0.70	5/8	0.7266	0.7266	0.57	$3.96\times$	0.974
`tpe_w2025`	-0.63	$[-1.65, +0.10]$	1.38	2/8	0.2891	0.5781	0.46	$2.45\times$	0.914
`random64_b`	-3.18	$[-3.78, -2.62]$	0.90	0/8	0.0078	**0.0469**	3.53	$2.88\times$	0.957

Effect span: 4.39 pp; paired-sd median (all six): 0.62 pp; span/sd ratio: $7.1\times$
Noise reduction: median $4.7\times$ (mean $5.2\times$), $\bar{\rho} = 0.969$; constant8 median $6.6\times$, $\bar{\rho} = 0.980$
Seed-blocked permutation test: global null rejected at $p < 0.0001$

The five candidates sharing the multiset $\{1, 1, 2, 4, 8, 16, 16, 16\}$ run from `tpe_perm` at $+1.09$ pp to `random64_b` at -3.18 pp, so permuting the same ranks with composition untouched produces almost the entire observed spread. `tpe_w2025`, the TPE search output, sits below the baseline in point estimate (-0.63 pp, CI $[-1.65, +0.10]$) with an interval containing zero and no rejection under Holm (raw $p = 0.2891$, Holm $p = 0.5781$); the interval does exclude the $+1$ to $+3$ pp range typically reported for adaptive allocation, ruling out an effect of that size *in this regime* without settling whether gains reported elsewhere are spurious. It also loses by 1.84 pp to an equal-budget random allocation of a different multiset (`random64_a`: $+1.21$ pp, CI $[+0.96, +1.43]$, Holm $p = 0.0469$).

4.2 Cross-Seed Stability and Effect Ordering

Search run at one seed selected the panel's least reproducible candidate. `tpe_w2025` was chosen by a TPE search at master seed 2025; re-evaluated over eight seeds it carries the largest paired sd of any candidate (1.38 pp against a 0.54 pp median over the other five, $2.6\times$ less stable), and at seed 2025 itself it ranks fifth of six. Single-seed selection did not merely pick a candidate that fails to

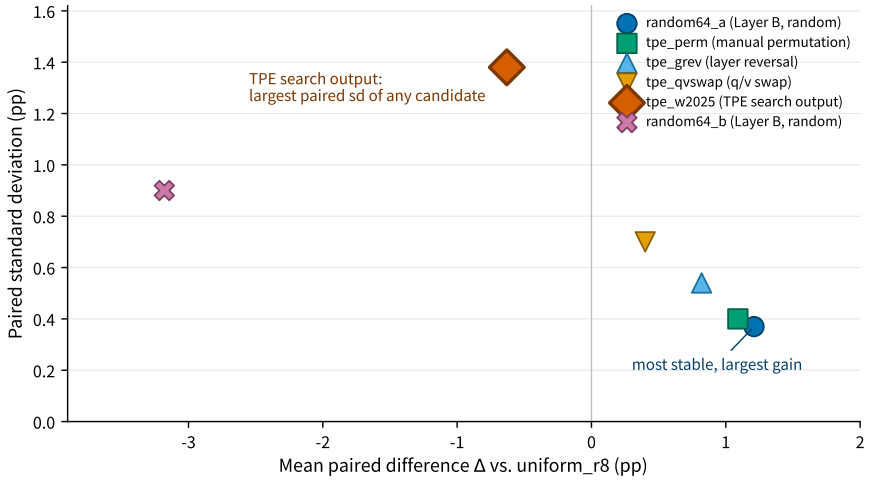

Figure 1: Stability vs. performance (MNLI constant2, 8 seeds, paired against `uniform_r8`). Each point is a candidate. The plot shows that high variance and near-zero effect coincide in the TPE search output (`tpe_w2025`, red diamond), while the two largest positive effects come with the two smallest paired sds and the reliably negative candidate has the largest SNR.

generalize: it picked the one whose effect is least determined by the allocation and most by the draw, which a single-seed objective has no way to see.

Ranking overall is not unstable: cross-seed Kendall τ_b between effect orderings averages $+0.663$ in constant2 (range $[+0.333, +1.000]$, no negative pairs among 28 comparisons), so the search output's failure to win is not a seed-randomness artefact — reproducibility coexists with poor performance. Stability varies sharply by candidate and is unaligned with effect size (Figure 1): `tpe_w2025` combines the largest paired sd with a near-zero effect, the two top performers have the two smallest paired sds, and `random64_b` is the most precisely estimated effect of all (SNR 3.53) even though its paired sd (0.90 pp) sits above the panel median.

4.3 Placement Effects Within Layer A

Layer A candidates (Table 1) share the multiset $\{1, 1, 2, 4, 8, 16, 16, 16\}$ and hence identical budgets, differing only in which gene receives which rank, so the design isolates placement by construction. The searched placement is not the best placement of its own multiset: `tpe_w2025` ranks fourth of five, and three manual permutations of the very ranks it selected do better. Since all five carry the same ranks in the same quantities, no importance-based argument about *which ranks to use* explains the spread — only *where* they sit can.

Placement is a first-order determinant of the measured effect: within one multiset it accounts for 4.27 of the panel's 4.39 pp, and the panel's two largest effects belong to a hand permutation (`tpe_perm`, $+1.09$ pp) and an equal-budget random draw (`random64_a`, $+1.21$ pp), both Holm-significant, while the searched output is not. What fails in this audit is therefore not the premise that placement matters, but the attribution of a single-seed gain to the method that chose the placement.

4.4 Regime-Conditional Effects

Effect magnitudes compress under strong adapter scaling. Because one candidate (`tpe_w2025`) collapses to chance accuracy in the `m777` block under constant8 (Section 4.6), that block is excluded, leaving a balanced 7×7 constant8 panel; the constant2 column keeps all 8 seeds; the collapsed cell itself is reported in Table 8 rather than entered here. Constant8 also shifts the level upward (median 84.83 vs. 81.11 pp over the seven sum-64 candidates, collapses excluded); the compression concerns *between-candidate differences*, and the paired estimator removes any per-seed level shift.

The candidate span shrinks to 1.41 pp from 4.39 pp, a 3.1× compression, and cross-seed Kendall τ_b falls from $+0.663$ to $+0.231$ (range $[-0.390, +0.810]$, 14% of pairs negatively correlated). Every positive effect vanishes: the three candidates beating `uniform_r8` by $+0.82$ to $+1.21$ pp under constant2 land within $[-0.12, +0.11]$ pp. Only `random64_b` retains a nominally significant harmful effect (-1.30 pp, Holm $p = 0.094$); nothing survives Holm correction at $\alpha = 0.05$, since $n = 7$ makes $6 \times 0.0156 = 0.094$ the smallest attainable corrected value. Ordering transfers only partially ($\tau = +0.60$ between regimes over six candidates, exact two-sided $p = 0.136$), so relative performance is not regime-invariant (Figure 4 in Appendix A.5): an ordering established under one scaling policy cannot be assumed under another, and stronger scaling shrinks between-candidate differences rather than moving all candidates uniformly.

4.5 Cross-Task Transfer of Candidate Ordering

If the ordering does not survive a change of scaling policy, does it survive a change of task? Replicating the constant2 panel on QNLI (P0E, 7 candidates × 8 seeds) under the same protocol, the ordering transfers exactly (Table 7): Kendall $\tau = +1.000$, all 15 pairs concordant, giving the identical permutation `random64_a` > `tpe_perm` > `tpe_grev` > `tpe_qvswap` > `tpe_w2025` > `random64_b`. Exactly two of the $6! = 720$ orderings attain $|\tau| = 1$ under a uniform-permutation null (the identity and its reversal), giving exact $p = 2/720 = 0.0028$. Seed-by-seed the cross-task τ averages $+0.649$ (range $[+0.43, +0.87]$, no negative blocks), so the agreement does not come from averaging.

The agreement is recipe-specific, not task-general. Extending to CoLA (Appendix A.1) breaks the perfect transfer: $\tau_{\text{MNLI–CoLA}} = \tau_{\text{QNLI–CoLA}} = +0.467$ (11 concordant of 15 pairs), with `tpe_perm` falling from second to fourth and `tpe_qvswap` from fourth to last. Ordering thus transfers perfectly between two same-type tasks and decays to a weak, non-significant correlation on a task differing in type, metric and class balance, suggesting that what transfers belongs to the shared recipe rather than to allocation quality. Three qualifications bound the MNLI–QNLI result: both tasks are GLUE sentence-pair classification under one recipe, so they are not independent evidence of task generality; magnitudes do not transfer where ordering does, QNLI effects running roughly half their MNLI counterparts with none of the three favourable candidates reaching $\alpha = 0.05$; and the `m42` exclusion is consequential — including the collapsed `tpe_perm` run drops its QNLI mean to -2.88 pp (paired sd 9.78 pp) and moves it from second to last, so comparing raw eight-seed averages recovers the wrong ordering.

Candidate ordering is therefore more robust to related task changes than to changes in scaling policy, but that establishes neither search quality nor task-general superiority—the best-ordered candidate is a random allocation.

4.6 Collapse as a Discrete Failure Mode

Criterion. A run is a *full collapse* when its metric falls at least 25 points below the median of all runs sharing its task and scaling regime, and a *partial degradation* at 8 points without reaching 25. A relative threshold avoids separate absolute cutoffs for a three-class accuracy, a two-class accuracy and an MCC; the reference medians are 80.61 and 84.23 pp (MNLI constant2/constant8, $n = 64$ each), 87.78 pp (QNLI constant2, $n = 56$) and 0.6354 MCC (CoLA constant2, $n = 56$). The rule yields two full collapses (-52.0 and -27.4 points) and no run in the partial band; no CoLA run is abnormal (worst -3.0 MCC points). The first is `tpe_w2025` at master seed 777 on MNLI under constant8; the second is `tpe_perm` at master seed 42 on QNLI under constant2. The QNLI event at accuracy 0.604 sits above the 0.50 majority baseline; what marks it as collapse is its 27.4-point deficit plus the spectral signature of Section 5, shared with the MNLI event and shown by no normal run at end of training.

Two features of Table 8 matter here. Neither event propagates: each occupies one candidate in one seed block, no other candidate collapses in the same block, and both collapsed cells are reported in Table 8 rather than entered into the panels below. And degradation is rare (2 of 301 configurations, 0.7%) yet catastrophic (52.0 and 27.4 pp), so an allocation can look sound on average while harbouring fragility that surfaces only under particular draws. For the affected comparisons the *entire*

seed block is dropped rather than the collapsed cell alone, keeping panels balanced: `m777` leaves the constant8 column of Table 6, `m42` the QNLI column of Table 7.

5 Spectral Diagnostics of Collapse

5.1 Motivation and Method

To characterize the failure mode of Section 4.6 we compute, for each adapter module at training completion, $\|\Delta \mathbf{W}\|_F = \|s \cdot \mathbf{BA}\|_F$, where s is the scaling factor, $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$ the trained down-projection and $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$ the trained up-projection. The 64 modules (32 layers $\times$ 2 projections) are partitioned into depth quartiles, each quartile taking its share of total update energy; rank-dependent metrics (stable rank, spectral entropy) are normalized by the module's rank r so that low rank is not confounded with low structural complexity.

Two collapsed adapters have end-of-training checkpoints: `tpe_w2025`@m777 on MNLI under constant8 (acc 0.322) and `tpe_perm`@m42 on QNLI under constant2 (acc 0.604). Table 12 lists each against its normal siblings — seven master-seed siblings on QNLI, five data-seed probes on MNLI that hold head and adapter initialization bitwise identical and vary only the data stream. No new training is involved: the QNLI checkpoints are reloaded from the P0E panel by protocol-and-seed cache key.

5.2 Positional Signature: Shallow-Layer Concentration

Both collapsed adapters concentrate *update energy in the two shallowest layers* — layers 0–1 hold 26.2% (MNLI) and 18.5% (QNLI) of total update energy against sibling medians of 3.25% and 1.95%, 22 and 17 sibling-range widths above the top of the range — while mid-depth layers receive negligible updates and total magnitude falls to 41% and 68% of the sibling median (Tables 13 and 14). We quote medians because a range from five points would shift with one outlying run.

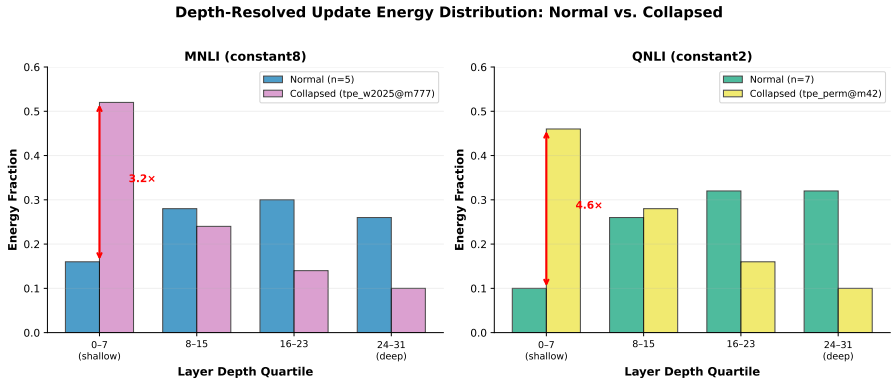

Figure 2: Depth profile of update energy. Per-layer fraction of total update energy $\|\Delta \mathbf{W}\|_F$ for the two collapsed adapters (MNLI `tpe_w2025`@m777 under constant8; QNLI `tpe_perm`@m42 under constant2) against normal runs sharing the same candidate, task and regime (seven master-seed siblings on QNLI; five data-seed probes with identical initialization on MNLI). Collapsed adapters concentrate energy in layers 0–1 (26.2% and 18.5% of the total, versus 3.2% and 2.0% medians normally) while mid-depth layers receive negligible updates.

Shape, not magnitude. The fraction $E_{0,1}/\sum_{\ell=0}^{31} E_\ell$ normalizes out total scale, so any mechanism scaling all updates by a common factor would lower $\|\Delta \mathbf{W}\|_F$ while leaving that profile intact. Instead shallow layers in collapsed runs take *more* absolute energy than in normal ones despite the lower total, and deep layers dramatically less — not global rescaling, but a pattern consistent with updates failing to propagate to depth rather than with a capacity shortfall, though two checkpoints cannot identify the mechanism further.

5.3 Interpretation and Attribution

The signature is qualitatively similar across the two events, which differ in task, candidate, seed and regime, though not in every metric ($\tau = +0.625$ between the two collapsed layer-wise energy profiles, against $+0.811$ between normal runs). It supports a *post-hoc* check: the layer 0–1 energy fraction needs only the adapter weights, and a value above 10% (against 2–3% normally) flags an adapter for inspection — a warning threshold from these experiments, not a predictor, since it identifies adapters that *have* collapsed and one probe crossed it at step 10 without collapsing (Appendix A.9). Cause or consequence cannot be settled from end-of-training snapshots (Appendix A.6).

On attribution we make no causal claim. The five Layer A candidates each carry exactly two $r = 1$ genes, yet only two of them collapse, so shared low-rank structure neither predicts nor explains the events; each event occupies a single seed and a single regime, with no other candidate collapsing in its block; and only two collapsed checkpoints exist, so the design cannot separate seed, placement, regime and task. Rank allocation is at most contributing, neither necessary nor sufficient.

6 Discussion

6.1 Implications for Adaptive Allocation Research

Two assumptions common in this literature fail in the audited setting.

Assumption 1: a search output is a superior allocation. Re-evaluated under seed decoupling, `tpe_w2025` is not reliably better than uniform ranks and loses to an equal-budget random draw.

Assumption 2: allocation effects transfer across scaling policies. Ordering transfers only partially across scaling policies ($\tau = +0.60$, $p = 0.136$); an allocation tuned under one policy carries no guarantee under another.

What this does and does not establish. We do not establish that published gains from learned rank allocations are spurious. We identify a concrete alternative explanation — allocation-versus-initialization confounding under uncontrolled seeding — sufficient to produce uniform-beating-looking results with no genuine allocation advantage. Distinguishing the two requires seed-decoupled re-evaluation, which we contribute.

6.2 Limitations

Scope and what is being audited. Findings are scoped to Llama-2-7B, three GLUE tasks at $N = 5{,}000$, and LoRA on `q_proj`/`v_proj`. We audit the *outputs* of both searches, not the procedures: the audit is in-protocol — same base model, task family, hyperparameters and step count as the search that produced these allocations — so the negative result is no artefact of a changed setting, though how those procedures would behave with seed-decoupled evaluation inside their own loops is untested. Evaluation-set size further bounds resolution, materially on CoLA (Appendix A.4).

Evidence base. Two collapse events among 301 configurations (0.7%) cannot support inferential statistics on collapse predictors, and the signature rests on two end-of-training checkpoints, which cannot say when shallow concentration emerged or whether it preceded degradation; one probe (Appendix A.9) diverges by step 10 without collapsing, leaving early signatures unvalidated.

7 Conclusion

Allocation search outputs do not reliably outperform uniform ranks here; placement alone induces large effects at fixed composition; and transfer depends on both scaling policy and task type. The largest detectable negative effect belongs to an equal-budget random candidate, and on CoLA the only Holm-significant effect is a manual permutation. This does not invalidate adaptive allocation; it raises the evidence standard that robust-transfer claims must meet.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

We used generative AI tools as general-purpose assistants in preparing this submission. Specifically, LLMs were used for (i) writing and editing assistance, including drafting and revising prose, improving clarity, and formatting LaTeX tables and figures; (ii) code assistance for the analysis pipeline (bootstrap resampling, sign tests, Holm correction, permutation tests, and plotting scripts); and (iii) internal-consistency checking of reported numbers across the manuscript. All experimental design decisions, all training runs, and all reported numerical results originate from the authors' own experiments. No synthetic or LLM-generated data were used: every result derives from actual fine-tuning runs on GLUE data. LLMs did not originate the research hypothesis or the statistical claims; all statistical derivations and all numbers appearing in the paper were verified by the authors against the primary experiment logs. The authors take full responsibility for the entire content of this paper, including any text or code produced with AI assistance.

ETHICS STATEMENT

This work is an audit of existing parameter-efficient fine-tuning methods and introduces no new model, dataset, or deployment artifact. All experiments use publicly available resources: the Llama-2-7B checkpoint under its community license and the GLUE benchmark (MNLI, QNLI, CoLA) under its published terms. No human subjects were involved, no personally identifiable information was processed, and no data was collected from users. Our findings are negative with respect to a widely adopted technique; we have taken care to scope the claims to the regimes actually tested (one base model, three classification tasks, $N = 5{,}000$, two constant scaling regimes) and to state explicitly where generalization is unwarranted (Section 6.1). We report the failure modes we observed, including runs that collapsed, rather than filtering them from the analysis. The principal societal consideration is compute cost: the audit consumed roughly 40–60 GPU-hours on A100 hardware, and we release the full run logs so that others need not repeat it.

REPRODUCIBILITY STATEMENT

Our protocol is designed for exact reproduction. Section 3.1 specifies the SHA-256 seed-derivation scheme in full, including the three key strings, so that every run in the paper can be regenerated bit-for-bit from its `master_seed`; the eight master seeds are listed in Section 3.4. Section 3.2 gives the complete gene-to-module mapping, and Table 1 lists all eight candidate chromosomes verbatim. Section 3.4 states the base checkpoint, the sampling procedure and size for each task, every optimizer hyperparameter, the resulting step count, and the exact composition of all 301 unique configurations (305 log records). The inference procedure — bootstrap resample count, sign test, Holm families, and permutation counts — is specified in Section 3.1, and the variance decomposition that motivates the design is reported in Appendix A.7. Code, the full per-run evaluation logs, and the analysis scripts that produce every table and figure in this paper will be released upon acceptance.

REFERENCES

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. Sparse low-rank adaptation of pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4133–4145, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.252.

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`.

Yihang Gao and Vincent Y. F. Tan. Statistical inference for rank allocation in low-rank adaptation. *arXiv preprint arXiv:2607.20205*, 2026.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Sotiris Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*, 2023.

Jiaqi Liu, Haidong Kang, Qihui Zhao, Guo Yu, and Jingchao Wang. RSRA: Training-free probing of representation sensitivity for efficient LoRA rank allocation, 2026. URL https://arxiv.org/abs/2607.09757. arXiv:2607.09757v2.

Zequan Liu, Jiawen Lyn, Wei Zhu, Xing Tian, and Yvette Graham. ALoRA: Allocating low-rank adaptation for fine-tuning large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 622–641, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.35.

Chuyan Zhang, Kefan Wang, and Yun Gu. Beyond low-rank tuning: Model prior-guided rank allocation for effective transfer in low-data and large-gap regimes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2507.00327.

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.

Yi Zhao, Qinghua Yao, Xinyuan Song, and Wei Zhu. Shaplora: Allocation of low-rank adaption on large language models via shapley value inspired importance estimation. In *Second Conference on Parsimony and Learning (CPAL)*, 2025.

A Supplementary Materials

A.1 Independent Verification on a Third Task: CoLA

MNLI and QNLI are both sentence-pair classification tasks scored by accuracy, so their agreement (Section 4.5) may reflect shared recipe structure rather than a task-general result. As an independent check we replicated the constant2 panel on CoLA—a single-sentence acceptability task, scored by MCC because of severe class imbalance (majority class 0.707), drawn from a pool of only 8,551 examples so that our fixed $N = 5{,}000$ covers 58.5% of it. All 56 runs (7 candidates $\times$ 8 seeds) completed without collapse; the lowest observed MCC is 0.605, far above the degradation threshold used in Section 4.6.

Table 3 reports paired differences against `uniform_r8` over the eight seeds. Exactly one candidate reaches Holm-corrected significance, and its effect is negative: `tpe_qvswap`—a *manually constructed* q/v swap of `tpe_w2025`'s rank multiset, not a search output (Table 1)—loses to the uniform baseline on all 8 of 8 seeds (mean $\Delta = -0.0148$ MCC, 95% CI $[-0.0238, -0.0074]$, raw sign $p = 0.0078$, Holm $p = 0.0469$). Its per-seed rank among the seven candidates is 6, 6, 6, 3, 6, 6, 7, 7—near-last in seven of eight blocks, never better than third. No candidate shows a significant *gain*: the largest positive effect (`random64_a`: $+0.0081$ MCC) has Holm $p = 1.0$. The TPE search output `tpe_w2025` is not reliably different from uniform rank-8 here (-0.0008 MCC, 4/8 seeds, Holm $p = 1.0$). The correct reading is therefore: on CoLA the search output is indistinguishable from uniform, whereas one manually constructed permutation of its rank multiset is significantly harmful. This is the third task on which the search output fails to beat uniform ranks, and the first on which any panel candidate is significantly worse.

Holm correction detail. We apply the step-down procedure with the monotonicity enforcement $\tilde{p}_{(i)} = \max\{\tilde{p}_{(i-1)}, \min(1, (m-i+1)\, p_{(i)})\}$. This matters here because the raw multiplier products are 0.0469, 0.3516, 1.1563, 0.8672, 1.4531, 1.0000: the fourth value must be raised to 1.0000 by the third. We further break ties among equal raw p-values in a fixed candidate order for byte-level reproducibility; in this panel `random64_a` and `random64_b` do tie at $74/256$, but after monotonization both map to 1.0000 regardless of tie order, so the reported table is invariant to that choice.

Table 3: CoLA paired-difference statistics (constant2, 8 seeds, vs. `uniform_r8`), all 56 runs completed without collapse. Effects are in MCC units (Matthews correlation coefficient) because of severe class imbalance (majority class 0.707). `tpe_qvswap`, a manually constructed q/v swap of `tpe_w2025`'s multiset rather than a search output, is the single Holm-significant result and is *harmful*, losing on 8/8 seeds; no candidate shows a significant gain, and the search output `tpe_w2025` is indistinguishable from uniform. Effect span 0.0229 MCC is 1.96× the paired-sd median (0.0117 MCC), the weakest span/noise ratio of the three tasks (MNLI 7.08×, QNLI 3.59×) but still detectable. Holm correction is over $m = 6$ with monotonicity enforced; sign p is the two-sided exact binomial test, whose attainable values at $n = 8$ are $\{0.0078, 0.0703, 0.2891, 0.7266, 1.0000\}$.

Candidate	Mean Δ (MCC)	95% CI (MCC)	Paired sd (MCC)	Wins	Sign p	Holm p	SNR
`random64_a`	+0.0081	[+0.0006, +0.0154]	0.0115	6/8	0.2891	1.0000	0.71
`tpe_grev`	+0.0073	[+0.0016, +0.0122]	0.0081	7/8	0.0703	0.3516	0.90
`tpe_w2025`	−0.0008	[−0.0117, +0.0089]	0.0157	4/8	1.0000	1.0000	0.05
`tpe_perm`	−0.0041	[−0.0086, +0.0006]	0.0074	3/8	0.7266	1.0000	0.56
`random64_b`	−0.0102	[−0.0192, −0.0014]	0.0139	2/8	0.2891	1.0000	0.73
`tpe_qvswap`	−0.0148	[−0.0238, −0.0074]	0.0120	0/8	0.0078	**0.0469**	1.24
Effect span: 0.0229 MCC; paired-sd median: 0.0117 MCC; span/sd ratio: 1.96×							
Only one candidate reaches Holm significance (`tpe_qvswap`, *harmful*); no candidate shows a significant *gain* over uniform rank-8							

Interpretation: CoLA provides independent third-task verification of the paper's central claim. `tpe_qvswap`—a manual q/v swap of `tpe_w2025`'s multiset, not a search output—is the only candidate reaching Holm-corrected significance, and its effect is *negative*: it loses to `uniform_r8` on all 8 of 8 seeds (per-seed ranks 6, 6, 6, 3, 6, 6, 7, 7), with mean $\Delta = -0.0148$ MCC and sign-test $p = 2/256 = 0.0078$ (the two-sided lower bound at $n = 8$). Relative to noise the effect span is 1.96×, weaker than MNLI (7.08×) or QNLI (3.59×) but clearly detectable.

Effect strength is attenuated, not absent. A cross-task comparison requires a scale-free metric, since CoLA is scored in MCC and MNLI/QNLI in accuracy percentage points. Expressing each task's paired effect span as a multiple of its paired-sd median gives 7.08× (MNLI: 4.39 pp span, 0.62 pp sd), 3.59× (QNLI: 1.96 pp, 0.55 pp), and 1.96× (CoLA: 0.0229 MCC, 0.0117 MCC). CoLA therefore shows the weakest allocation signal of the three tasks relative to its own noise floor—but a 1.96× span remains detectable, and the strongest effect in the panel is significantly harmful. We deliberately avoid attributing this attenuation to data sufficiency alone: CoLA differs from MNLI/QNLI simultaneously in task type (acceptability judgment vs. inference), evaluation metric (MCC vs. accuracy), and class balance (majority class 0.707), in addition to pool coverage (58.5% vs. 1.3–4.8%). With a single task at high coverage these four factors are fully confounded, and we make no claim that allocation effects disappear at data sufficiency.

What CoLA adds to the audit. The result strengthens rather than qualifies the main finding. On MNLI and QNLI the search output was merely *indistinguishable* from uniform rank-8; on a third task, under the same protocol and a metric chosen for CoLA's class imbalance, it remains indistinguishable while a manually constructed permutation of its own multiset is *significantly worse* than the uniform baseline. Across three tasks and 240 paired panel runs, no search output achieves a Holm-significant gain over uniform ranks, and at $\alpha = 0.05$ the Holm-significant losses are the manual permutation on CoLA and the random equal-budget candidate `random64_b` on MNLI; on QNLI the latter reaches only the $\alpha = 0.10$ level (0.094, the smallest value attainable at $n = 7$).

A.2 Search-Space Granularity and the Role of the Surrogate

Why eight genes. Per-layer allocation over 32 layers × 2 projections with five admissible ranks spans $5^{64} \approx 10^{44}$ points, far beyond what real-hardware evaluation can cover at any budget we could spend; quartile-level genes reduce this to 5^8. The granularity is also the one the signal lives at: LoRA's own ablation finds $\mathbf{W}_q + \mathbf{W}_v$ the best projection pair (Hu et al., 2021), and depth-segment rather than single-layer structure is what the allocation literature's importance profiles resolve, so eight genes = four depth segments × two sensitive projections is the smallest expression that keeps both. What the coarse genes cost is resolution, not expressiveness: the audit's own Layer A places five permutations of a single multiset 4.27 pp apart at fixed composition (Section 4.3), so quartile-level genes demonstrably move the metric. Whether a finer parameterization would find more is a

question this design cannot answer, and any such gain would have to clear the 0.4–1.4 pp paired sd of this panel to be detectable at all.

The surrogate screens, it does not extrapolate. In SAGA-RA the random-forest / Gaussian-process surrogate ranks a GA population and forwards only the top-k chromosomes to real training; every surrogate prediction is overwritten by the next generation's measured accuracy, and no surrogate output enters any number reported in this paper. It is a screening device inside the search loop, not an outer-loop model used to extrapolate over unevaluated regions — which is also why coarse genes cost little: the surrogate never has to be accurate away from the points that were actually trained.

A.3 Run Inventory and Log-Record Accounting

Training uses 2 epochs at per-device batch size 8 with gradient accumulation 2. Because the trainer forms optimizer steps by floor-dividing the micro-batches per epoch by the accumulation factor, this gives $\lfloor \lceil 5000/8 \rceil / 2 \rfloor \times 2 = \lfloor 625/2 \rfloor \times 2 = 624$ optimizer steps, identically for all three tasks.

The main MNLI panel (P0B) is 8 candidates $\times$ 8 master seeds $\times$ 2 scaling regimes = 128 cells. Two of those cells—`uniform_r8` and `tpe_w2025` at master seed 2025, in both regimes, so four cells in total—were executed in the pilot phase (P0A) under an identical protocol hash and are reused rather than re-run; we verified that their protocol digests, seed derivations and hyperparameters match the P0B configuration exactly, so the 128-cell analysis panel costs 124 new runs. For paired inference (Table 2) we drop `sagara_w2025` as a capacity confound, leaving $7 \times 8 \times 2 = 112$ cells. QNLI (P0E) and CoLA (P0G) each add 7 candidates $\times$ 8 seeds under constant2 = 56 runs. Spectral analysis trains nothing additional: it reloads 8 QNLI panel checkpoints and the 6 unique MNLI data-seed probes. Supporting work comprises 30 variance-decomposition probes (Appendix A.7), 20 native-regime runs, 9 pilot runs and 6 unique data-seed probes, i.e. 65 unique supporting configurations, so $124 + 56 + 56 + 65 = 301$; Table 4 lists the groups.

A note on the released logs. The `master_seed` field is null on 142 of the 305 log records (the P0D, P0E, P0G groups), because those runners recorded the derived seeds without echoing the master value. The master seed is recoverable for every such record — from the `_m<seed>` token in the run tag, and independently from the `(seed_head, seed_lora, seed_data)` triplet, which the derivation scheme maps one-to-one from m — and we verified the two routes agree on all affected records before using them. The released logs will carry the field populated.

One data-seed probe is itself a collapse whose evaluation was written to the log five times. Those four redundant re-evaluations are the entire difference between the 305 evaluation records in the log and the 301 unique configurations behind them, and every per-run statistic in this paper is computed on the unique configurations. Each run takes 8–12 GPU-minutes on an A100.

A.4 Per-Candidate Noise Reduction in Both Regimes

Evaluation-set size. Evaluation sets differ in size: MNLI (9,815) and QNLI (5,463) give binomial standard errors of 0.40 and 0.44 pp, an order of magnitude below the paired seed-to-seed variance of those panels (raw sd 2.4–3.4 pp), so finite-sample noise is not what limits them. CoLA (1,043 examples, MCC-scored) is the exception: its sampling variability is close to the ± 0.010–0.015 MCC scale of the effects themselves, which is why the CoLA conclusions are stated as ordinal agreement plus one significant loss rather than as calibrated magnitudes.

Beyond LoRA. The protocol should transfer to any fine-tuning comparison where a shared initialization source dominates run-to-run variance; how much is gained depends on how much of that variance is genuinely common-mode, which must be measured rather than assumed — the decomposition below is how we measured it here.

Table 5 gives the raw and paired standard deviations behind the reduction and ρ columns of Table 2, for constant2 and constant8 alike. The constant8 mean is inflated by one near-degenerate entry: `tpe_grev` has a paired sd of 0.17 pp over 7 seeds ($\rho = 0.999$), and excluding it lowers the constant8 mean from $7.16\times$ to $5.46\times$. Figure 3 plots the same quantities. We therefore use medians ($4.7\times$

Table 4: Run inventory, in log-record units. The MNLI panel contributes 124 newly trained cells and reuses 4 pilot cells (`uniform_r8` and `tpe_w2025` at master seed 2025, both regimes) under an identical protocol hash, so the 128-cell analysis panel of Section 4.1 costs 124 new runs. Spectral analysis (Section 5) trains nothing additional: it reloads 8 QNLI panel checkpoints and the 6 unique MNLI data-seed probes. The log contains 305 evaluation records over 301 unique configurations; the difference is four redundant re-evaluations of the single collapsed data-seed probe (10 records, 6 unique configurations), and all reported statistics use the unique configurations.

Group	Purpose	Records	Section
MNLI main panel	8 candidates × 8 seeds × 2 regimes, newly trained	124	4.1
QNLI panel	7 candidates × 8 seeds, constant2	56	4.5
CoLA panel	7 candidates × 8 seeds, constant2	56	A.1
Variance probes	seed-source decomposition	30	A.7
Native regime	$s = \alpha/r$ exploration	20	3.3
Pilot	protocol validation (4 cells reused above)	9	—
Data-seed probes	data-seed sweep, spectral analysis (6 unique)	10	5
Total records		**305**	
Unique configurations		**301**	

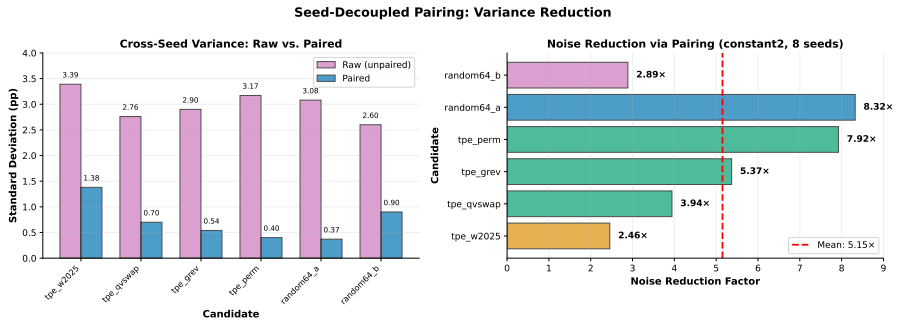

Figure 3: Noise reduction via seed-decoupled pairing. Left: raw vs. paired standard deviations. Right: per-candidate reduction factors, $2.5\times$ to $8.3\times$ in constant2 (median $4.7\times$).

constant2, $6.6\times$ constant8) as the summary statistic throughout the paper, and the honest statement is that constant8 pairing is at least as effective as constant2 pairing rather than markedly better.

Since the standard error of a mean scales as $n^{-1/2}$, these median reductions correspond to $22\times$ and $43\times$ reductions in variance, so matching the power of our 8-seed paired design without pairing would take roughly 173–344 unpaired seeds per candidate. This is a variance-equivalence calculation under the measured common-mode assumption, not an empirical power analysis: it presumes the shared component is exactly common-mode, which $\rho \geq 0.914$ supports but does not guarantee at smaller effect sizes.

A.5 Relocated Main-Text Exhibits

The exhibits below support results stated in full in the main text and are placed here for space.

Table 5: Noise reduction achieved by the seed-decoupled protocol. Pairing removes the shared score-head component, which accounts for 94.1% of run-to-run variance in the constant2 panel (Appendix A.8), reducing paired standard deviations by $2.5-8.3\times$ in constant2 (median $4.7\times$, mean $5.2\times$) and $3.1-15.6\times$ in constant8 (median $6.6\times$, mean $7.2\times$, the mean inflated by the single near-degenerate entry marked †). Cross-candidate correlation ρ averages 0.969 (constant2) and 0.980 (constant8), confirming that head noise appears as a near-identical offset across candidates within each seed block.

Candidate	**Raw sd (pp)**	**Paired sd (pp)**	**Reduction**	**Correlation ρ**
		Constant2 regime		
`random64_a`	3.08	0.37	$8.31\times$	0.993
`tpe_perm`	3.17	0.40	$7.98\times$	0.993
`tpe_grev`	2.90	0.54	$5.35\times$	0.984
`tpe_qvswap`	2.76	0.70	$3.96\times$	0.974
`tpe_w2025`	3.39	1.38	$2.45\times$	0.914
`random64_b`	2.60	0.90	$2.88\times$	0.957
Constant2, mean / median:			$5.16\times$ / $4.65\times$	$\bar{\rho} = 0.969$
		Constant8 regime (7 seeds; `m777` block excluded)		
`random64_a`	2.64	0.35	$7.58\times$	0.993
`tpe_perm`	2.86	0.52	$5.54\times$	0.992
`tpe_grev`	2.59	0.17	$15.62\times^{\dagger}$	0.999
`tpe_qvswap`	2.66	0.87	$3.05\times$	0.945
`tpe_w2025`	2.77	0.79	$3.50\times$	0.961
`random64_b`	2.44	0.32	$7.65\times$	0.992
Constant8, mean / median:			$7.16\times$ / $6.56\times$	$\bar{\rho} = 0.980$

† `tpe_grev` under constant8 has a near-degenerate paired sd (0.17 pp over 7 seeds, $\rho = 0.999$), which inflates its ratio; we report it for completeness but use the median as the summary statistic for that regime. All raw and paired standard deviations are sample estimates with $n - 1$ denominators, computed on the panels described in Table 2 and Table 6.

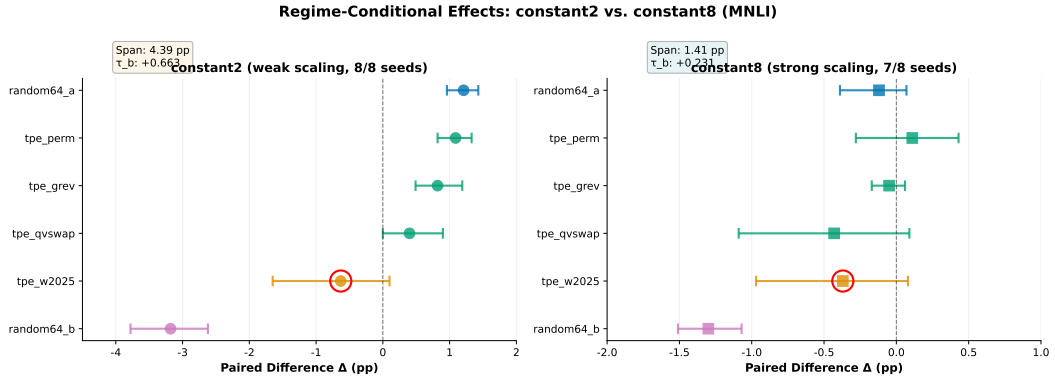

Figure 4: Regime-conditional effects: constant2 vs. constant8 (MNLI main panel). Between-candidate effect span compresses $3.1\times$ under strong scaling; the TPE search output (red outline) shifts position but remains near zero in both regimes.

Table 6: Regime-conditional effects on MNLI: mean paired differences against `uniform_r8` under constant2 (8 seeds) and constant8 (7 seeds, excluding the collapsed `m777` block; Section 4.6). Scaling policy, seed inclusion and collapse exclusion all differ from Table 2, so the two are not stackable. Two-sided exact sign tests, Holm over $m = 6$; at $n = 7$ the attainable raw values are $\{0.0156, 0.1250, 0.4531, 1.0000\}$, making 0.094 the smallest reachable corrected value.

	Constant2			Constant8		
Candidate	**Mean Δ (pp)**	**95% CI (pp)**	**Holm p**	**Mean Δ (pp)**	**95% CI (pp)**	**Holm p**
`random64_a`	$+1.21$	$[+0.96, +1.43]$	**0.047**	-0.12	$[-0.39, +0.07]$	1.000
`tpe_perm`	$+1.09$	$[+0.82, +1.33]$	**0.047**	$+0.11$	$[-0.27, +0.43]$	1.000
`tpe_grev`	$+0.82$	$[+0.49, +1.19]$	**0.047**	-0.05	$[-0.16, +0.06]$	1.000
`tpe_qvswap`	$+0.40$	$[+0.00, +0.90]$	0.727	-0.43	$[-1.09, +0.09]$	1.000
`tpe_w2025`	-0.63	$[-1.65, +0.10]$	0.578	-0.37	$[-0.98, +0.08]$	1.000
`random64_b`	-3.18	$[-3.78, -2.62]$	**0.047**	-1.30	$[-1.50, -1.07]$	0.094

Effect span: 4.39 pp (constant2) vs. 1.41 pp (constant8), compression ratio $3.1\times$
Cross-seed Kendall τ_b: mean $+0.663$ (constant2) vs. $+0.231$ (constant8)
Between-regime rank correlation: $\tau = +0.60$ over 6 candidates (exact two-sided $p = 0.136$, not significant)

Table 7: Cross-task transfer of candidate ordering: MNLI (constant2, 8 seeds) vs. QNLI (constant2, 7 seeds after excluding the `m42` block, in which `tpe_perm` collapsed; Section 4.6). Ranks agree on all six candidates (Kendall $\tau = +1.000$, exact $p = 0.0028$), while effect magnitudes shrink by roughly half. Holm correction is over $m = 6$ within each task.

	MNLI			QNLI		
Candidate	**Mean Δ (pp)**	**Rank**	**Holm p**	**Mean Δ (pp)**	**Rank**	**Holm p**
`random64_a`	$+1.21$	1	**0.047**	$+0.61$	1	0.625
`tpe_perm`	$+1.09$	2	**0.047**	$+0.58$	2	0.625
`tpe_grev`	$+0.82$	3	**0.047**	$+0.41$	3	1.000
`tpe_qvswap`	$+0.40$	4	0.727	$+0.01$	4	1.000
`tpe_w2025`	-0.63	5	0.578	-0.19	5	1.000
`random64_b`	-3.18	6	**0.047**	-1.35	6	0.094
Effect span:		4.39 pp		1.96 pp		
`uniform_r8` accuracy:		80.86 pp		87.90 pp		

Between-task Kendall $\tau = +1.000$ over 6 candidates (15/15 concordant), exact $p = 0.0028$

Table 8: All degraded runs: 2 of the 301 unique configurations (0.7%), both full collapses. No run of an analyzed candidate falls in the 8–25 point partial band. "Deviation" is measured against the regime median of Section 4.6. The two events are single (candidate, seed) cells that do not propagate: neither shares its seed block with another degradation, and neither candidate is degraded at any other seed.

Task	**Candidate**	**Seed**	**Regime**	**Accuracy**	**Loss**	**Deviation (pp)**
MNLI	`tpe_w2025`	777	constant8	0.322	1.14	-52.0
QNLI	`tpe_perm`	42	constant2	0.604	0.656	-27.4

Reference medians (Section 4.6): MNLI 84.23 pp constant8, 80.61 pp constant2;
QNLI 87.78 pp constant2; CoLA 0.6354 MCC
Chance levels: MNLI $1/3 \approx 0.333$ with $\ln 3 = 1.099$; QNLI 0.50 with $\ln 2 = 0.693$
Full-collapse rate: $2/301 = 0.7\%$ of unique configurations

Note: The MNLI event falls below chance accuracy with a loss slightly *above* $\ln 3$; the QNLI event sits just above the 0.50 majority baseline and is classified as a collapse by its 27.4-point deficit together with the spectral signature of Section 5, which it shares with the MNLI event (Table 13). Comparison by allocation type is descriptive only: before the `m777` exclusion the MNLI panel holds 1 degradation among its $7 \times 8 \times 2 = 112$ heterogeneous cells and 0 among its $1 \times 8 \times 2 = 16$ `uniform_r8` cells (Fisher exact, two-sided $p = 1.000$). With two events we draw no conclusion about whether heterogeneous allocations are intrinsically more fragile, and the two checkpoints cannot separate seed, placement, regime and task (Section 5).

A.6 Rank, Projection and Spectral Structure of Collapsed Adapters

These measurements support the attribution argument of Section 5 and rest on the same two collapsed checkpoints, so they are reported as descriptive evidence rather than tested claims.

The deficit is not localized to a rank group or a projection type. The ratio of collapsed to normal median $\|\Delta\mathbf{W}\|_F$ is roughly uniform across rank groups (MNLI 0.31–0.54 over $r \in \{1, 2, 4, 8, 16\}$; QNLI 0.64–0.89) and across projection types (MNLI q_proj 0.36, v_proj 0.40; QNLI 0.76 for both). These are medians of per-module ratios and need not match the whole-adapter ratios of Section 5.2 (0.41 MNLI, 0.68 QNLI), which aggregate $\|\Delta\mathbf{W}\|_F$ and are pulled down by the high-rank deep modules carrying most of the norm. Both views agree that no single rank group or projection type carries the deficit, which is consistent with a depth-wise phenomenon—the chromosome sets per-layer capacity while collapse changes which capacities are used—but with two events we do not treat it as proof that collapse is independent of rank.

The updates become more isotropic, not less. Normalized stable rank and normalized spectral entropy are *higher* in collapsed adapters than in normal runs (MNLI 0.339 vs. 0.285 median; QNLI 0.301 vs. 0.295). This tells against the simple "rank collapse" reading in which singular values concentrate on a single direction, and looks more like an early-stage or shallow-dominated optimization state than convergence to a degenerate low-rank solution.

Causality is bidirectional on the available evidence. Shallow-layer concentration could be the *cause* of collapse (deep layers fail to participate, representational capacity is insufficient, the model falls back to constant output) or the *consequence* (the model converges to constant output, deep-layer gradients vanish, only shallow layers continue updating). Separating these requires intermediate checkpoints, which do not exist for the collapsed runs; end-of-training snapshots also cannot say whether divergence was gradual or abrupt, or whether adaptive learning-rate adjustment or layer-wise gradient clipping would prevent it.

A.7 Variance Decomposition of Randomness Sources

To motivate the seed-decoupled protocol (Section 3.1), we quantify how much cross-run variance each randomness source contributes. We fix a reference configuration—`uniform_r8` under constant2 on MNLI at the single master seed 2025, whose accuracy is 84.371 pp—and vary *one* source at a time over ten levels, holding the other two at their reference values. This yields 30 probe runs (P0D). The reference value is a single-seed measurement and is therefore not comparable to the 80.86 pp reported for `uniform_r8` in Table 7, which is a mean over the eight master seeds; seed 2025 happens to be the most favourable of the eight (the per-seed range is 76.59–84.37 pp). We use accuracies at a common master seed only within a probe sweep, never across the probe and the panel.

Table 9: Single-source variance decomposition on MNLI (10 levels per source, 30 runs total). Score-head initialization accounts for 92.2% of the total variance across the three sources; adapter initialization and data ordering together contribute 7.8%. Because the head seed is shared across candidates within a master seed, this dominant component is common-mode noise that paired differences cancel algebraically.

Source	Mean (pp)	sd (pp)	Variance (pp^2)	Share
Score-head initialization	81.70	2.099	4.404	$\mathbf{92.2}\%$
Data order / dropout	83.67	0.518	0.268	5.6%
LoRA initialization	84.44	0.321	0.103	2.2%

Score-head initialization dominates: its variance (4.404 pp^2) exceeds the combined variance of the other two sources (0.371 pp^2) by more than an order of magnitude. This is the empirical basis for pairing: since all candidates within a master seed share the same head seed, the head-induced offset appears as a near-identical additive term and cancels in $\Delta(a, m)$.

The probe design above is deliberately conservative—it sweeps ten levels of each source around a single reference configuration—and it therefore recovers the *share* of variance attributable to each

source more reliably than its absolute magnitude. Because the ten head-seed levels are not the eight master seeds used in the main panel, the probe's absolute head variance (4.404 pp^2) is not directly comparable to the main panel's cross-seed spread. We therefore re-estimate the components directly on the panel.

A.8 Panel Variance Components

We fit a two-way random-effects decomposition to the main MNLI panel, treating master seed and candidate as crossed random factors,

$$y_{cm} = \mu + \alpha_c + \beta_m + \varepsilon_{cm}, \qquad \alpha_c \sim \mathcal{N}(0, \sigma^2_{\text{cand}}), \quad \beta_m \sim \mathcal{N}(0, \sigma^2_{\text{seed}}), \quad \varepsilon_{cm} \sim \mathcal{N}(0, \sigma^2_{\text{resid}}), \tag{5}$$

where y_{cm} is the accuracy of candidate c under master seed m. With one observation per cell, σ^2_{resid} absorbs both the candidate-by-seed interaction and any residual run-to-run noise; the decomposition is therefore a conservative attribution, since any genuine interaction inflates the idiosyncratic term rather than the shared one. We solve the ANOVA moment equations $\hat{\sigma}^2_{\text{seed}} = (\text{MS}_{\text{seed}} - \text{MS}_{\text{resid}})/I$ and $\hat{\sigma}^2_{\text{cand}} = (\text{MS}_{\text{cand}} - \text{MS}_{\text{resid}})/J$ over all $I = 7$ sum-64 candidates and J seeds, truncating at zero. Because this design is balanced, the ANOVA estimator coincides with REML.

Table 10: Panel variance components for the MNLI main panel, by scaling regime. The seed component is the part of run-to-run variance that is shared across candidates within a seed block and therefore cancels in paired differences; the residual component is what survives pairing. Shares are computed over the run-to-run noise terms ($\sigma^2_{\text{seed}} + \sigma^2_{\text{resid}}$), excluding the candidate effect, which is signal rather than noise.

Regime	$\hat{\sigma}^2_{\text{seed}}$ (pp^2)	$\hat{\sigma}^2_{\text{cand}}$ (pp^2)	$\hat{\sigma}^2_{\text{resid}}$ (pp^2)	**Shared share**
constant2 (7×8)	8.425	2.262	0.526	$\mathbf{94.1}\%$
constant8 (7×7)[‡]	6.735	0.196	0.223	$\mathbf{96.8}\%$

[‡] The `m777` seed block is excluded under constant8 because two candidates collapse to chance accuracy there (Section 4.6); including it would attribute the collapse to seed variance. Note that $\hat{\sigma}^2_{\text{cand}}$ falls by an order of magnitude from constant2 to constant8, the variance-components counterpart of the $3.1\times$ effect-span compression in Table 6.

The panel estimate puts the shared component at 94.1% of run-to-run variance under constant2 and 96.8% under constant8—close to the probe's 92.2% in share, but roughly twice as large in absolute terms (8.425 vs. 4.404 pp^2). The panel figure is the one consistent with the rest of the paper: $\hat{\sigma}^2_{\text{seed}} + \hat{\sigma}^2_{\text{resid}} = 8.95$ pp^2 implies a raw per-run sd of 2.99 pp, matching the 2.60–3.39 pp raw standard deviations in Table 5, and the implied correlation $8.425/8.951 = 0.941$ sits inside the measured range $\rho \in [0.914, 0.993]$. We therefore quote 94.1% throughout the paper and retain the probe (Table 9) for what only it can establish: that the shared component is specifically the *score head*, not data ordering (5.6%) or adapter initialization (2.2%), which the panel decomposition cannot separate because all three vary together across master seeds.

One caveat bounds this estimate. The additive model predicts a uniform paired-sd reduction of $\sqrt{(\hat{\sigma}^2_{\text{seed}} + \hat{\sigma}^2_{\text{resid}})/2\hat{\sigma}^2_{\text{resid}}} = 2.9\times$ under constant2, whereas the measured per-candidate reductions range from $2.5\times$ to $8.3\times$. The spread is expected—$\hat{\sigma}^2_{\text{resid}}$ is pooled across candidates while each candidate's paired sd is estimated from 8 differences—but its magnitude indicates that the idiosyncratic term is genuinely heteroscedastic across candidates: `tpe_w2025` (paired sd 1.38 pp) is an order of magnitude noisier than `random64_a` (paired sd 0.37 pp). The variance components should thus be read as a panel average, not as a per-candidate noise model.

A.9 Trajectory Analysis (Preliminary)

One exploratory probe tracked a partially degraded run with dense checkpoint sampling to test whether the spectral signature emerges before accuracy degradation. The run exhibited shallow-layer divergence at step 10 (1.6% of training) but did not reach full collapse (loss remained well below $\ln K$), suggesting the signature may precede failure with false positives.

Due to the single-trajectory design and incomplete collapse, we report this as a proof-of-concept only. Validating the signature as an early-stopping criterion requires a controlled study with multiple collapse trajectories and dense sampling, which remains future work.

A.10 Native-Regime Corpus: Per-Method Results

Table 11: Allocation search under the native regime ($s = \alpha/r$, coupled seeds). *Accuracy* is the mean over the independent training runs in the native corpus. Baselines: `uniform_r8` (the field default) and `uniform_r4`. Differences are in percentage points (pp).

Method	Accuracy	vs `uniform_r8` (pp)	vs `uniform_r4` (pp)
GA	0.801766	+0.98	+0.36
TPE	0.802921	+1.10	+0.48
SAGA-orig	0.802615	+1.07	+0.45
SAGA-RA	0.799932	+0.80	+0.18
`uniform_r8` (default)	0.791951	—	−0.62
`uniform_r4`	0.798166	+0.62	—

A.11 Spectral Tables

Tables 12–14 give the per-run spectral summary, the collapsed-versus-sibling comparison, and the structure of the per-module deficit referred to in Section 5. Panel A of Table 12 is a data-seed probe: head and adapter initialization are bitwise identical across its six runs, so it isolates the data stream, whereas Panel B varies the master seed and therefore all three streams together. The two reference groups are consequently not the same quantity, and ranges in Table 13 are read within a panel.

Table 12: Spectral summary of every checkpoint used in the diagnostics. Panel A: the six MNLI `constant8` runs of `tpe_w2025`, one per data-seed probe; head and adapter initialization are bitwise identical across the panel, so only the data stream varies. Panel B: the eight QNLI `constant2` runs of `tpe_perm`, one per master seed. $\|\Delta\mathbf{W}\|_F$ is in scaled units at end of training; the layer 0–1 share is of total update energy; stable rank and spectral entropy are per-run medians over the 64 modules, normalized by module rank. No module receives zero update in any run. Collapsed runs are marked †.

Panel	**Run**	**acc**	$\|\Delta\mathbf{W}\|_F$	**layers 0–1**	**stable rank**	**entropy**
A: MNLI, `constant8`	`ds1523753046`†	0.322	5.75	26.2%	0.339	0.807
	`ds133462716`	0.845	14.53	2.9%	0.285	0.691
	`ds256552569`	0.849	13.92	3.2%	0.269	0.643
	`ds461321679`	0.832	13.86	2.9%	0.302	0.662
	`ds64122173`	0.844	13.85	3.9%	0.279	0.746
	`ds926333634`	0.846	14.23	3.4%	0.303	0.631
B: QNLI, `constant2`	`m42`†	0.604	3.12	18.5%	0.301	0.632
	`m11`	0.890	4.97	2.0%	0.295	0.566
	`m2025`	0.880	5.06	1.8%	0.255	0.592
	`m22`	0.881	4.94	1.8%	0.258	0.648
	`m33`	0.886	4.44	2.1%	0.301	0.625
	`m44`	0.882	4.61	1.7%	0.316	0.655
	`m55`	0.900	4.27	2.6%	0.319	0.563
	`m777`	0.877	4.27	2.3%	0.287	0.656

Table 13: Collapsed runs against their normal siblings. "Siblings" are the 5 other probes of Panel A and the 7 other runs of Panel B in Table 12; the bracketed figure is the sibling range, and the parenthetical one gives how far outside that range the collapsed value sits, in units of the range's width ("—" if inside). Normalized stable rank and entropy are scaling-invariant; $\|\Delta\mathbf{W}\|_F$ is not and is read within a panel only.

Metric	MNLI collapsed	MNLI siblings	QNLI collapsed	QNLI siblings
$\|\Delta\mathbf{W}\|_F$ (total)	5.75 ($12.0\times$ low)	13.92 $[13.85, 14.53]$	3.12 ($1.4\times$ low)	4.61 $[4.27, 5.06]$
layers 0–1 share	26.2% ($22.1\times$ high)	3.25% $[2.9, 3.9]\%$	18.5% ($17.4\times$ high)	1.95% $[1.7, 2.6]\%$
stable rank (median)	0.339 ($1.1\times$ high)	0.285 $[0.269, 0.303]$	0.301 (—)	0.295 $[0.255, 0.319]$
stable rank (min module)	0.066 (—)	0.067 $[0.065, 0.071]$	0.064 ($1.0\times$ low)	0.069 $[0.067, 0.071]$
spectral entropy (median)	0.807 ($0.5\times$ high)	0.662 $[0.631, 0.746]$	0.632 (—)	0.625 $[0.563, 0.656]$
score head $\|W\|_F$	2.208 (—)	2.207 $[2.207, 2.208]$	1.816 (—)	1.807 $[1.78, 1.82]$
modules with zero update	0	0	0	0

Table 14: Structure of the per-module magnitude deficit, as the ratio of the collapsed module to the sibling median of the same module. The noise floor is the smallest pairwise ratio observed among sibling runs, so a module counts only if its deficit exceeds the variation siblings show among themselves. The spread of each grouping is the range of its subgroup medians.

Deficit summary	MNLI	QNLI
modules beyond own noise floor	$60/64$ (94%)	$29/64$ (45%)
per-module ratio, median	0.393	0.756
per-module ratio, range	$[0.194, 1.272]$	$[0.414, 2.156]$
coefficient of variation	0.477	0.378
modules with ratio < 0.15	0	0
by rank r=1/2/4/8/16	$0.41/0.40/0.31/0.54/0.38$	$0.89/0.66/0.86/0.76/0.64$
spread	0.227	0.249
by depth 0–7/8–15/16–23/24–31	$0.40/0.31/0.39/0.48$	$0.76/0.60/0.76/0.89$
spread	0.166	0.289
by projection (q/v)	$0.36/0.40$	$0.76/0.76$
spread	0.042	0.002