RETHINKING LAYER SELECTION FOR LOW-RANK ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Layer selection concentrates a low-rank adaptation budget on layers judged important by a pretrained model. We show that internal coordinates can change these recommendations while leaving the available adaptation functions unchanged. Our *capacity-preserving audit* preserves both the network function and the set of LoRA functions at every fixed placement. Gradient-based selections change across the tested architectures and calibration tasks, whereas contraction and block-ablation controls retain their predicted invariances. Controlled calibration experiments and norm balancing connect this sensitivity to internal scale. Training every selected set on the original backbone then measures its adaptation value: random-placement controls improve on the canonical gradient selection, and fixed sets exhibit different utility across tasks. Finally, comparisons with static placements, AdaLoRA, and FoRA match final effective adapter parameters, training steps, and learning-rate search. Full-depth, uniform low-rank LoRA achieves the highest accuracy in these comparisons, including against adaptive rank allocation that retains every layer. These results connect score interpretation to practical allocation: test coordinate sensitivity, measure the utility of the selected layers, and compare against distributed adaptation under a common parameter budget.

1 INTRODUCTION

Low-rank adaptation (LoRA) trains small adapters within a frozen network (Hu et al., 2022). A fixed parameter budget can support low-rank adapters across many layers or higher-rank adapters in a selected few. Layer-selection methods use gradients, sensitivity, or representation statistics to identify promising locations (Saket, 2026; Zhou et al., 2025; Xu et al., 2026); adaptive methods also redistribute rank during training (Zhang et al., 2023). These approaches raise a common question: which properties of a pretrained model make one allocation preferable to another?

A layer's score is often interpreted as its importance, but the numerical value can depend on the coordinates used to represent its weights. An invertible internal transformation can rescale weights and gradients while leaving the network function unchanged (Dinh et al., 2017; Neyshabur et al., 2015). This distinction matters for layer selection because a ranking determines where the entire adaptation budget is placed. Equivalent representations provide a controlled way to test what a score measures, before asking whether its preferred layers are useful for training.

Can equivalent models recommend different LoRA layers, and what do those recommendations gain over distributed adaptation?

Our **capacity-preserving audit** changes internal coordinates while holding the task, calibration examples, and every block's input-output map fixed. It pairs changes to attention values with inverse changes to output projections, and applies compensated transformations inside SwiGLU. Beyond preserving the pretrained function, these transformations preserve the set of bounded-rank LoRA functions at every fixed placement (Theorem 2). The audit therefore compares scores under a common adaptation-capacity reference.

The audit and training study answer complementary questions. A gradient score measures sensitivity in the optimizer's coordinates and may be useful in those coordinates. Invariance tests whether its interpretation also survives an equivalent representation. Preserving each fixed-placement adaptation

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.

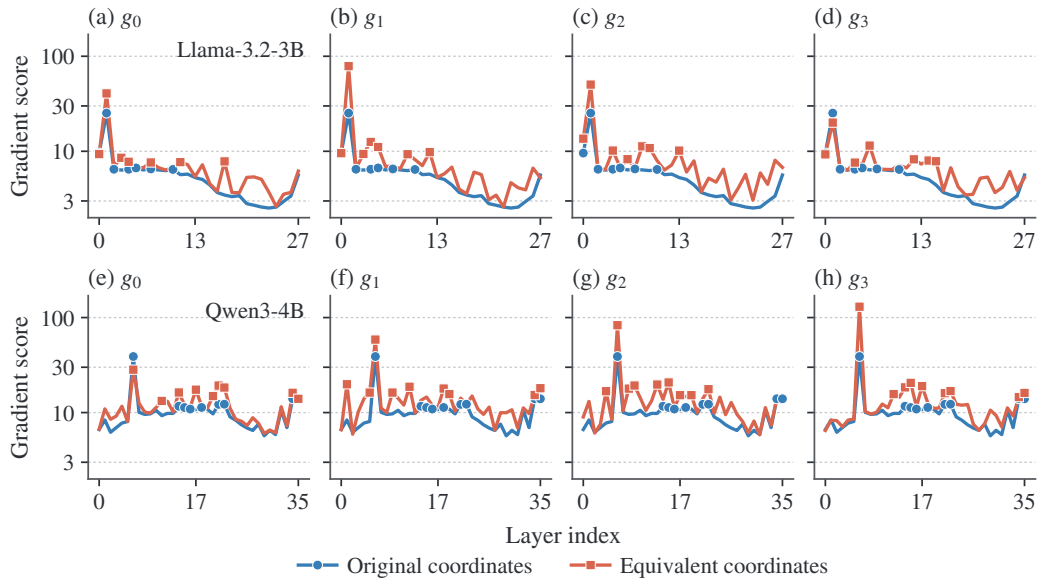


Figure 1: **Equivalent representations change Gradient profiles and selected layers.** Gradient profiles under four exact gauges (g_0–g_3, seeds 18000–18003; bound $b = 2$), using the same 32 calibration examples. Markers identify the top-7 layers for Llama and top-9 for Qwen. Table 2 reports selection overlap J and paired downstream accuracy changes Δ.

set makes this test specific to LoRA: the same locations retain the same attainable deterministic functions, even when a selector recommends a different subset. Evaluating the recommendations in common training coordinates then measures the value of the locations themselves.

The resulting score changes have concrete consequences. Gradient-selected sets change across all six tested model–task configurations. Transformation laws predict which statistics move: Euclidean gradient norms depend on non-orthogonal coordinates, while matrixwise weight–gradient contractions remain invariant. Block ablation supplies a complementary control because each block map is preserved. The certified numerical audit agrees with these predictions, and Figure 1 shows how different representations recommend different layers.

To measure utility, we select layers on the equivalent checkpoints and train every selected set on the *original* backbone. This holds optimizer coordinates fixed. The resulting accuracy profiles differ by model and downstream task. Depth- and overlap-controlled random placements improve on the canonical Gradient selection in the Llama comparison. Balancing internal scales reduces gradient-score dispersion and recovers the evenly spaced reference set. Together with a crossed calibration study, these controls connect the coordinate dependence of a score to the locations it recommends and their training value.

The final comparison places these diagnostics in a practical context. We evaluate static placements, AdaLoRA, and FoRA with equal final effective adapter parameters, a common training schedule, and equal validation search. Full-depth LoRA achieves the highest accuracy, including against adaptive allocation that preserves full layer coverage. Distributed, uniform low-rank adaptation thus provides a strong reference for both layer selection and rank redistribution.

Our contributions are as follows:

1. A **capacity-preserving audit** separates coordinate-dependent scores from function-based interpretations while retaining every fixed-placement LoRA adaptation set.
2. Fixed-backbone training, random-placement controls, and coordinate balancing connect changed scores to the utility and geometry of the selected layers.
3. Matched-budget comparisons with static placements and adaptive algorithms establish the strength of distributed, uniform low-rank adaptation under common training and validation protocols, separating the value of uniform rank allocation from that of full layer coverage.

2 LAYER SCORES AND FUNCTION-PRESERVING TRANSFORMATIONS

2.1 FROM A SCORE TO AN ALLOCATION

Consider a transformer with L blocks. An allocation chooses a layer set S and rank r_ℓ for each layer. Adapters attach to the seven projection types in each block (q, k, v, o, and MLP up, gate, down). The budget is the sum of adapter sizes, so fewer layers permit larger ranks. The main top-k comparisons use $k = L/4$ and a single score computed from a frozen checkpoint.

For standard additive LoRA, a rank-$r_{\ell,p}$ update to projection p factors into two trainable matrices. With frozen biases, its parameter count is

$$P(S, r) = \sum_{\ell \in S} \sum_{p} r_{\ell,p} \big(d^{\text{in}}_{\ell,p} + d^{\text{out}}_{\ell,p}\big). \tag{1}$$

The projection shapes are constant across layers within each tested model. Assigning a common rank within each layer therefore makes total layer rank $\sum_\ell r_\ell$ proportional to parameter count. This is the budget used to compare full-depth and quarter-depth placements. Algorithms that change the parameterization or distribute ranks separately across projections require their own accounting.

We distinguish the score from the rule that uses it. Gradient averages per-example gradient norms; Fisher averages squared norms. SNIP aggregates absolute elementwise weight-gradient products (Lee et al., 2019b). Contraction sums absolute matrixwise weight–gradient pairings. We evaluate it as an invariant control and as a selector with a first-order loss interpretation. Read replaces a block's output with its input and measures the change in next-token logit. Trained Gradient and Fisher selectors use full-block parameters; checkpoint profiles use the seven projection weights. Appendix B specifies parameter sets and estimators. The orbit table tests four scalar scores and a rank-8 matrix-sketch control. These complement deployed selectors through distinct interventions.

2.2 AN INVARIANCE TEST FOR SCORE INTERPRETATION

A transformer has multiple parameterizations implementing the same function. For example, a bilinear product is unchanged when one factor is multiplied by an invertible matrix and the other by its inverse. We use this freedom to vary weights within a block without modifying input-output maps.

Definition 1 (Function-preserving transformation)**.** Let f_θ denote the network function. An invertible map g on parameter space is function-preserving if $f_{g\theta}(x) = f_\theta(x)$ for every admissible parameter vector θ and input x. For a group $\mathcal{G}$ of such maps, the orbit of θ is $\{g\theta \mid g \in \mathcal{G}\}$.

We also call these maps gauge transformations. The group used in this paper acts inside transformer blocks and preserves layer identities. We keep the query and key projections and the residual-stream coordinates fixed.

Theorem 1 (Internal transformations used in the audit)**.** *Consider deterministic RoPE attention with a linear value path and a SwiGLU MLP. The following changes preserve the block function, with biases transformed consistently when present. For each KV head h, choose an invertible matrix S_h and set*

$$W'_{V,h} = S_h W_{V,h}, \qquad W'_{O,i} = W_{O,i} S_h^{-1}$$

for every query head i sharing that KV head. For the MLP, an invertible diagonal matrix D gives

$$W'_{\text{up}} = D W_{\text{up}}, \qquad W'_{\text{down}} = W_{\text{down}} D^{-1},$$

with the gate unchanged. A common permutation of the gate and up channels, compensated by the inverse permutation in down, also preserves the block.

The attention transformation leaves attention weights fixed and cancels between value and output. The MLP scaling cancels around the elementwise gate, and a shared permutation commutes with that gate. These arguments do not require treating SiLU as homogeneous. Appendix C gives the constructions and proofs. We use orthogonal, diagonal, and condition-controlled general matrices to test the resulting invariances.

Preserving the pretrained function is one requirement for a placement audit. We also require the available updates to agree at each fixed set of layers. The following result establishes this stronger property for standard additive LoRA.

Table 1: **Different scores have different invariances.** Qwen3-0.6B across 140 certified orbit points. Each class reports largest relative deviation and smallest top-k Jaccard; ✓ inv. indicates numerical agreement with invariance. Scores use affected projection modules (App. D).

Layer score	Definition	Perm.	Diag.	Orth.	GL
Gradient	$N^{-1}\sum_n \|g_{n,A_\ell}\|_2$	✓	× dev. 1.85 J=0.75	✓	× dev. 0.42 J=1.00
Fisher	$N^{-1}\sum_n \|g_{n,A_\ell}\|_2^2$	✓	× dev. 7.13 J=0.75	✓	× dev. 1.01 J=0.75
SNIP	$N^{-1}\sum_n \sum_{m\in A_\ell} \|W_m \odot G_{n,m}\|_1$	✓	✓	× dev. 0.15 J=0.75	× dev. 1.21 J=0.75
Contraction	$N^{-1}\sum_n \sum_{m\in A_\ell} \|\langle W_m, G_{n,m}\rangle\|$	✓	✓	✓	✓

Theorem 2 (Fixed-placement LoRA adaptation sets)**.** *Fix a layer set S and rank bounds $r_{\ell,p}$ for each of the seven target projections $p \in \{q, k, v, o, \mathrm{up}, \mathrm{gate}, \mathrm{down}\}$ in every $\ell \in S$. Let $\mathcal{F}_{S,r}(\theta)$ be the deterministic forward functions obtained by adding weight updates of rank at most $r_{\ell,p}$ at these projections, with no updates outside S. For every internal gauge g in Theorem 1, with base biases transported consistently when present,*

$$\mathcal{F}_{S,r}(\theta) = \mathcal{F}_{S,r}(g\theta).$$

The theorem gives a common capacity reference for the audit: each fixed placement admits the same forward functions in the two representations. We test whether a layer score recommends the same placement under this change, then evaluate the selected placements by finetuning. A constructive argument explains the equality. Under an invertible matrix action $W' = AWB$, transport an allowed update as

$$\Delta W' = A\Delta W B, \qquad \mathrm{rank}(\Delta W') = \mathrm{rank}(\Delta W).$$

This transport leaves zero updates zero and is invertible at each selected projection. The structural cancellation in Theorem 1 holds for the updated block as well as the pretrained block. Hence the transported updates realize the same functions, and the inverse transport gives the reverse inclusion. Appendix C.3 supplies the full proof, including shared attention heads and gate permutations. The equality concerns attainable deterministic functions; initialization and optimizer dynamics are evaluated separately.

Definition 2 (Function-only score)**.** With the calibration data fixed, a layer score depends only on the network function if $s_\ell(\theta) = s_\ell(\widetilde{\theta})$ whenever two parameterizations of the same architecture satisfy $f_\theta = f_{\widetilde{\theta}}$ and layer identities are aligned. Such a score must satisfy $s_\ell(g\theta) = s_\ell(\theta)$ for every transformation in the audit group.

A certified score change identifies parameterization dependence. Stable scores show agreement under the tested transformations. These outcomes assess a function-only interpretation under a specified transformation set, while a separate training comparison assesses the score's utility for the chosen optimizer.

2.3 Predicted responses of the scores

The gradient transformation determines how each score responds. Multiplying a weight matrix by a scalar c changes its gradient by c^{-1} under the compensated transformation. A Euclidean norm retains this scale dependence. Pairing the weight with its gradient cancels it. SNIP also cancels diagonal scaling element by element, but taking an elementwise absolute value does not commute with a general rotation. These laws hold for each calibration example and therefore for the corresponding averaged estimators. Proofs and matrix-sketch definitions appear in Appendix C.

Proposition 1 (Transformation laws)**.** *Assume the full internal transformation preserves the differentiable calibration loss, $\mathcal{L}(g\theta) = \mathcal{L}(\theta)$ for every admissible θ. Write $G = \nabla_W \mathcal{L}$. Under a left action $W' = SW$, the gradient becomes $G' = S^{-\top}G$. Under a right action $W' = WT^{-1}$, it becomes $G' = GT^\top$. A one-sided action preserves $\|G\|_F$ for all G if and only if its acting matrix is orthogonal. SNIP is invariant under permutations and diagonal rescaling, but not under general rotations. The pairing $\mathrm{tr}(W^\top G)$ is invariant under either invertible action, and so is the sum of its matrixwise absolute values.*

A scalar example makes the distinction concrete. Apply a common value scaling $W'_V = cW_V$ and compensate it with $W'_O = c^{-1}W_O$, for $c > 0$, with value biases transported when present. The attention output is unchanged, while the two affected gradient terms satisfy

$$\|G'_V\|_F^2 + \|G'_O\|_F^2 = c^{-2}\|G_V\|_F^2 + c^2\|G_O\|_F^2. \tag{2}$$

Thus the same input-output map can have different gradient magnitudes. Each matrixwise weight–gradient pairing remains unchanged because its two scale factors cancel. This example links the transformation laws to layer ranking: a norm-based score measures sensitivity in the chosen coordinates, whereas the contraction measures a quantity preserved by this change. The downstream comparison then tests the utility of the layer sets these scores recommend.

The audit measures both score changes and changes in the induced top-k set. The latter is important for allocation. A numerical change that leaves the ranking intact has a different consequence from one that recommends training different layers.

3 Function-preserving changes alter selected layers

We first evaluate the transformation laws on a dense orbit of Qwen3-0.6B. The 7 settings comprise an orthogonal V/O transformation, three general V/O settings with condition number $\kappa \in \{2, 4, 8\}$, diagonal rescaling in V/O and in the MLP, and an MLP channel permutation. Each setting uses the same number of sampled transformations, giving 140 orbit points in total.

Before comparing scores, we certify numerical equivalence. Local module outputs receive identical inputs and must agree within the specified tolerance; float64 local checks and evaluation-precision logits assess the implementation (App. D). These checks accompany the exact identities of Theorem 1.

Table 1 agrees with the predicted response of each score. Gradient and Fisher retain orthogonal invariance and change under non-orthogonal transformations. SNIP retains diagonal and permutation invariance but changes under rotations. Contraction and the matrix-sketch control remain invariant. All 35 cells in the full seven-setting audit match their predicted class. Fisher's relative deviation reaches 7.13 under MLP rescaling; Gradient's selection overlap falls to 0.75.

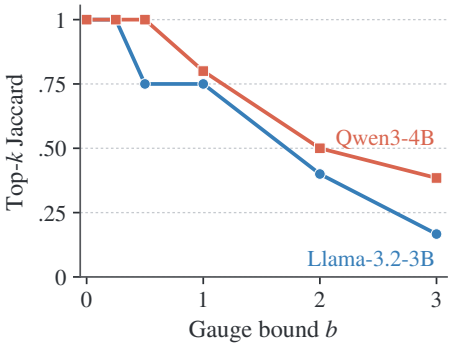


Figure 2: **Selection overlap under rescaling.** Top-k Jaccard versus gauge bound b (seed 18000).

The cross-model study covers five checkpoints in six model–calibration-task configurations. Gradient-selected sets change in every configuration. With mathematics calibration on Llama, for example, overlap is 0.167, replacing 5 of 7 selected layers (App. Figure 4). Read remains unchanged: preserving every block map also preserves the ablated network obtained by replacing a block with identity.

Sensitivity along a controlled scale. Figure 2 varies the scalar bound $b = \max_\ell |\log_2 c_\ell|$ while keeping calibration examples and the transformation seed fixed. Llama's top-7 overlap is 1.000, 0.750, 0.400, and 0.167 at bounds 0, 0.5, 2, and 3. Qwen's top-9 overlap at those bounds is 1.000, 1.000, 0.500, and 0.385. Thus a bounded, exactly compensated change suffices to alter which layers receive adapters. Checkpoint channel-norm ratios provide a separate scale reference (App. R).

Separating coordinates from calibration samples. A 5×5 Llama study crosses five scalar gauges with five batches of 32 examples. Coordinates account for 82.4% of the observed score sum of squares, versus 4.1% for batches, and remain significant against their interaction ($F_{4,16} = 46.82$, $p < 10^{-7}$; App. K).

Finding 1. Equivalent models can recommend different layers. Internal transformations preserve the network function and fixed-placement adaptation capacity, yet change Gradient selections. Contraction and block ablation provide invariant controls.

Table 2: **Changed selections on the same backbone.** Gauge rows show mean ± sample SD over seeds 42–44; Δ is seed-paired. Random rows show mean Jaccard J and the range of mean accuracies across four sets (App. M). All arms use rank eight on the original backbone.

Selection	Selected layer IDs	J	CS-Avg (%)	Paired Δ (pp)
Llama-3.2-3B ($k = 7$ of $L = 28$)				
Original	$\{0, 1, 2, 4, 5, 7, 10\}$	1.000	63.83 ± 0.97	$+0.000 \pm 0.000$
Gauge 18000	$\{0, 1, 3, 4, 7, 11, 17\}$	0.400	65.34 ± 1.81	$+1.508 \pm 0.854$
Gauge 18001	$\{0, 1, 3, 4, 5, 9, 12\}$	0.400	66.10 ± 0.65	$+2.271 \pm 0.587$
Gauge 18002	$\{0, 1, 4, 6, 8, 9, 13\}$	0.273	67.72 ± 0.74	$+3.889 \pm 1.057$
Gauge 18003	$\{0, 1, 4, 6, 12, 14, 15\}$	0.273	65.71 ± 1.44	$+1.874 \pm 1.028$
Random (4 sets)	Layers 0–17	0.368	$[65.42, 65.81]$	—
Qwen3-4B ($k = 9$ of $L = 36$)				
Original	$\{6, 14, 15, 16, 18, 21, 22, 34, 35\}$	1.000	82.10 ± 0.08	$+0.000 \pm 0.000$
Gauge 18000	$\{6, 11, 14, 17, 20, 21, 22, 34, 35\}$	0.500	81.94 ± 0.25	-0.154 ± 0.199
Gauge 18001	$\{1, 5, 6, 9, 12, 18, 19, 34, 35\}$	0.286	81.95 ± 0.19	-0.148 ± 0.179
Gauge 18002	$\{4, 6, 8, 9, 13, 15, 17, 19, 22\}$	0.200	82.12 ± 0.18	$+0.021 \pm 0.128$
Gauge 18003	$\{6, 12, 14, 15, 17, 21, 22, 34, 35\}$	0.636	82.04 ± 0.15	-0.060 ± 0.080
Random (4 sets)	Layers 0–35	0.186	$[81.99, 82.11]$	—

4 Testing the utility of changed selections

The audit establishes that a selector's recommendation can change under equivalent representations. To measure the consequence for adaptation, we hold the training backbone fixed and vary only the selected layer set.

For each of Llama-3.2-3B and Qwen3-4B, we obtain a canonical Gradient top-k set and four sets from certified scalar gauges. We discard the transformed weights and train every set on the original checkpoint. Layer count, rank eight, target projections, data, and training seeds 42–44 are shared within each model. Table 2 lists the layer IDs, overlap with the canonical set, and seed-paired accuracy differences.

Changed selections have model-dependent utility. All four transformed selections improve the observed mean on Llama, with paired gains from $+1.508$ to $+3.889$ percentage points. Their Jaccard overlaps range from 0.273 to 0.400. Qwen's overlaps range from 0.200 to 0.636, but its mean accuracy changes are much smaller, from -0.154 to $+0.021$ points. A substantial change in membership can therefore have a small accuracy effect, while another checkpoint is sensitive to the allocation. Figure 3 displays the individual training seeds alongside these summaries.

Random placements put the gains in context. Llama's canonical set lies within layers 0–10, while its transformed sets reach layers 12–17. Four random controls use the same depth cap of 17 and Jaccard overlap in $[0.25, 0.40]$ with the canonical set. They achieve mean accuracies of 65.42–65.81%, compared with 63.83% for canonical Gradient, and improve accuracy in all twelve seed-paired comparisons. Their mean gains range from $+1.587$ to $+1.977$ points. These controls establish useful alternatives within the same depth-constrained region without gradient ranking. Appendix M gives the IDs and individual runs. On Qwen, four random sets yield 81.99–82.11%, close to the canonical mean of 82.10%.

Transfer to generated mathematics answers. The five Llama layer sets are fixed from commonsense calibration before GSM8K finetuning. Every arm uses the original backbone, rank eight, one epoch, and the same three training seeds. The g_0 and g_1 sets improve accuracy at every seed on both tasks. Their mean GSM8K gains are $+1.188$ and $+1.744$ points. The g_2 set, which has the largest commonsense gain, yields a mean GSM8K change of $+0.152$ points, below both g_0 and g_1. Thus the same fixed layer sets can have different utility orderings across downstream tasks (App. P).

Standardizing internal scales. A complementary Llama experiment balances the norms of compensated MLP and attention factors before scoring. Across the original checkpoint and four positive

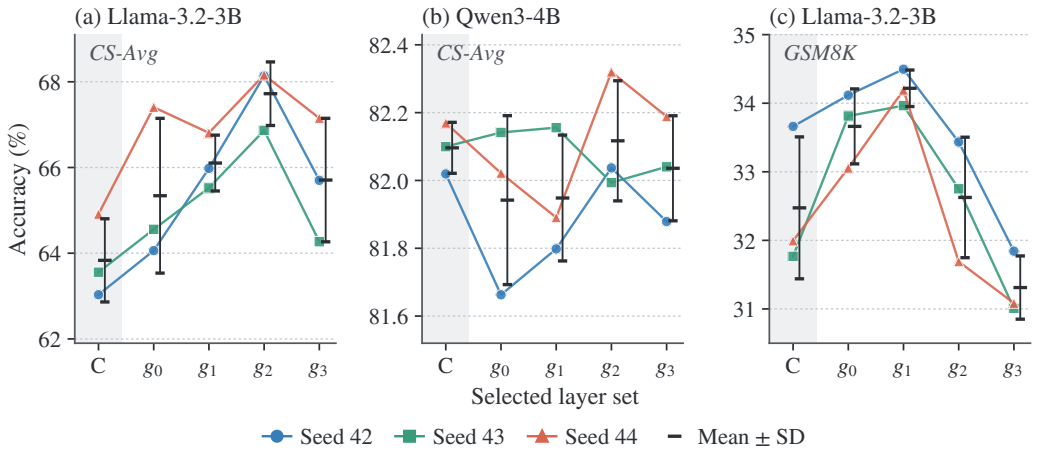


Figure 3: **Training outcomes vary across selections, models, and tasks.** Lines pair training seeds; black bars show mean ± sample SD. C denotes the canonical set, and axes use different ranges. All GSM8K runs use the same first-answer extraction rule (App. P).

scalar gauges, the depth-wise coefficient of variation falls from 0.49–0.58 to 0.08. Direct top-7 sorting recovers $\{0, 4, 8, 12, 16, 20, 24\}$ in all five representations, matching the uniform-quarter reference. This shared set reaches $66.22 \pm 0.74\%$ on the original backbone, compared with 63.83% for canonical Gradient (App. R, Table 20). The intervention preserves the pretrained function while changing the score profile and its recommendation, connecting the early-layer preference to internal parameter scale.

Finding 2. Score sensitivity and adaptation utility are distinct. Equivalent representations produce layer sets with different training outcomes. Fixed-backbone comparisons, random references, and cross-task transfer measure the utility of those recommendations directly.

5 Controlled Allocation Comparisons

5.1 Reproducible Initialization and Explicit Budgets

We compare full-depth LoRA, uniform quarter-depth placement, Gradient and Contraction top-k, AdaLoRA (Zhang et al., 2023), and FoRA (Park et al., 2026). All six methods match the final effective adapter parameter count. We count the retained A/B factors using Equation 1; AdaLoRA's singular values are absorbed into a factor before evaluation. At budget $2L$, full depth assigns rank two to every layer and static quarter-depth methods assign rank eight to $L/4$ layers. The common final budgets are 3,039,232 parameters on Llama and 4,128,768 on Qwen (App. R).

All methods use the same Commonsense170K subset of approximately 10,000 training examples and 1,000 optimization steps. Standard LoRA uses AdamW and the constant multiplier $\alpha/r = 2$; AdaLoRA redistributes rank over all layers, while FoRA combines Fisher-based selection with orthogonal updates. Seeds are set before adapter construction, with initialization-hash checks for Full-depth and Uniform-quarter. Table 3 reports mean and sample SD over seeds 42–44 at learning rate 2×10^{-4}.

CS-Avg averages BoolQ, PIQA, HellaSwag, WinoGrande, ARC-Easy, ARC-Challenge, and OpenBookQA. We score fixed answer-label continuations by mean token log probability, using the same evaluation catalog for every arm. All parameter counts refer to the final adapter representation; Appendix R details the training schedules and rank allocations.

Full-depth LoRA has the highest observed mean on both models: 68.32% on Llama and 82.37% on Qwen. Its margins over the strongest alternative, AdaLoRA, are 1.61 and 0.17 points. AdaLoRA retains nonzero adapters in every layer, so full-depth uniform allocation also improves on an adaptive allocation with the same coverage. On Llama, AdaLoRA (66.71%), Contraction (65.55%), and

Table 3: **Fixed learning rate.** CS-Avg (%), mean ± sample SD over seeds 42–44 at 2×10^{-4} and matched final adapter budget $2L$.

Allocation	Llama	Qwen
	$2L$	$2L$
Full-depth	**68.32 ±1.07**	**82.37 ±0.12**
Uniform quarter	66.22 ±0.74	82.16 ±0.07
AdaLoRA	66.71 ±0.82	82.20 ±0.09
FoRA	65.12 ±0.95	82.13 ±0.08
Gradient top-k	63.83 ±0.97	82.10 ±0.08
Contraction top-k	65.55 ±2.13	82.08 ±0.07

Table 4: **Selected learning rate.** CS-Avg (%) at seed 42 after validation selection of 4×10^{-4} from $\{1, 2, 4\} \times 10^{-4}$.

Allocation	Llama		Qwen	
	$2L$	$8L$	$2L$	$8L$
Full-depth	**72.31**	**74.25**	**83.55**	**84.37**
Uniform quarter	70.45	73.08	82.87	83.69
AdaLoRA	70.82	73.40	82.95	83.78
FoRA	69.45	73.05	82.80	83.42
Gradient top-k	68.80	72.94	82.76	83.29
Contraction top-k	70.84	72.72	82.86	83.57

FoRA (65.12%) all improve on Gradient top-k (63.83%). The comparison establishes the strength of uniform rank allocation alongside the benefits of broader depth coverage.

5.2 Equal learning-rate search

We next match the learning-rate search over $\{1, 2, 4\} \times 10^{-4}$ at budgets $2L$ and $8L$. Validation seed 45 chooses a rate for each model–budget–method cell; the selected configurations are evaluated at independent training seed 42. All 24 cells select 4×10^{-4}. Table 4 reports the resulting accuracies, with full-depth ranks two and eight and static quarter-depth ranks eight and 32.

Full depth leads in all four model–budget settings. Its margin over the strongest alternative is 1.47 points at Llama $2L$ and 0.85 at Llama $8L$; the corresponding Qwen margins are 0.60 and 0.59. The comparison therefore retains a full-depth advantage after using the same validation search for every method. A follow-up at Llama $2L$ reports a three-seed mean of $72.20 \pm 0.29\%$ for full depth and $70.26 \pm 0.50\%$ for Contraction. An expanded 8×10^{-4} check gives worse validation losses for the tested Llama arms (App. R). These results support the ranking within the evaluated search and budgets.

5.3 Coverage and task-level controls

A broader placement survey includes Qwen2.5-0.5B and compares full, uniform, score-selected, contiguous, and random allocations. Among quarter-depth placements, uniform spacing exceeds the best contiguous block mean by 0.86–6.01 points across the three models. This comparison holds layer count and rank fixed, providing a direct descriptive reference for the role of coverage (App. I). The survey's five-run initialization protocol is distinct from the controlled three-seed experiments; each result is summarized within its own protocol.

The same survey also reports per-task differences between full depth and Gradient. Excluding HellaSwag, full depth retains positive six-task mean differences of 1.85, 3.55, and 0.28 points on Qwen2.5, Llama, and Qwen3, respectively (App. N). Thus its survey advantage is supported by the remaining tasks as well. The reported metric is the specified answer-label evaluation; Appendix R records evaluation and implementation details.

Finding 3. Full-depth LoRA is a strong allocation reference. It achieves the highest observed means in the controlled core comparison and the highest seed-42 accuracy in every evaluated model–budget setting after equal learning-rate search, including comparisons with AdaLoRA and FoRA.

6 Implications for evaluating layer selection

The transformation laws clarify a score's interpretation, explicit layer IDs record its recommendations, and fixed-backbone training measures their utility. Controlled allocation comparisons place that utility in the context of an explicit budget.

Use invariance to clarify a score's interpretation. An invariant statistic can be interpreted consistently across the tested internal coordinate changes. A coordinate-dependent statistic measures a sensitivity in the chosen parameterization and should be evaluated in that role. The scalar example in Equation 2 provides a simple check, while block ablation and contraction supply complementary controls. Balancing positive scales is a useful additional diagnostic; its measured effect should be distinguished from the choice of allocation rule.

Include simple allocation references. Full-depth low-rank LoRA is the strongest accuracy reference in the controlled comparisons, and uniform quarter-depth placement provides a score-free reference at fixed layer count. Random placements help establish how much a scoring rule contributes beyond the geometry of its selected set. Reporting layer IDs alongside accuracy makes these comparisons concrete: a new ranking matters through the locations and ranks it ultimately assigns.

State which resource is held fixed. Our six-method comparisons match final effective adapter parameters and optimization steps. Adaptive pruning additionally has a changing training-time allocation, which can be reported alongside the final budget. In the recorded Qwen2.5 measurement, full depth takes 274 seconds for calibration plus training, compared with 171 seconds for Gradient top-k (App. K). A deployment choice can weigh accuracy and cost together. Five-seed tests retain all nine advantages after Holm correction ($p < 0.01$); DoRA/PiSSA and 8B results extend the allocation comparison (Apps. S–U).

7 Related work

Selecting layers and allocating rank. PEFT reduces trainable parameter count (Hu et al., 2022; Houlsby et al., 2019; Mangrulkar et al., 2022). Layer and rank allocation use selective finetuning (Lee et al., 2019a; Shi et al., 2025), rank-component pruning (Zhang et al., 2023), dynamic sampling (Pan et al., 2024), adapter statistics (Zhou et al., 2025; Xu et al., 2026), orthogonal updates (Park et al., 2026), or representation-sensitivity probes (Liu et al., 2026). We isolate layer scoring from these broader adaptation algorithms.

Saliency and parameterization. Gradient norms and weight–gradient pairings assess pruning saliency (Lee et al., 2019b; Molchanov et al., 2019; Saket, 2026). Pruning studies highlight simple baselines and contextual evaluation (Frankle & Carbin, 2019; Frankle et al., 2021; Liu et al., 2019; 2022; Michel et al., 2019). Our contribution tests layer-selection rules under exact function-preserving rewrites, evaluating downstream utility separately.

Neural-network symmetries. Function and parameterization distinction appears in natural gradients and sharpness (Amari, 1998; Dinh et al., 2017). Recent work studies transformer symmetries in optimization (da Silva et al., 2025) and gauge-invariant LoRA updates in differential privacy (Wang & Zhang, 2026). Our audit acts on backbone coordinates and connects fixed-placement adaptation-set invariance (Theorem 2) to layer recommendations and their measured training utility.

Activation scale across depth. Pretrained LLMs exhibit massive activations and outlier features (Sun et al., 2024; Dettmers et al., 2022), alongside position-specific attention sinks (Xiao et al., 2024). Our audit leaves residual-block boundaries fixed, so across-layer activation profiles and internal function-preserving interventions provide complementary evidence.

8 Conclusion

Layer selection should be assessed through both representation and training. Our capacity-preserving audit shows that equivalent pretrained networks can recommend different layers while retaining the same LoRA function set at every fixed placement. Crossed calibration controls and norm balancing connect this sensitivity to internal scale; fixed-backbone training, random placements, and task transfer establish the utility of the recommendations. Under matched final adapter budgets and common training protocols, full-depth uniform low-rank LoRA outperforms the tested static and adaptive alternatives. Together, these results provide a principled evaluation of layer selection: identify coordinate dependence, measure allocation utility, and compare against distributed adaptation at a common budget.

AI USE STATEMENT

AI tools were used primarily for manuscript writing, organization, and language editing. They also assisted with figure and table preparation, the interpretation and presentation of results, and proof exposition and consistency checks. The authors reviewed and corrected the AI-assisted material and take responsibility for the final manuscript, including its claims, experiments, and references.

REPRODUCIBILITY STATEMENT

Score definitions and estimators are specified in Appendix B; the compensated transformations, adaptation-set proofs, and certification checks are given in Appendices C and D. Section 5 and Appendices H and R describe calibration, layer selection, initialization, training, evaluation, and parameter accounting. Selected layer sets and task-transfer protocols appear in Appendices M and P; the extended seed, factorization, and scale comparisons are documented in Appendices S–U. The accompanying source package includes the manuscript, frozen figure inputs, tabulated results, and numerical verification scripts.

REFERENCES

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Marvin F. da Silva, Felix Dangel, and Sageev Oore. Hide & seek: Transformer symmetries obscure sharpness & Riemannian geometry finds it. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55591–55613. PMLR, 2025.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Pruning neural networks at initialization: Why are we missing the mark? In *International Conference on Learning Representations*, 2021.

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019.

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.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5254–5276. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.319.

Jaejun Lee, Raphael Tang, and Jimmy Lin. What would elsa do? freezing layers during transformer fine-tuning. *arXiv preprint arXiv:1911.03090*, 2019a.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. SNIP: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2019b.

Jiaqi Liu, Haidong Kang, Qihui Zhao, Guo Yu, and Jingchao Wang. RSRA: Training-free probing of representation sensitivity for efficient LoRA rank allocation. *arXiv preprint arXiv:2607.09757*, 2026.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024.

Shiwei Liu, Tianlong Chen, Xiaohan Chen, Li Shen, Decebal Constantin Mocanu, Zhangyang Wang, and Mykola Pechenizkiy. The unreasonable effectiveness of random pruning: Return of the most naive baseline for sparse training. In *International Conference on Learning Representations*, 2022.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art parameter-efficient fine-tuning methods, 2022.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3846.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11264–11272, 2019.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Rui Pan, Xiang Liu, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, and Tong Zhang. LISA: Layerwise importance sampling for memory-efficient large language model fine-tuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Juneyoung Park, Seongbae Lee, Han-Sang Lee, Kyuho Lee, Minjae Kim, Seungheon Hyeon, Kiduk Kwon, Seongwan Kim, and Jaeho Lee. FoRA: Fisher-orthogonal rank adaptation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2605.29317*, 2026.

Abdulmalek Saket. Aletheia: Gradient-guided layer selection for efficient LoRA fine-tuning across architectures. *arXiv preprint arXiv:2604.15351*, 2026.

Guangyuan Shi, Zexin Lu, Xiaoyu Dong, Wenlong Zhang, Xuanyu Zhang, Yujie Feng, and Xiao-Ming Wu. Understanding layer significance in LLM alignment. In *Second Conference on Language Modeling*, 2025.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024.

Hong Wang and Kelly Wang. Complete characterization of gauge symmetries in transformer architectures. In *Proceedings of the 4th (2025) and 3rd (2024) NeurIPS Workshops on Symmetry and Geometry in Neural Representations*, volume 282 of *Proceedings of Machine Learning Research*, pp. 622–663. PMLR, 2026.

Shihao Wang and Xueru Zhang. PRISM: Gauge-invariant tangent-space differentially private LoRA. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Yichen Xu, Yuyang Liang, Shan Dai, Tianyang Hu, Tsz Nam Chan, and Chenhao Ma. Understanding and guiding layer placement in parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2602.04019*, 2026.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

Hongyun Zhou, Xiangyu Lu, Wang Xu, Conghui Zhu, Tiejun Zhao, and Muyun Yang. LoRA-drop: Efficient LoRA parameter pruning based on output evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5530–5543. Association for Computational Linguistics, 2025.

A Interpreting the audit and training comparisons

Protocols and evaluation roles. Table 5 separates the Llama-3.2-3B results by experiment. The broad survey reports five recorded runs at 2×10^{-4}; its Full-depth and Gradient means are $67.75 \pm 1.00\%$ and $61.37 \pm 0.96\%$. The core placement comparison resets initialization seeds before adapter construction and reports three-seed means of $68.32 \pm 1.07\%$ and $63.83 \pm 0.97\%$ at the same learning rate. The coordinate-matched experiment crosses backbone coordinates and selected locations under its own recorded initialization protocol. The validation-selected comparison uses 4×10^{-4} and an independent seed-42 final evaluation, giving 72.31% and 68.80%. Its additional three-seed Full-depth mean is $72.20 \pm 0.29\%$. These experiments retain their own records and uncertainty units. The six-method core and selected-rate comparisons match final effective adapter parameter counts and optimization steps within each model. Appendix R specifies the A/B representation used for counting, including the absorption of AdaLoRA's singular values before evaluation.

Two questions, two tests. The audit tests a necessary condition for interpreting a score as a property of the network function. A change on a certified orbit rules out that interpretation. A coordinate-dependent score may still guide an optimizer operating in those coordinates, so training utility is evaluated separately. Read and contraction provide invariant controls; the fixed-original-backbone study compares changed recommendations without changing optimizer coordinates. The coordinate-by-selection experiment in Appendix L provides a complementary comparison of both coordinates and locations.

Allocation and resources. Uniform-quarter and contiguous-quarter placements fix layer count, rank, and parameter count, directly comparing adapter locations in the depth survey (Appendix I). The controlled placement comparison in Section 5.1 fixes the initialization procedure, and the later grid matches validation-search budgets. Runtime and memory in Appendix K describe their recorded protocols. All claims of highest accuracy refer to observed means within the specified evaluation and search conditions.

Evidence across tasks. The certified audit and checkpoint profiles examine score behavior across their specified model sets. The depth survey covers three backbones and seven commonsense tasks, whereas the controlled placement and equal-grid comparisons cover Llama-3.2-3B and Qwen3-4B. A separate Llama experiment finetunes on GSM8K with fixed commonsense-derived selections (Appendix P); Appendix O examines selection stability under mathematics calibration.

B Score definitions and estimators

Scores are evaluated at fixed backbone weights on a shared calibration sample. For the placement comparison, this sample consists of 32 training-corpus examples selected by the index rule in App. H. It is not a separate held-out test set. Let $\mathcal{L}_n$ be the mean causal language-modeling loss on the tokens of example n, let $G_{W,n} = \nabla_W \mathcal{L}_n$, and let N be the number of examples retained after tokenization. We distinguish the full parameter vector θ_ℓ of block ℓ from its seven LoRA target projections

$$\mathcal{W}_\ell = \{W_q, W_k, W_v, W_o, W_{\text{up}}, W_{\text{gate}}, W_{\text{down}}\}.$$

The full vector also includes normalization parameters and any biases. For module scores, a linear map with bias is represented by $\widetilde{W} = [W\ b]$ and $\widetilde{G}_{W,n} = [G_{W,n}\ \nabla_b \mathcal{L}_n]$. Without a bias, these reduce to W and $G_{W,n}$. The scoring scope can therefore exceed the adapted parameters. LoRA trains low-rank additions to the projection weights with `bias=none`, leaving the original biases and normalization parameters fixed.

The checkpoint score profiles are a separate weight-only measurement over the seven projections, excluding biases and normalization parameters. The full-block definitions below describe the deployed selectors.

Gradient and Fisher. The deployed layer scores are

$$s_\ell^{\text{Grad}} = \frac{1}{N} \sum_{n=1}^{N} \|\nabla_{\theta_\ell} \mathcal{L}_n\|_2, \qquad s_\ell^{\text{Fisher}} = \frac{1}{N} \sum_{n=1}^{N} \|\nabla_{\theta_\ell} \mathcal{L}_n\|_2^2.$$

The norm is taken after stacking all parameter gradients within a block, and before averaging over examples. Gradient is neither a sum of matrix norms nor the norm of the mean gradient. Fisher is

Table 5: **Llama-3.2-3B protocol reference.** Every row uses budget $2L$ within its own experiment. CS-Avg (%) is mean ± sample SD across the listed seeds, or one final result at seed 42.

Experiment	Learning rate	Seeds	Full-depth	Gradient
Survey (Table 10)	2×10^{-4}	42–46	67.75 ± 1.00	61.37 ± 0.96
Core (Table 3)	2×10^{-4}	42–44	68.32 ± 1.07	63.83 ± 0.97
Coordinates θ (Table 14)	2×10^{-4}	42–44	68.03 ± 0.69	62.64 ± 1.79
Coordinates $g\theta$ (Table 14)	2×10^{-4}	42–44	68.82 ± 0.77	64.54 ± 0.80
Selected LR (Table 4)	4×10^{-4}	42	72.31	68.80

the trace of the empirical matrix $N^{-1}\sum_n g_{\ell,n}g_{\ell,n}^\top$, where $g_{\ell,n} = \nabla_{\theta_\ell}\mathcal{L}_n$. It is not $\|N^{-1}\sum_n g_{\ell,n}\|_2^2$. Consequently, Fisher and Gradient need not induce identical rankings. Their stored core top-k sets coincide on 2 of the 3 models and differ on Llama-3.2-3B. These definitions audit ambient Euclidean gradient and empirical-Fisher summaries. They do not establish the behavior of natural-gradient metrics or every component of a method using related statistics (Saket, 2026; Park et al., 2026; Zhang et al., 2023).

SNIP and contraction. For a specified module set $\mathcal{M}_\ell$, we use

$$s_\ell^{\mathrm{SNIP}}(\mathcal{M}_\ell) = \frac{1}{N}\sum_n \sum_{W\in\mathcal{M}_\ell} \|\widetilde{W} \odot \widetilde{G}_{W,n}\|_1,$$

$$s_\ell^{\mathrm{Ctr}}(\mathcal{M}_\ell) = \frac{1}{N}\sum_n \sum_{W\in\mathcal{M}_\ell} |\langle \widetilde{W}, \widetilde{G}_{W,n}\rangle|.$$

Here $\|\cdot\|_1$ is the sum of absolute entries. Our SNIP-style statistic averages per-example connection-saliency magnitudes (Lee et al., 2019b), rather than taking the magnitude after averaging gradients as in the original minibatch criterion. We retain the short label SNIP in the audit. Contraction takes the absolute value of the signed module pairing before summing modules and averaging examples. The signed pairing is the derivative at $\varepsilon = 0$ of the loss when that module's weight and bias are multiplied by $1 + \varepsilon$. It therefore measures a specified first-order perturbation, as in Taylor saliency (Molchanov et al., 2019). Absolute values prevent cancellation between modules or examples. The deployed contraction uses all seven projections, $\mathcal{M}_\ell = \mathcal{W}_\ell$, including their biases.

Read. Read measures the effect of removing a block's residual contribution. Let $z_\theta(x_n)$ be the next-token logit vector at the final prompt position and $t_n = \arg\max_t z_\theta(x_n)_t$. Let $z_\theta^{(-\ell)}(x_n)$ be the logits obtained by replacing block ℓ's output with its input. The implemented score is

$$s_\ell^{\mathrm{Read}} = \frac{1}{N}\sum_n \left| z_\theta(x_n)_{t_n} - z_\theta^{(-\ell)}(x_n)_{t_n} \right|.$$

The target token is chosen by the unmodified model and held fixed during ablation. Read is a block-ablation logit statistic, rather than a norm of the residual write. Its internal-gauge invariance follows from the block-boundary intervention (App. E). It does not replicate the adapter-output norm of LoRA-drop (Zhou et al., 2025) or the representation probe of RSRA (Liu et al., 2026).

Module scope in the orbit audit. Table 7 aggregates only modules affected by the tested family. These are V, O for a value/output transformation, $\mathrm{up}, \mathrm{down}$ for MLP scaling, and $\mathrm{gate}, \mathrm{up}, \mathrm{down}$ for an MLP permutation. Writing this set as $\mathcal{A}_\ell$, the audited Gradient and Fisher are

$$\frac{1}{N}\sum_n \left(\sum_{W\in\mathcal{A}_\ell} \|\widetilde{G}_{W,n}\|_F^2 \right)^{1/2}, \qquad \frac{1}{N}\sum_n \sum_{W\in\mathcal{A}_\ell} \|\widetilde{G}_{W,n}\|_F^2.$$

SNIP and contraction use the same affected-module set. The cross-model deployed scores instead use the full-block definitions above. The allocation diagnostic in App. G also uses all block parameters for Gradient and Fisher, and all seven projections for the other module-based quantities. These scopes are not interchangeable.

Random projection control. The orbit audit also uses a rank-8 random projection of a contraction matrix. For a left action, set $C_{W,n} = \widetilde{W}^\top \widetilde{G}_{W,n}$. For a right action, set $C_{W,n} = G_{W,n} W^\top$ using the weight matrix alone. The bias is unchanged by the right action and is omitted from this matrix control, although its pairing remains in the scalar contraction. With a fixed probe Ω_W whose independent Gaussian entries have variance $1/8$, the score is

$$s_\ell^{\mathrm{sketch}}(\mathcal{M}_\ell) = \frac{1}{N} \sum_n \sum_{W \in \mathcal{M}_\ell} \|C_{W,n} \Omega_W\|_F.$$

The products are evaluated as $\widetilde{W}^\top (\widetilde{G}_{W,n} \Omega_W)$ on the left and $G_{W,n}(W^\top \Omega_W)$ on the right. Probe seeds are fixed by layer and module and reused along the orbit. Untouched Q, K modules use the left form in the full-layer diagnostic. The control tests a matrix invariant and is not a deployed LoRA utility estimator.

C Function-preserving transformations and their scope

We establish sufficient internal gauges for the audit. The constructions preserve each block's input-output map and do not require a classification of the model's maximal symmetry group. The checkpoints use RoPE attention and SiLU-SwiGLU MLPs, with grouped-query attention or QK-RMSNorm in some architectures. The proofs concern deterministic evaluation without dropout. Attention probabilities depend on Q, K, and the attention kernel does not normalize the value path.

C.1 The attention value and output sector

Theorem 3 (A value/output gauge subgroup)**.** *Represent token features as columns. Let X contain the input tokens and let $V_h = W_{V,h} X + b_{V,h} \mathbf{1}^\top$. Query head i uses KV head $h = \lfloor i/g \rfloor$, where g is the GQA group size, and has output $W_{O,i} V_h A_i^\top$, with A_i its token-attention matrix. For arbitrary $S_h \in \mathrm{GL}(d_v)$, the transformations*

$$W'_{V,h} = S_h W_{V,h}, \qquad b'_{V,h} = S_h b_{V,h}, \qquad W'_{O,i} = W_{O,i} S_h^{-1} \quad \textit{when } \lfloor i/g \rfloor = h$$

preserve the attention output. They define a gauge action of $\prod_{h=1}^{n_{\mathrm{kv}}} \mathrm{GL}(d_v)$.

Proof. The transformation leaves Q, K, and hence A_i, unchanged. Since $V'_h = S_h V_h$,

$$W'_{O,i} V'_h A_i^\top = W_{O,i} S_h^{-1} S_h V_h A_i^\top = W_{O,i} V_h A_i^\top.$$

The identity holds for every head and input. Summing heads and keeping the additive output bias fixed preserves the block map. Composition and inversion follow from those of the matrices S_h. □

This construction acts independently within each KV head. General dense mixing across heads need not preserve their distinct attention patterns. That observation motivates our block-diagonal construction but does not exclude special cases with coincident patterns, head permutations with joint parameter transport, or other symmetries (Wang & Wang, 2026).

The audit draws orthogonal matrices by QR decomposition of Gaussian matrices, positive diagonal matrices, and general invertible matrices with condition numbers $\kappa \in \{2, 4, 8\}$. Within these orbit families, the geometric mean of the singular values is normalized to one. GQA repetition follows the implementation's expand-then-reshape order and is checked by the round-trip test in App. D. The separate cross-model experiment uses reciprocal scalar factors.

Proposition 2 (Restrictions on pre-RoPE query/key maps)**.** *Without QK normalization, consider paired maps $q' = Mq$ and $k' = M^{-\top} k$ before RoPE. They preserve attention inner products at all positions for arbitrary q, k exactly when*

$$M^\top R_t^\top R_u M^{-\top} = R_t^\top R_u \quad \textit{for all positions } t, u.$$

Thus arbitrary $\mathrm{GL}(d_k)$ maps are not generally admissible, although a nontrivial commuting subgroup can remain.

Proof. The transformed inner product is $q^\top M^\top R_t^\top R_u M^{-\top} k$. Equality for all q, k is equivalent to the displayed matrix identity. For standard RoPE rotations, this requires compatibility with their relative rotations. Reciprocal scalar maps satisfy this condition in the absence of QK normalization. □

QK-RMSNorm introduces additional constraints. In particular, its nonzero ε prevents arbitrary input scalings from cancelling through normalization. This does not eliminate every query/key symmetry. The reported families act on V/O and the MLP and make no completeness claim for the query/key sector.

C.2 The SwiGLU sector

Write

$$\mathrm{MLP}(x) = W_{\mathrm{down}}\big(\mathrm{SiLU}(W_{\mathrm{gate}}x + b_{\mathrm{gate}}) \odot (W_{\mathrm{up}}x + b_{\mathrm{up}})\big) + b_{\mathrm{down}}.$$

Lemma 1 (Dense mixing with a fixed gate)**.** *If $B(s \odot Au) = s \odot u$ for every $s, u \in \mathbb{R}^{d_{\mathrm{ff}}}$, then A is invertible and diagonal and $B = A^{-1}$.*

Proof. Setting $s = \mathbf{1}$ gives $BA = I$. The full identity then says $A^{-1}\,\mathrm{diag}(s)A = \mathrm{diag}(s)$ for every s. A matrix commuting with every diagonal matrix is diagonal. □

Lemma 2 (Scalar rescaling of a SiLU gate)**.** *For nonzero scalars a, b, c, if $c^{-1}b\,\mathrm{SiLU}(at) = \mathrm{SiLU}(t)$ for every t, then $a = 1$ and $b = c$.*

Proof. Since $\mathrm{SiLU}(t) = t/2 + t^2/4 + O(t^4)$ near zero, equality of the linear and quadratic coefficients gives $ba/c = 1$ and $ba^2/c = 1$. Hence $a = 1$ and $b = c$. □

Theorem 4 (A SwiGLU gauge subgroup)**.** *Let P be a hidden-channel permutation and D an invertible diagonal matrix. The following transformations preserve the MLP map*

$$\begin{aligned} W'_{\mathrm{up}} &= PDW_{\mathrm{up}}, & b'_{\mathrm{up}} &= PDb_{\mathrm{up}}, \\ W'_{\mathrm{gate}} &= PW_{\mathrm{gate}}, & b'_{\mathrm{gate}} &= Pb_{\mathrm{gate}}, \\ W'_{\mathrm{down}} &= W_{\mathrm{down}}(PD)^{-1}, & b'_{\mathrm{down}} &= b_{\mathrm{down}}. \end{aligned}$$

They generate the monomial subgroup $\mathfrak{S}_{d_{\mathrm{ff}}} \ltimes (\mathbb{R} \setminus \{0\})^{d_{\mathrm{ff}}}$.

Proof. Let $s = \mathrm{SiLU}(W_{\mathrm{gate}}x + b_{\mathrm{gate}})$ and $u = W_{\mathrm{up}}x + b_{\mathrm{up}}$. Elementwise SiLU commutes with permutations, so the transformed hidden product is $(Ps) \odot (PDu) = PD(s \odot u)$. The inverse factor in W'_{down} cancels PD, and the output bias is unchanged. □

Lemma 1 rules out dense mixing when the gate coordinates are fixed and its identity is required for arbitrary s, u. Lemma 2 rules out the stated scalar compensation through SiLU. These lemmas justify restrictions on the chosen construction; they do not classify all parameter-dependent degeneracies or joint transformations of an arbitrary trained MLP.

Residual coordinates. All audited families keep the residual-stream basis fixed. Generic rotations or rescalings of that basis need not commute with learned elementwise normalization weights or with the normalization ε. Coordinated residual permutations can nevertheless remain symmetries when surrounding parameters are transported consistently. We exclude residual-basis changes from the scope of this audit, without claiming that the architecture excludes every such symmetry.

C.3 Fixed-placement LoRA adaptation sets

Proof of Theorem 2. Let Δ collect the weight updates allowed by the fixed layer set S and the rank bounds $r_{\ell,p}$. Standard additive LoRA represents every matrix of rank at most its factor width; any nonzero fixed scaling can be absorbed into a factor. The gauge acts within each block and leaves the selected layer identities unchanged.

For the value/output gauge, let $U_V = \mathrm{blkdiag}(S_h)_h$ act on the complete value projection and $U_O = \mathrm{blkdiag}(S_{\lfloor i/g_{\mathrm{GQA}} \rfloor})_i$ act on the input coordinates of the complete output projection. Transport the updates by

$$\Delta W'_V = U_V \Delta W_V, \qquad \Delta W'_O = \Delta W_O U_O^{-1}, \qquad \Delta W'_Q = \Delta W_Q, \qquad \Delta W'_K = \Delta W_K.$$

Here g_{GQA} is the number of query heads per KV head; the repetition in U_O matches each query head to its shared KV head. For a SwiGLU gauge with permutation P and invertible diagonal D, transport the remaining updates by

$$\Delta W'_{\mathrm{up}} = PD\Delta W_{\mathrm{up}}, \qquad \Delta W'_{\mathrm{gate}} = P\Delta W_{\mathrm{gate}}, \qquad \Delta W'_{\mathrm{down}} = \Delta W_{\mathrm{down}}(PD)^{-1}.$$

The same formulas apply independently in each selected layer; zero updates outside S remain zero. Each update is multiplied only by invertible matrices, so its rank bound is preserved. The maps also have inverses and thus biject the admissible update tuples.

Transform the base biases as in Theorems 3 and 4; LoRA adds no bias updates here. These gauges are structural symmetries for every admissible updated block, not only for the frozen checkpoint. In particular, the query and key updates are identical in both representations, so their attention matrices agree; the value/output factors still cancel for those matrices. Likewise, SiLU commutes with the gate permutation and the up/down factors cancel for every updated MLP. Consequently, $f_{\theta+\Delta} = f_{g\theta+T_g\Delta}$ for the displayed transport T_g. Applying its inverse proves both inclusions of the function sets. This is equality of deterministic representable functions at a fixed placement; the training procedure is not part of the set. □

C.4 Transformation laws of the scores

Theorem 5 (Gradient transformation). *Let $\mathcal{L}_n$ be preserved by the full gauge action. For a matrix factor, a left action $W' = SW$ gives $G' = S^{-\top}G$. A right action $W' = WT^{-1}$ gives $G' = GT^{\top}$.*

Proof. For the left action, $dW' = S\,dW$ and $\langle G', S\,dW\rangle = \langle G, dW\rangle$ for every dW. Thus $S^{\top}G' = G$. For the right action, $dW' = dW\,T^{-1}$ gives $G'T^{-\top} = G$. The argument applies separately to each example and to transported biases. □

Corollary 1 (Reciprocal scaling). *Under a gauge whose left action on a factor is $W' = cW$, $G' = c^{-1}G$ and $W' \odot G' = W \odot G$. The partner factor has reciprocal weight scaling and opposite gradient scaling.*

A full-block norm combines these factors with untouched parameters, including Q, K, normalization parameters, and the gate under up/down scaling. Changing internal scales can therefore alter layer norms by different amounts and change their ranking. This identity describes variation along a gauge orbit. It does not by itself identify the cause of a cross-layer correlation in the original checkpoint.

Theorem 6 (Isometry criterion for Euclidean scores). *For a left action $W' = SW$,*

$$\|G'\|_F^2 = \mathrm{tr}\big(G^{\top}S^{-1}S^{-\top}G\big) = \mathrm{tr}\big(G^{\top}(S^{\top}S)^{-1}G\big).$$

This equals $\|G\|_F^2$ for every G if and only if S is orthogonal. For the right action $W' = WT^{-1}$,

$$\|G'\|_F^2 = \mathrm{tr}(G^{\top}G\,T^{\top}T),$$

which is invariant for every G if and only if T is orthogonal.

Proof. Substitute the gradient transformations from Theorem 5. Equality of the quadratic form for every column, or every row, of G requires its metric matrix to be the identity. This is equivalent to orthogonality of the corresponding transformation. □

The condition holds example by example and survives averaging. Gradient and Fisher are invariant under the audited orthogonal and permutation actions. Non-isometric diagonal or general linear actions need not preserve them, although a particular gradient can coincide accidentally.

Proposition 3 (SNIP under monomial actions). *The entrywise quantity $\|W \odot G\|_1$ is invariant under one-sided monomial actions, including signed permutations and invertible diagonal scalings. It is not invariant under general rotations.*

Proof. A diagonal left action multiplies row i of W by d_i and row i of G by d_i^{-1}. A right action gives the corresponding cancellation by column. A permutation reorders the entrywise products.

For a counterexample outside this subgroup, take

$$W = \begin{pmatrix} 1 & 2 \\ 3 & 5 \end{pmatrix}, \qquad G = \begin{pmatrix} 2 & -1 \\ 4 & 3 \end{pmatrix}.$$

Initially $\|W \odot G\|_1 = 31$, $\mathrm{tr}(W^\top G) = 27$, and $\|G\|_F^2 = 30$. For $S = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, the left action gives SNIP 41 and squared gradient norm 25, while $W'^\top G' = W^\top G$. For the orthogonal matrix $Q = \frac{1}{5}\left(\begin{smallmatrix} 3 & -4 \\ 4 & 3 \end{smallmatrix}\right)$, SNIP becomes 27 while the gradient norm and trace remain unchanged. □

Remark 1 (Scalar and matrix contractions). The scalar pairing $\mathrm{tr}(W^\top G)$ is invariant under either one-sided action. Under a left action, the full matrix $W^\top G$ is unchanged. Under a right action, $GW^\top$ is unchanged, whereas $W^\top G$ transforms by similarity as $T^{-\top}(W^\top G)T^\top$. Thus the appropriate invariant matrix admits many scalar summaries, including spectral functions and functions of individual entries. An off-diagonal entry is not a spectral function, and entries of the similarity-transformed matrix need not remain fixed. The contraction used here is one invariant first-order summary, with no uniqueness or universal utility claim.

D Orbit certification and complete audit results

Certification. The Qwen3-0.6B audit contains 140 points from 7 transformation families with 20 seeds per family. Numerical checks compare local sector outputs in float64, reconstruct the original tensors after an inverse transformation, and compare held-out logits at evaluation precision. The construction leaves parameters outside the transformed sector unchanged; this is not a separate recorded byte check. The recorded maximum deviations are approximately 6.4×10^{-12} for local outputs, 4×10^{-15} for reconstruction, and 10^{-5} for held-out logits. Local module equivalence is the executable admission condition, and all reported points passed it. Round-trip errors are recorded separately; end-to-end logits provide a numerical diagnostic. The algebra proves equality of deterministic functions for all inputs, while finite checks verify the implementation at tested inputs and precisions.

The separate cross-model audit uses power-of-two reciprocal factors. For the opening Llama example, the checked logits are bitwise identical at the 994 recorded positions. This observation concerns the evaluated logits. The transformed weights themselves differ, and exact representability of the factors alone does not guarantee bitwise equality for arbitrary floating-point computations.

Audit units and aggregation. Table 7 crosses Gradient, Fisher, SNIP, absolute module contraction, and the random projection control with seven transformation families, giving 35 audit cells. Read belongs to the separate cross-model audit in Table 6. Each full-orbit entry uses affected-module aggregation and reports the maximum relative score deviation, minimum depth-profile Spearman correlation, and minimum top-k Jaccard across that family's orbit points. Near-zero reference scores follow a fixed exclusion rule for relative errors, detailed in the released scoring code. Table 1 summarizes the four scalar scores over four transformation classes, taking the worst reported value among the included families. These cells check transformation laws and are not independent finetuning experiments.

D.1 Bounded scalar transformation path

Figure 2 uses the same 32 calibration examples at each intensity for Llama-3.2-3B and Qwen3-4B. A fixed gauge seed generates per-layer scalar factors c_ℓ for an exact SwiGLU up/down compensation, with $|\log_2 c_\ell| \le b$ at bounds $b \in \{0, 0.25, 0.5, 1, 2, 3\}$. The top-$k$ set is computed from the same full-block Gradient statistic at every point. Its Jaccard overlap with the canonical set is, in bound order, $1, 1, 0.750, 0.750, 0.400, 0.167$ for Llama ($k = 7$) and $1, 1, 1, 0.800, 0.500, 0.385$ for Qwen ($k = 9$). Each transformed checkpoint passes the deterministic FP64 forward check; the maximum absolute logit difference across transformed points is 1.53×10^{-13} for Llama and 5.24×10^{-13} for Qwen. The scalar map has condition number one at every layer, so the amplitude bound, rather than that condition number, limits the intervention. The recorded resampling statistic compares resampled

Table 6: **Cross-model audit of deployed scores.** Spearman ρ and top-k Jaccard compare the original checkpoint with its transformed representation under the combined V/O and MLP gauge. Gradient changes the selected set in every listed configuration. Read uses the block-ablation definition in App. B.

		Gradient		Read	
Model	Task	ρ	Jaccard	ρ	Jaccard
Qwen2.5-0.5B	commonsense	0.79	0.50	1.00	1.00
Qwen3-0.6B	commonsense	0.52	0.27	1.00	1.00
Llama-3.2-3B	commonsense	0.80	0.56	1.00	1.00
Llama-3.2-3B	math	0.65	0.17	1.00	1.00
Mistral-7B	commonsense	0.85	0.45	1.00	1.00
OLMo-2-7B	commonsense	0.20	0.23	1.00	1.00

Table 7: **Complete orbit audit on Qwen3-0.6B.** Columns show seven transformation settings with equally many certified points, 140 in total. The three row groups report maximum relative score deviation, minimum Spearman ρ, and minimum top-k Jaccard for four scores and the sketch control. Scores aggregate only affected modules. Table 1 summarizes the corresponding invariance classes.

Score	V/O Orth.	V/O GL κ=2	V/O GL κ=4	V/O GL κ=8	V/O Diag.	SwiGLU Diag.	MLP Perm.
Maximum relative deviation							
Gradient	1.66e-06	5.07e-02	1.86e-01	4.17e-01	1.22e+00	1.85e+00	2.51e-06
Fisher	3.41e-06	1.04e-01	4.08e-01	1.01e+00	3.92e+00	7.13e+00	5.38e-06
SNIP	1.51e-01	2.46e-01	5.65e-01	1.21e+00	1.24e-06	1.31e-06	1.06e-06
Contraction	1.37e-05	9.73e-06	9.93e-06	1.69e-05	1.05e-05	6.06e-06	7.32e-06
Sketch r=8	1.71e-06	1.13e-06	1.47e-06	1.58e-06	1.16e-06	1.47e-06	5.42e-06
Minimum Spearman ρ							
Gradient	1.000	0.994	0.992	0.991	0.973	0.976	1.000
Fisher	1.000	0.998	0.997	0.993	0.974	0.970	1.000
SNIP	0.997	0.997	0.995	0.994	1.000	1.000	1.000
Contraction	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Sketch r=8	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Minimum top-k Jaccard							
Gradient	1.000	1.000	1.000	1.000	0.750	0.750	1.000
Fisher	1.000	1.000	1.000	0.750	0.750	0.750	1.000
SNIP	0.750	0.750	0.750	0.750	1.000	1.000	1.000
Contraction	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Sketch r=8	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Orth. = orthogonal; Diag. = channel-diagonal rescaling; Perm. = channel permutation. Sketch r=8 is the rank-8 matrix sketch.

scores *after* each transformation with the canonical set; it includes the transformation effect and is not an independent calibration-only reference.

E READ UNDER INTERNAL GAUGES

Every gauge in Theorems 3 and 4 preserves each block map at a fixed residual boundary. Replacing any one of those block maps by the identity therefore yields the same ablated network before and after the gauge. Both the original logits and the ablated logits in the definition of Read remain fixed. Their absolute difference, its calibration average, and the resulting ranking are invariant.

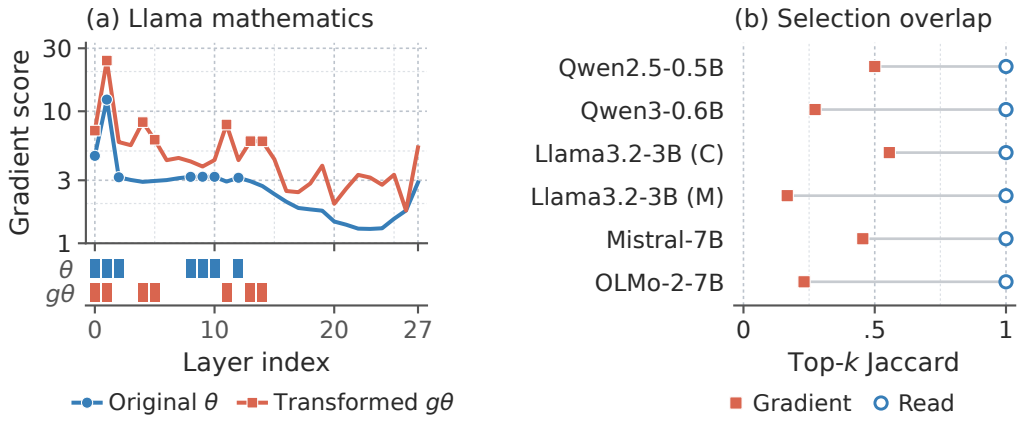


Figure 4: **Cross-model selection audit.** (a) An internal reparameterization of Llama-3.2-3B preserves all 994 checked logits but changes the Gradient profile. On mathematics calibration, top-7 selection replaces 5 layers, with Jaccard overlap 0.167. (b) Gradient selections change in all six model-task combinations. Read retains its set under the same block-preserving transformations. C and M denote commonsense and mathematics; other rows use commonsense.

This argument establishes invariance under the specified internal gauges. It neither classifies residual-coordinate transformations nor shows that Read estimates the benefit of inserting an adapter. Full-depth LoRA exceeds Read top-k by $+1.42$ and $+4.62$ pp on Qwen2.5-0.5B and Llama-3.2-3B, respectively. The Qwen3-4B mean contrast is smaller (Table 11). Invariance under function-preserving transformations is necessary for a function-only interpretation of a score. Passing this audit is not sufficient for that interpretation globally, and does not determine whether a placement helps under a particular optimizer.

F Score profiles and residual scale

Table 8 reports 7 measured profiles from six checkpoints, including Llama-3.2-3B on both commonsense and math text. Each profile uses the seven projection weights only. Maximum absolute activation is measured at the block output for each sample, then averaged across samples. RMS takes the root mean square over hidden coordinates, then averages over tokens and samples. Gradient, Fisher, and SNIP correlate negatively with both summaries on Qwen2.5-0.5B, Llama-3.2-3B, and Mistral-7B. The strength depends on the model and the summary.

The distinction between the two residual summaries matters. Qwen3-0.6B has weak max-absolute correlations but strong negative RMS correlations. Qwen3-4B has near-zero Gradient and Fisher correlations under both summaries, while SNIP retains a negative RMS correlation. OLMo-2-7B also retains negative RMS correlations despite weak max-absolute correlations. Across the seven profiles, the measured rank correlations between Gradient and SNIP range from 0.50 to 0.93. These observations support an association between residual scale and some score profiles. They do not identify residual growth as the sole cause of a layer ranking. In particular, internal gauges change parameter coordinates while preserving the boundary residuals.

G Proportional allocation with a minimum rank

A score alone does not specify which layers receive adapters or how much rank each receives. On the original checkpoints, Gradient selects $\{0, 1, 2, 4, 5, 7, 10\}$ on Llama-3.2-3B, while Qwen3-4B's 9 layers span indices 6–35 (depth profiles and residual-scale correlations appear in Appendix F). Contraction illustrates the distinction between a score and an allocation rule because it passes the internal audit and has a direct first-order interpretation: each signed matrixwise pairing is the first-order loss derivative for a multiplicative change to that matrix, and summing absolute pairings gives a nonnegative layer score. This describes the score, rather than how to distribute a rank budget.

Table 8: **Spearman correlations across depth.** Each score profile is compared with residual maximum absolute activation and residual RMS on the indicated model and corpus. A negative value means that layers with larger residual statistics tend to receive smaller scores.

		vs. residual max-abs			vs. residual RMS		
Model	Corpus	Grad	Fisher	SNIP	Grad	Fisher	SNIP
Qwen2.5-0.5B	commonsense	-0.64	-0.61	-0.46	-0.54	-0.52	-0.97
Qwen3-0.6B	commonsense	-0.22	-0.20	-0.11	-0.81	-0.83	-0.99
Llama-3.2-3B	commonsense	-0.83	-0.82	-0.74	-0.89	-0.88	-0.87
Qwen3-4B	commonsense	+0.14	+0.11	+0.20	+0.04	+0.01	-0.30
Mistral-7B-v0.3	commonsense	-0.55	-0.54	-0.43	-0.88	-0.89	-0.71
OLMo-2-7B	commonsense	+0.13	+0.14	-0.03	-0.31	-0.30	-0.53
Llama-3.2-3B	math	-0.77	-0.78	-0.60	-0.84	-0.85	-0.74

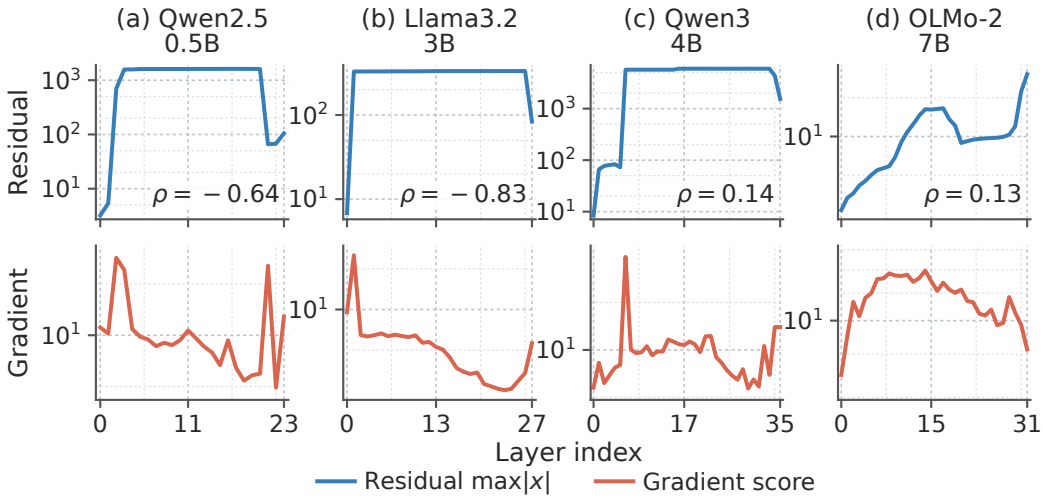


Figure 5: **Gradient profiles and residual magnitude vary together on some checkpoints.** Within each model panel, the upper plot shows the sample-averaged maximum absolute block-output activation, and the lower plot shows Gradient over the seven projection weights. The across-layer Spearman correlations are -0.64 and -0.83 on Qwen2.5-0.5B and Llama-3.2-3B, compared with 0.14 and 0.13 on Qwen3-4B and OLMo-2-7B. These are checkpoint profiles, not interventions on residual magnitude.

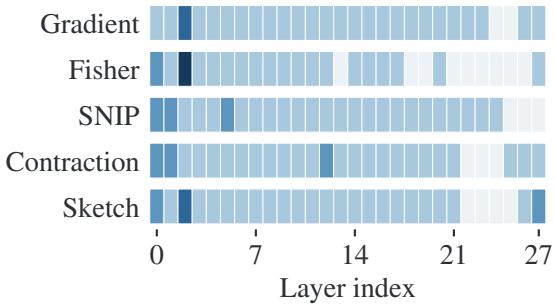


Figure 6: **Score-proportional ranks.** Qwen3-0.6B, budget $2L$. Darker cells denote higher ranks.

Figure 6 distributes total rank $2L$ proportionally using Hamilton rounding, with at least rank one per layer. Contraction has coefficient of variation 0.405 and assigns ranks 1 to 3. Top-k instead assigns rank eight to a quarter of the layers and zero elsewhere; the layerwise values are given below. Top-k

Table 9: **Hamilton allocation on Qwen3-0.6B.** All scores receive the same $2L$ rank budget with a minimum of rank one per layer. Bold entries differ from uniform rank 2 by more than one. The minimum rank imposes full-layer coverage.

Quantity	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Gradient	2	2	**4**	2	2	2	2	2	2	2	2	2	2	2
Fisher	3	2	**10**	2	2	2	2	2	2	2	2	2	2	1
SNIP	3	3	2	2	2	3	2	2	2	2	2	2	2	2
Contraction	3	3	2	2	2	2	2	2	2	2	2	2	3	2
rank-8 matrix sketch	3	2	**4**	2	2	2	2	2	2	2	2	2	2	2

Quantity	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Gradient	2	2	2	2	2	2	2	2	2	2	1	1	2	2
Fisher	2	2	2	2	1	1	2	1	1	1	1	1	1	2
SNIP	2	2	2	2	2	2	2	2	2	2	2	1	1	1
Contraction	2	2	2	2	2	2	2	2	1	1	1	2	2	2
rank-8 matrix sketch	2	2	2	2	2	2	2	2	1	1	1	1	2	3

uses the ranking to choose a subset, while proportional allocation uses score magnitudes and, in this diagnostic, forbids zero-rank layers.

The allocation diagnostic uses Qwen3-0.6B with $L = 28$ layers and a total budget of $2L$ rank units. It assigns rank one to every layer before distributing the remaining L units in proportion to a nonnegative score profile. For $\sum_j s_j > 0$, define

$$a_\ell = \frac{L s_\ell}{\sum_j s_j}, \qquad r_\ell = 1 + \lfloor a_\ell \rfloor + \mathbf{1}\{\ell \text{ receives a remaining unit}\}.$$

Hamilton's method gives the remaining units to the largest fractional parts of a_ℓ, with ties resolved by the implementation's fixed ordering. Consequently $\sum_\ell r_\ell = 2L$. Half of the budget already guarantees full-layer placement before scores enter the allocation.

Table 9 uses the full-layer aggregation defined in App. B. Each score assigns a rank differing by at most one from uniform rank 2 on at least $L - 1$ layers. Contraction has a depth coefficient of variation 0.405 before allocation, and its allocated ranks range from 1 to 3. Fisher produces one larger early-layer allocation of 10, reflecting its squared-gradient peak.

Under this rule the contraction allocation is broadly distributed, but it is not exactly uniform. The result does not establish equal scores, absence of a ranking, or optimality of full-depth training. The top-k policy is a different allocation rule, giving rank 8 to a quarter of the layers and zero to the rest. Training experiments compare placement choices separately from the invariance audit.

H Data, training, and selection rules

This section specifies the depth and placement survey in Appendix I. The fixed-backbone and GSM8K experiments use their separately stated training protocols (Appendices M and P).

Data and calibration. Training uses a fixed subset of the Commonsense170K instruction mixture (Hu et al., 2023), with $13{,}500$ training examples and $1{,}500$ validation examples from the shared seeded split. Evaluation uses the complete BoolQ, PIQA, HellaSwag, WinoGrande, ARC-Easy, ARC-Challenge, and OpenBookQA evaluation files from LLM-Adapters, totaling $20{,}465$ examples. These are not all official test splits; BoolQ uses its validation split. The evaluator scores fixed answer-label continuations by mean token log probability with task-specific formats. It retains at most the final 480 prompt tokens rather than generating free-form answers or scoring full answer texts. CS-Avg is the unweighted mean of the seven task accuracies.

Each model's selectors use the same 32 source examples, given by the prefix of the original training-subset index list. This prefix is not the first 32 examples of the final training split, and it is not

asserted to be disjoint from training. Gradient, Fisher, and contraction truncate calibration text to 256 tokens. Deployed Read uses a maximum of 128. All four additional statistical controls use the training maximum of 384 tokens for calibration. Examples shorter than eight tokens are skipped, and each estimator averages over the examples it actually uses. Selected sets and source indices are stored in criteria artifacts that training jobs validate at startup.

Shared training configuration. All arms on a given model use AdamW with learning rate 2×10^{-4}, a cosine schedule, warmup ratio 0.03, batch size 8, gradient accumulation 2, and one epoch. The maximum training sequence length is 384, gradient clipping is 1.0, and LoRA dropout is 0.05. Setting $\alpha = 2r$ keeps the adapter multiplier α/r constant. Adapters attach to all seven projection matrices in each selected block. Arms share Trainer seed labels and data order, and evaluation uses the same example catalog. Adapter creation precedes setting these seed labels, so the labels do not certify controlled adapter initialization, even when ranks and selected sets coincide.

Parameter budgets. For k selected layers of a model with depth L, the intended rank is $2L/k$, rounded to an integer. Because module shapes are constant across layers, the adapter parameter count is

$$P(k, r) = kr \sum_{W \in \mathcal{W}_\ell} \big(d_{\text{in}}(W) + d_{\text{out}}(W)\big).$$

Full, half, and quarter placements use ranks $2, 4, 8$ and match exactly. On Qwen2.5-0.5B, the eighth arm also matches at rank 16, with $1,099,776$ trainable parameters.

The label *eighth* denotes the sparsest geometry arm. On Llama-3.2-3B it selects 3 of 28 layers at rank 19, giving $3,093,504$ parameters against $3,039,232$ at full depth. This is 1.79% more parameters, within a tolerance of one rank unit per layer. On Qwen3-4B it selects 4 of 36 layers at rank 18 and matches the full-depth parameter count. The trend analysis uses these realized fractions.

Placement arms. The geometry arms use full depth and uniformly spaced half, quarter, and nominal eighth sets. Contiguous controls place the same quarter-layer budget in early, middle, or late blocks. Seven calibrated selectors use $k = L/4$, including Gradient, Read, contraction, and the four rules below. The three random quarter sets per model each use one training seed and supply a descriptive range. That range measures variation among those sets, rather than a confidence interval over training seeds. The uniform-quarter versus contiguous comparison holds k, rank, and parameter count fixed. The geometry ladder changes both the number of adapted layers and the rank per layer.

Additional statistical controls. Table 10 names each control by its implemented statistic. These four controls measure block sensitivity, gradient energy, linear-readout fit, and gradient magnitude standardized across examples. They are computed on the frozen backbone and then used to select layers for the common training procedure. The following definitions specify the statistic and calibration protocol for each control.

For *input-perturbation sensitivity*, let $B_\ell(H)$ be the block output for its input hidden-state tensor H. With one Gaussian perturbation $\delta_{\ell,n}$ per example and layer, whose entrywise standard deviation is 0.001, the score is

$$s_\ell^{\text{pert}} = \frac{1}{N} \sum_n \frac{\|B_\ell(H_{\ell,n} + \delta_{\ell,n}) - B_\ell(H_{\ell,n})\|_F}{\|\delta_{\ell,n}\|_F}.$$

This measures the finite-difference response at the block boundary. It is distinct from Layer Card's projected-residual, activation-energy, and coupling diagnostics (Xu et al., 2026).

For the *empirical Fisher trace*, the rule is

$$s_\ell^{\text{Fisher}} = \frac{1}{N} \sum_n \|\nabla_{\theta_\ell} \mathcal{L}_n\|_2^2,$$

including all block parameters with gradients. Subsequent training uses the common AdamW configuration; this score selects the same layer sets as Gradient on all reported models.

For *in-sample probe accuracy*, collect each retained token's block-output feature $h_{\ell,j}$ and next-token target y_j. Fit a bias-free vocabulary-sized linear probe U_ℓ using 24 full-batch SGD steps with learning

rate 0.05 on

$$\mathcal{L}_{\text{probe}}(U_\ell) = -\frac{1}{M}\sum_{j=1}^{M} \log \text{softmax}(U_\ell h_{\ell,j})_{y_j}.$$

The layer score is its accuracy on the same tokens used to fit the probe,

$$s_\ell^{\text{probe}} = \frac{1}{M}\sum_j \mathbf{1}\{\arg\max_t (U_\ell h_{\ell,j})_t = y_j\}.$$

The statistic measures how well a linear readout fits the calibration tokens. RSRA instead bases rank allocation on spectral effective rank, virtual low-rank perturbations, and Fréchet representation displacement (Liu et al., 2026).

For *standardized gradient magnitude*, let $v_{\ell,n} = \|\nabla_{\theta_\ell}\mathcal{L}_n\|_2$, $\bar{v}_\ell = N^{-1}\sum_n v_{\ell,n}$, and $\widehat{\sigma}_\ell^2 = (N-1)^{-1}\sum_n (v_{\ell,n} - \bar{v}_\ell)^2$. The score is the one-sample statistic against zero

$$s_\ell^{\text{std}} = \frac{\bar{v}_\ell}{\widehat{\sigma}_\ell/\sqrt{N}}.$$

The implementation falls back to $\bar{v}_\ell$ when there is a single sample or the statistic is nonfinite, and uses zero if there are no usable samples. It ranks layers by this statistic and does not use its p-value for selection. This frozen-gradient statistic differs from StatLoRA's uncertainty estimates over component-score trajectories during training, including historical means and covariance (Gao & Tan, 2026).

I Depth and placement survey

This depth survey spans Qwen2.5-0.5B, Llama-3.2-3B, and Qwen3-4B, with score-free layer geometries, selected quarter-depth sets, and contiguous controls under the shared LoRA parameter budget. Table 10 reports the complete seven-task CS-Avg results. Its five run labels control data order and Trainer randomness, describing observed multi-run variation across models. The three random-quarter sets are each evaluated at a single common training seed.

The sparse-uniform arm uses $(k, L, r) = (3, 24, 16)$, $(3, 28, 19)$, and $(4, 36, 18)$ for Qwen2.5-0.5B, Llama-3.2-3B, and Qwen3-4B, respectively. Llama rank rounding gives $3{,}093{,}504$ trainable parameters versus $3{,}039{,}232$ for full-depth ($+1.79\%$).

In the recorded means, full-depth placement leads the scored arms on all three models. At fixed quarter-layer count and rank, uniform placement exceeds the mean of the three contiguous alternatives on each model; Figure 7 shows the underlying run spread. The broad survey describes these placement patterns and task scores. Appendix J defines its paired descriptive contrasts, while the fixed-backbone experiment in Appendix M uses a controlled initialization procedure for its own matched-seed comparison.

J Paired comparisons and uncertainty reporting

Best-contiguous descriptive comparison. The supplementary descriptive contrast compares uniform-quarter mean accuracy with the largest of the early, middle, and late five-run means. The differences are 0.86346, 6.01319, and 1.28299 percentage points on Qwen2.5-0.5B, Llama-3.2-3B, and Qwen3-4B, respectively. The selected reference blocks are early, middle, and middle. The maximum is taken after averaging the five runs for each block. This defines a descriptive comparison with the strongest observed contiguous placement.

For arms a, b on one model, let $\Delta_s = A_a(s) - A_b(s)$, where $A_x(s)$ is the macro-average accuracy for run label s on the fixed evaluation files. We report the mean and sample standard deviation of five paired differences. These summarize the recorded runs, not a confidence interval for new models or evaluation populations.

There are five contrasts per model. Three compare full depth with Gradient, Read, and contraction. The fourth compares uniform quarter placement with the mean of the three contiguous placements.

Table 10: **Depth and placement survey across models.** CS-Avg (%) over five runs (mean ± sample SD); random sets report the range at seed 42. Parameter counts are matched across arms within each model.

	Qwen2.5-0.5B	Llama-3.2-3B	Qwen3-4B
Score-free allocations			
Full-depth (all L, r=2)	**53.81 ±0.52**	**67.75 ±1.00**	**82.39 ±0.35**
Uniform 1/2 (r=4)	53.37 ±0.27	66.42 ±1.01	82.27 ±0.19
Uniform 1/4 (r=8)	53.04 ±0.33	66.56 ±1.46	82.11 ±0.30
Sparse uniform	52.89 ±0.36	60.53 ±1.00	81.33 ±0.25
Random quarter sets (3, seed 42)	[50.8, 52.6]	[61.4, 66.9]	[81.3, 81.5]
Audited selectors (quarter of layers at matched budget)			
Gradient top-k	51.97 ±0.39	61.37 ±0.96	82.08 ±0.14
Read top-k	52.39 ±0.46	63.13 ±1.38	82.23 ±0.09
Contraction $\langle W, G \rangle$ top-k	52.62 ±0.59	63.98 ±1.11	82.17 ±0.26
Additional statistical controls (quarter of layers)			
Input-perturbation sensitivity	50.89 ±0.36	55.64 ±1.46	79.93 ±0.26
Empirical Fisher trace	52.01 ±0.25	61.85 ±2.08	82.00 ±0.23
In-sample probe accuracy	43.40 ±0.78	53.50 ±0.37	78.63 ±0.11
Standardized gradient magnitude	51.85 ±0.68	53.13 ±0.57	78.85 ±0.08
Contiguous placements (quarter of layers)			
Early block	52.18 ±0.54	59.92 ±1.67	79.78 ±0.41
Middle block	52.10 ±0.53	60.55 ±0.74	80.83 ±0.13
Late block	43.48 ±0.48	53.38 ±0.37	77.89 ±0.11

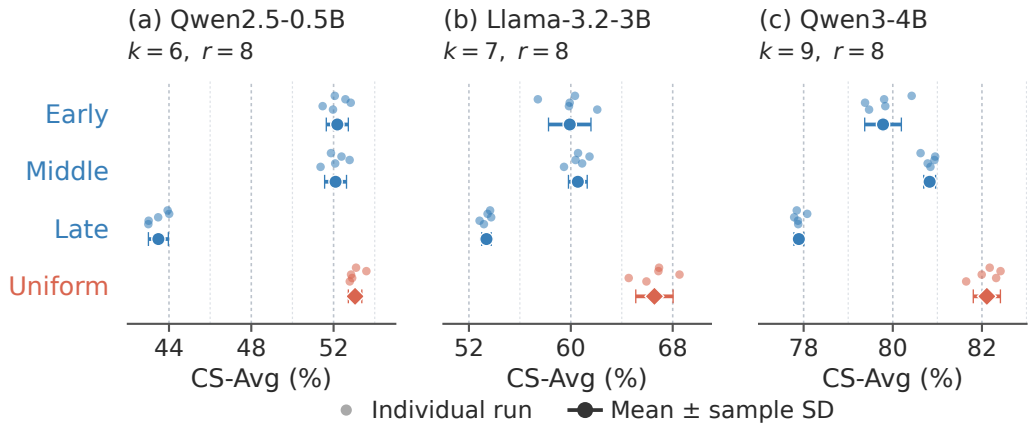


Figure 7: **Original survey of quarter-depth placement geometry.** Each panel compares early, middle, late, and uniformly spaced quarter-layer sets on one model. Large markers and horizontal bars show five-run means and sample standard deviations; small points show each run. Uniform placement is coral. Layer count, rank, and parameter count are equal within each panel. These are descriptive run summaries from the depth survey.

For the uniform-allocation sequence, let $x_j = k_j/L$. Its per-run slope is

$$\widehat{\beta}_s = \frac{\sum_j (x_j - \bar{x})(A_j(s) - \bar{A}_s)}{\sum_j (x_j - \bar{x})^2}.$$

The fifth contrast is the average slope, in percentage points per unit adapted-layer fraction. It is not an endpoint difference and does not require every adjacent mean to increase.

Table 11: **Paired effects across recorded training runs.** Mean and sample standard deviation of five run-level differences. Effects are in percentage points; the coverage slope is in percentage points per unit adapted-layer fraction. These are descriptive summaries.

Contrast	Qwen2.5-0.5B	Llama-3.2-3B	Qwen3-4B
Full depth − Gradient	$+1.84 \pm 0.38$	$+6.38 \pm 0.78$	$+0.30 \pm 0.29$
Full depth − Read	$+1.42 \pm 0.38$	$+4.62 \pm 1.46$	$+0.15 \pm 0.33$
Full depth − Contraction	$+1.19 \pm 0.42$	$+3.77 \pm 1.30$	$+0.21 \pm 0.36$
Uniform $1/4$ − block mean	$+3.79 \pm 0.37$	$+8.62 \pm 1.49$	$+2.61 \pm 0.23$
Coverage slope	$+1.04 \pm 0.43$	$+6.02 \pm 1.32$	$+0.94 \pm 0.31$

Uncertainty reporting. All arms use the same evaluation items. The reported sample standard deviation captures variation across the five recorded runs on those fixed files. Paired differences preserve the run labels across arms. Figure 7 displays the five run accuracies for each placement, alongside its mean and sample standard deviation. These descriptive summaries support comparisons of observed effect sizes; population-level confidence intervals and significance tests are not reported.

Scope of the depth survey. The depth survey contains 210 runs in total, together with five paired contrasts per model and additional single-seed random controls. The separate coordinate-matched comparison in Appendix L contains 36 runs, and the fixed-backbone and GSM8K transfer experiments are reported under their own protocols in Appendices M and P.

K Costs and calibration robustness

Table 12 reports wall-clock time and peak memory on a single H100 for Qwen2.5-0.5B. Gradient and contraction require a backward scoring pass, while Read requires repeated forward passes with block ablations. In the recorded measurement, Gradient calibration takes 4 s and Read takes 23 s. The in-sample probe control fits linear readouts, and input-perturbation sensitivity uses perturbed forward sweeps. The latter can take approximately 18 minutes on the largest measured model.

Full-depth training takes 274 s, compared with 167 s for Gradient top-k, plus its calibration pass. The resulting calibration-plus-training times are 274 and 171 s, a ratio of about 1.60, with a peak-memory difference below 2 GB. Together with Table 10, these measurements describe the trade-off between accuracy and runtime under an equal trainable-parameter budget.

The calibration-size check was performed on Qwen2.5-0.5B. Increasing the sample from 32 to 128 examples gives top-k Jaccard 1.00 for Gradient, 1.00 for Read, and 0.71 for Fisher. Gradient and Read select the same sets in this comparison. This supports stability for the measured sample-size change on this model. It does not eliminate calibration uncertainty on other draws or models, or change the coordinate-dependence result.

Calibration draws and coordinate changes. An additional 5×5 study on Llama-3.2-3B crosses five independently drawn calibration batches of 32 examples with five certified gauges at bound $b = 2$. Each configuration produces a vector of 28 layer scores. Table 13 decomposes the observed sum of squares over the resulting $5 \times 5 \times 28$ score array. The coordinate factor contributes 82.4%, and the calibration-batch factor contributes 4.1%. These percentages describe observed score variation rather than estimated population variance components.

For the main-effect tests, gauges and batches are treated as crossed random factors. The analysis compares each main-effect mean square with the gauge-by-batch interaction mean square. Equivalently, both mean squares can be computed on the 25 cell means: averaging over layers scales their sums of squares by the same factor and leaves their ratio unchanged. The denominator has 16 degrees of freedom; the 28 layer scores are not counted as independent experimental replications for these tests. With $MS_{G\times B} = 12.70/16 = 0.79375$,

$$F_G = \frac{148.65/4}{12.70/16} = 46.819, \qquad F_B = \frac{7.39/4}{12.70/16} = 2.328.$$

Table 12: **Cost accounting on Qwen2.5-0.5B.** Runs use one H100 and equal trainable parameter counts. Calibration is a one-time scoring cost. Training time covers one epoch. The table reports these costs separately.

Arm	Params	Calib. (s)	Train (s)	Peak mem (GB)
Full-depth (all L, r=2)	1,099,776	0	274	11.4
Gradient top-k	1,099,776	4	167	9.8
Read top-k	1,099,776	23	170	9.8
Contraction $\langle W, G\rangle$ top-k	1,099,776	4	125	9.8

Using the $F_{4,16}$ reference gives $p = 1.233 \times 10^{-8}$ for coordinates and $p = 0.101$ for batches. The coordinate effect therefore remains pronounced relative to the interaction between the two sampled factors. The primary orbit audit independently fixes the calibration examples within each comparison, measuring recommendation changes on identical data.

The reported descriptive mean pairwise selection overlaps are 0.88 across calibration draws at fixed coordinates and 0.33 across coordinate changes on shared data. The score decomposition and the fixed-data orbit comparisons provide the quantitative calibration controls used in the main text.

Infrastructure and evaluation. Experiments use a single node with H100-class GPUs. A Qwen2.5-0.5B run takes approximately five minutes for training and five minutes for evaluation on the full catalog. Orbit certification and score recomputation use CPU or single-GPU jobs.

L Coordinate-matched placement experiments

For Qwen2.5-0.5B and Llama-3.2-3B, the coordinate-matched experiment crosses two backbone representations with two Gradient-selected sets. The representations are the original checkpoint θ and a certified $g\theta$ obtained using per-head V/O transformations with $\kappa = 4$ and a diagonal MLP rescaling. The sets are $S(\theta)$ and $S(g\theta)$. Each of the four combinations is trained, and full-depth training is evaluated on both representations as a reference (Table 14).

Full-depth means on the original and transformed representations are 54.04 and 54.44 on Qwen2.5-0.5B, and 68.03 and 68.82 on Llama-3.2-3B. These are similar observed means for two representations, not an equivalence test. The set of matrices of rank at most r is closed under invertible left and right multiplication. Nevertheless, adapter initialization, AdamW updates, regularization, and stochastic training are not generally equivariant under these transformations. Equality of the frozen functions therefore does not imply identical training trajectories.

Across the selector cells, the maximum reported shift is 0.7 pp on Qwen2.5-0.5B and 1.9 pp on Llama-3.2-3B. On Llama, the two sets share 5 of 7 members and both concentrate in early layers. Every selector cell remains below full depth on its own backbone by at least 1.7 and 4.3 pp, respectively. While individual cells exhibit moderate score shifts (up to 1.9 pp on Llama), full depth outperforms every selective combination on both representations in this comparison.

M Utility of gauge-derived selections on a fixed backbone

The coordinate-matched study above changes both the representation used for selection and, in some cells, the backbone used for training. Here we isolate the selection decision. For each of Llama-3.2-3B and Qwen3-4B, we score the original checkpoint and four exactly equivalent checkpoints obtained by independently seeded per-layer SwiGLU scalar gauges with $|\log_2 c_\ell| \le 2$. The same 32 commonsense calibration examples and full-block Gradient rule produce five quarter-depth sets per model. Each of those sets is then trained *on the original checkpoint*; no transformed weights enter adaptation. The gauges passed deterministic float64 forward checks before selection. Table 2 lists every frozen set, including the selected layer IDs and its Jaccard overlap with the original selection.

Additional random placements. Four random layer sets per model provide references beyond the gauge-derived sets (Table 15). Llama's rejection-sampling rule requires $\max \ell \le 17$ and Jaccard overlap in $[0.25, 0.40]$ with the canonical Gradient set. The final four sets have overlaps 0.400, 0.400,

Table 13: **Coordinate and calibration effects in the crossed study.** Observed score sums of squares (SS) for five gauges, five batches, and 28 layers. Main-effect $F_{4,16}$ tests use the interaction mean square. Shares describe the observed SS; they are not random-effect variance estimates.

Source	SS	Share (%)	df	F	p
Gauge	148.65	82.4	4	46.82	1.23×10^{-8}
Calibration batch	7.39	4.1	4	2.33	0.101
Gauge $\times$ batch	12.70	7.0	16	—	—
Within-cell remainder	11.67	6.5	675	—	—
Total	180.41	100.0	699	—	—

Table 14: **Coordinate-matched 2×2 comparison.** CS-Avg is reported as mean $\pm$ standard deviation over three training seeds. Rows identify the backbone representation and the placed set. Full-depth references are trained on both representations. The results concern these two representations and these selected sets.

Backbone	LoRA placement	Qwen2.5-0.5B	Llama-3.2-3B
θ (canonical)	full-depth	54.04 $\pm$0.24	68.03 $\pm$0.69
$g\theta$ (gauged)	full-depth	54.44 $\pm$0.48	68.82 $\pm$0.77
θ	$S(\theta)$ Gradient top-k	51.54 $\pm$0.40	62.64 $\pm$1.79
θ	$S(g\theta)$ Gradient top-k	52.23 $\pm$0.27	63.60 $\pm$1.29
$g\theta$	$S(\theta)$ Gradient top-k	52.15 $\pm$0.14	64.54 $\pm$0.80
$g\theta$	$S(g\theta)$ Gradient top-k	52.76 $\pm$0.35	64.48 $\pm$1.52

0.273, and 0.400, averaging 0.368. Their three-seed mean accuracies span 65.42–65.81%, compared with 63.83% for canonical Gradient. Table 16 records all runs and the seed-paired changes; every pair favors the random control. Averaging the four sets within each training seed gives an accuracy of $65.59 \pm 0.94\%$ and a paired gain of 1.757 ± 0.586 pp across the three seeds. This aggregation keeps the training seed as the uncertainty unit while averaging over the four fixed sets.

Across the twelve individual random-set runs, the descriptive mean and sample SD are $65.59 \pm 0.82\%$; the corresponding twelve paired differences are 1.757 ± 0.525 pp. These pooled dispersions describe the observed runs, not twelve independent replications of the shared canonical baseline. The four Qwen sets have means of 81.99–82.11% and overlaps 0.200, 0.059, 0.200, and 0.286, averaging 0.186. Their layer IDs define the comparison directly. The Llama depth cap and overlap interval control the region and amount of replacement, while preserving multiple placement geometries within that region.

All sets within a model use the same seven LoRA target projections, rank 8, training data, evaluation suite, and seeds 42–44. The trainer resets the seed immediately before constructing PEFT adapters. For each gauge-derived set, its accuracy is paired with the original-set result at the same training seed. Table 2 reports means and sample standard deviations over these three seeds, with evaluation items held fixed. The four Llama gauge-derived means exceed the original mean, with paired differences of $+1.508$ to $+3.889$ pp. The Qwen gauge-derived means span 81.94–82.12%, compared with 82.10% for the original selection. These results evaluate the utility of the changed recommendations separately from the size of the selection change.

N PER-TASK RESULTS AND TASK-EXCLUSION SUMMARIES

Table 17 reports the mean full-depth minus Gradient difference on each task. On Qwen2.5-0.5B, six of seven task means favor full depth, and on Llama-3.2-3B, all seven tasks favor full depth. On Qwen3-4B, five of seven task directions favor full depth, yielding an overall positive macro-average difference.

Table 17 summarizes the original five-run survey under its answer-label evaluation rule. Avg-7 includes all seven tasks; Avg-6 excludes HellaSwag and averages the remaining six. The reported Full-depth minus Gradient differences on Avg-6 are $+1.85$, $+3.55$, and $+0.28$ pp on Qwen2.5-0.5B,

Table 15: **Additional random placements on the original backbone.** Zero-based layer IDs, Jaccard overlap J with the original Gradient set, and three-seed CS-Avg means. Layer count and rank match the gauge-derived arms within each model. Table 16 gives the individual Llama runs.

Set	Selected layer IDs	J	CS-Avg (%)
Llama-3.2-3B: seven layers, rank eight			
S_1	$\{0, 2, 4, 7, 9, 13, 16\}$	0.400	65.58
S_2	$\{0, 1, 5, 8, 10, 14, 17\}$	0.400	65.42
S_3	$\{1, 2, 4, 6, 9, 12, 15\}$	0.273	65.81
S_4	$\{0, 2, 5, 7, 11, 13, 17\}$	0.400	65.55
Qwen3-4B: nine layers, rank eight			
S_1	$\{4, 7, 12, 16, 19, 22, 26, 31, 34\}$	0.200	82.04
S_2	$\{2, 8, 13, 17, 20, 23, 27, 30, 35\}$	0.059	82.02
S_3	$\{5, 9, 11, 15, 18, 21, 25, 29, 33\}$	0.200	82.11
S_4	$\{3, 6, 10, 14, 19, 22, 28, 32, 35\}$	0.286	81.99

Table 16: **Individual runs for the random-placement control.** Llama-3.2-3B CS-Avg (%) at rank eight and learning rate 2×10^{-4}. Summaries are means and sample SDs over seeds 42–44; paired Δ subtracts Canonical at the same seed.

	Training seed				
Selection	42	43	44	Mean $\pm$ SD (%)	Paired Δ (pp)
Canonical	63.03	63.56	64.91	63.83 ± 0.97	–
S_1	65.42	64.68	66.64	65.58 ± 0.99	$+1.747 \pm 0.635$
S_2	65.18	64.71	66.37	65.42 ± 0.86	$+1.587 \pm 0.512$
S_3	65.68	64.89	66.86	65.81 ± 0.99	$+1.977 \pm 0.660$
S_4	65.31	64.76	66.58	65.55 ± 0.93	$+1.717 \pm 0.542$

Table 17: **Mean full depth minus Gradient by task.** Five-run mean differences (pp) across seven tasks; Avg-6 excludes HellaSwag (HS). WG denotes WinoGrande and OBQA denotes OpenBookQA.

Model	ARC-C	ARC-E	BoolQ	HS	OBQA	PIQA	WG	**Avg-7**	**Avg-6**
Qwen2.5-0.5B	$+2.68$	$+4.12$	$+1.08$	$+1.83$	$+1.76$	$+1.59$	-0.16	**+1.84**	**+1.85**
Llama-3.2-3B	$+3.74$	$+1.63$	$+3.30$	$+23.38$	$+1.76$	$+3.29$	$+7.55$	**+6.38**	**+3.55**
Qwen3-4B	$+0.63$	$+0.38$	-0.72	$+0.44$	$+0.64$	-0.32	$+1.06$	**+0.30**	**+0.28**

Llama-3.2-3B, and Qwen3-4B. Thus the positive survey-average difference persists when that one task is omitted. These summaries retain the answer-label protocol for the remaining tasks. Appendix R separately documents a complete-ending HellaSwag evaluator.

O SELECTION STABILITY UNDER MATHEMATICS CALIBRATION

For Llama-3.2-3B, we additionally examine Gradient selections under mathematics calibration data, applying the same function-preserving gauge transformations as the main audit. Under mathematics calibration, the selected layer sets exhibit pronounced sensitivity to coordinate changes across depth regions (overlap $J = 0.167$, replacing 5 of 7 layers, as illustrated in Figure 4).

The downstream GSM8K generation experiment in Appendix P evaluates the previously established commonsense-selected layer sets on mathematics generation, testing the utility of the fixed selections on a second task without recalibration.

Table 18: **GSM8K transfer of frozen commonsense-selected layers.** All fifteen runs use the same first-answer extraction rule on 1,319 questions. Δ subtracts Original at the same training seed.

Seed	Selection	Acc. (%)	Δ (pp)
42	Original	33.66	–
42	Gauge 18000	34.12	+0.45
42	Gauge 18001	34.50	+0.83
42	Gauge 18002	33.43	-0.23
42	Gauge 18003	31.84	-1.82
43	Original	31.77	–
43	Gauge 18000	33.81	+2.05
43	Gauge 18001	33.97	+2.20
43	Gauge 18002	32.75	+0.99
43	Gauge 18003	31.01	-0.76
44	Original	31.99	–
44	Gauge 18000	33.06	+1.06
44	Gauge 18001	34.19	+2.20
44	Gauge 18002	31.69	-0.30
44	Gauge 18003	31.08	-0.91

P GSM8K GENERATION WITH TRANSFERRED LAYER SETS

This experiment uses exactly the five Llama-3.2-3B layer sets from Appendix M: the original Gradient set and four sets selected after the certified scalar gauges. Their layer identities are fixed from commonsense calibration before this task; no GSM8K example is used to recompute scores or choose an arm. All five adapters train from the same original backbone, at rank 8 and learning rate 2×10^{-4} for one epoch, with seeds 42, 43, and 44. The shared split has $6{,}725$ training and 748 validation examples. Evaluation uses the same $1{,}319$ GSM8K test questions for every arm.

All fifteen evaluations apply the same first completed-answer #### extraction rule to the original generated responses. For fourteen runs, the first and last answer markers agree. Gauge 18003 at seed 44 contains later generated markers; using the common first-answer rule gives 31.08%. Table 18 and Figure 3 therefore report all five sets under the same evaluator, with training seed as the unit of pairing and uncertainty.

Gauge 18000 and 18001 improve accuracy at every seed, with mean paired gains of 1.188 and 1.744 pp. Gauge 18002 has the largest commonsense gain and a GSM8K mean change of 0.152 pp. Gauge 18003 reaches $31.312 \pm 0.461\%$, with a paired change of -1.162 ± 0.574 pp. Thus a fixed selection's utility ordering depends on the downstream task, even when its layer identities and evaluation protocol are held fixed.

Q ETHICS AND REPRODUCIBILITY

Ethics. This work studies public checkpoints and benchmarks without human participants or collection of personal data. Its focus is the allocation of adaptation parameters. Training and calibration costs are reported separately under their recorded measurement conditions.

Reproducibility. Score definitions, proofs, and protocols appear in the appendix. The source package includes figure inputs, numerical checks, and separately identified historical and revised records.

R ADDITIONAL PROTOCOL AND BASELINE CHECKS

Initialization and shared placement settings. The controlled comparisons in Tables 3 and 4 train all six methods on the same approximately $10{,}000$-example Commonsense170K subset, with a common batch schedule and exactly 1,000 optimization steps per run. The fixed-rate comparison uses seeds 42–44 and a learning rate of 2×10^{-4}. Standard LoRA placement experiments set initialization seeds before constructing each PEFT adapter. Rebuilding with seed 42 reproduces the initial-adapter hash, while seed 43 changes it, in the Full-depth and Uniform-quarter checks on both models. This checks repeatability within each configuration. The earlier five-run survey and its training split are described separately in Appendices H and I.

Scoring scope and calibration. Gradient aggregates whole-block parameter gradients, whereas LoRA trains seven projection types. Recomputing Gradient on the same 32 calibration examples using only projection weights selects identical quarter-depth sets ($7/7$ on Llama and $9/9$ on Qwen; Jaccard 1.0). Increasing calibration size from 32 to 128 gives Gradient/Fisher/Read/Contraction top-k Jaccards of $1.00/0.75/0.75/0.56$ on Llama and $1.00/1.00/0.80/0.80$ on Qwen. These are separate checks of scoring scope and sample size.

Learning-rate selection and the additional candidate. The comparison covers two models, two matched final effective budgets, and six methods, giving 24 model–budget–method choices. Each uses seed 45 to select the minimum validation loss among $\{1, 2, 4\} \times 10^{-4}$; all choose 4×10^{-4}. Table 4 gives the independent seed-42 final evaluations. Standard Full-depth ranks are 2 and 8, while quarter-depth ranks are 8 and 32. On Llama at budget $2L$, an additional three-seed comparison gives $72.20 \pm 0.29\%$ for Full-depth and $70.26 \pm 0.50\%$ for Contraction, a mean difference of 1.94 pp. Full-depth scores at seeds 42–44 are 72.31, 71.88, and 72.42%; the paired Contraction scores are 70.84, 69.93, and 70.02%. The two configurations retain the same ordering at each seed.

At Llama budget $2L$, increasing the learning rate from 4×10^{-4} to 8×10^{-4} changes validation losses from $1.35/1.36/1.37/1.36/1.36/1.37$ to $1.54/1.55/1.58/1.56/1.57/1.59$, in the order Full-depth, Uniform, Gradient, Contraction, AdaLoRA, and FoRA. Thus the selected rate has lower validation loss for all six arms in this upper-grid check. Standard LoRA placements use $\alpha = 2r$, keeping the adapter multiplier $\alpha/r = 2$ at every tested rank. AdaLoRA sets $\alpha/r_{\text{target}} = 2$, and FoRA uses $\alpha/r = 2$. These settings and the common validation search specify the scaling and learning-rate controls used in the comparison.

Additional adaptation algorithms and final budgets. Tables 3 and 4 compare AdaLoRA (Zhang et al., 2023) and FoRA (Park et al., 2026) with the four static LoRA placements. All six methods have the same final effective adapter parameter count within each model and budget (Table 19). For retained factors on the seven target projections, this count is

$$P_{\text{eff}} = \sum_{\ell,p} r_{\ell,p}\big(d^{\text{in}}_{\ell,p} + d^{\text{out}}_{\ell,p}\big).$$

For AdaLoRA, we absorb the learned diagonal singular values into one retained factor before evaluation: $B = P_U \operatorname{diag}(E)$ and $A = P_V$. We count this final A/B representation. The singular-value parameters are therefore included through the resulting factors rather than counted as a separate evaluation tensor. Training-time parameter counts, optimizer states, and compute are separate resources.

- **AdaLoRA.** Weight–gradient sensitivity controls pruning of an SVD-parameterized update. At budget $2L$, the implementation starts with rank eight and prunes during steps 200–800 of the common 1,000-step run toward an average effective rank of two. On Llama, all 196 projection matrices retain ranks between one and four, with total retained rank 392; all 28 layers remain active. Qwen likewise retains nonzero adapters in all 36 layers. At budget $8L$, the initial rank is 16 and the target average rank is eight. The larger initial allocation affects training-time cost, while the final effective representation matches the counts in Table 19.
- **FoRA.** Fisher information selects the layers; the Llama comparison adapts seven of 28 layers. Factor B satisfies the Stiefel constraint $B^\top B = I_r$. Cayley-Adam updates B, AdamW updates A, and QR re-orthogonalization is applied every 100 steps. The recorded orthogonality residual remains below 10^{-5}.

Full-depth has the highest observed accuracy in every model–budget setting. Its advantage over AdaLoRA also holds when both allocations retain every layer, distinguishing uniform rank allocation from adaptive rank redistribution at full depth.

Table 19: **Matched final effective adapter budgets.** Each count is shared by Full-depth, Uniform-quarter, Gradient, Contraction, AdaLoRA, and FoRA. Counts include all seven target projections; AdaLoRA's singular values are absorbed into its retained factors before counting the final representation.

Model	Budget	Final effective parameters	Training steps
Llama-3.2-3B	$2L$	3,039,232	1,000
	$8L$	12,156,928	1,000
Qwen3-4B	$2L$	4,128,768	1,000
	$8L$	16,515,072	1,000

Complete-ending HellaSwag scoring. A retained scoring-format check evaluates all $10{,}042$ HellaSwag questions and $40{,}168$ nonempty endings. It scores the complete ending after a common answer prefix by either mean or summed token log probability, using a response instruction different from the answer-label protocol. For the seed-42, $2L$, validation-selected checkpoints, Full-depth/Gradient answer-label accuracy is $82.39/79.93\%$ on Llama and $89.08/88.77\%$ on Qwen. Complete-ending token-mean scores are $26.31/26.87\%$ and $33.31/33.40\%$, and token-sum scores are $28.65/29.02\%$ and $42.50/45.90\%$, respectively. These results specify a separate evaluator for one task; the reported seven-task means use the answer-label rule. The Avg-6 summary in Appendix N changes task inclusion while retaining that rule.

Table 20: **Norm balancing standardizes layer scores across scalar gauges.** Llama-3.2-3B, with the same 32 calibration examples. CV is the sample SD divided by the mean across 28 layers. All five balanced representations select $\{0, 4, 8, 12, 16, 20, 24\}$ by direct top-seven ranking; J_{bal} compares each selection with the balanced original. The shared set matches Uniform-quarter and scores $66.22 \pm 0.74\%$ over seeds 42–44 on the original backbone (rank eight, 2×10^{-4}, 1,000 steps).

Representation	Original top-7	CV before	CV after	J_{bal}
θ	$\{0, 1, 2, 4, 5, 7, 10\}$	0.58	0.08	1.000
$g_0\theta$	$\{0, 1, 3, 4, 7, 11, 17\}$	0.52	0.08	1.000
$g_1\theta$	$\{0, 1, 3, 4, 5, 9, 12\}$	0.49	0.08	1.000
$g_2\theta$	$\{0, 1, 4, 6, 8, 9, 13\}$	0.54	0.08	1.000
$g_3\theta$	$\{0, 1, 4, 6, 12, 14, 15\}$	0.51	0.08	1.000

Measured checkpoint scales. For nonzero MLP factors, define channel ratios $\rho_i = \|W_{\mathrm{up}}[i,:]\|_2 / \|W_{\mathrm{down}}[:,i]\|_2$. The magnitude of the corresponding balancing scale is $|\log_2 c_i| = \frac{1}{2}|\log_2 \rho_i|$. The reported Llama median, 90th percentile, and maximum are 0.72, 1.68, and 2.54; the Qwen values are 0.65, 1.52, and 2.38. The Llama 90th percentile corresponds to a scale magnitude of about 3.2 and a reciprocal-factor norm disparity of about 10.3. These summaries place the bounded interventions beside measured within-checkpoint scale differences. In the fixed scalar path, Llama's overlap changes at $b = 0.5$, and Qwen's changes at $b = 1$ (Appendix D.1). Channel imbalance and a layerwise scalar intervention are distinct quantities; the comparison concerns their numerical scale.

Norm balancing and its scope. For a nonzero MLP up row u_i and corresponding down column d_i, choose

$$c_i = \sqrt{\frac{\|d_i\|_2}{\|u_i\|_2}}, \qquad u_i' = c_i u_i, \qquad d_i' = d_i / c_i.$$

This is a function-preserving diagonal gauge and equalizes the paired norms. Both norms must be nonzero for the formula to apply. For grouped-query attention, one consistent head-level construction groups all output blocks $W_{O,j}$ sharing KV head h. With $o_h = (\sum_{j \in I_h} \|W_{O,j}\|_F^2)^{1/2}$ and $v_h = \|W_{V,h}\|_F$, the common positive factor $s_h = \sqrt{o_h / v_h}$ scales $W_{V,h}$ and inversely scales every $W_{O,j}$

in that group. This construction balances the value norm against the aggregate output norm and preserves the shared-head cancellation.

The MLP construction removes positive channel-scale freedom algebraically. If (u_i, d_i) is first replaced by $(a_i u_i, d_i/a_i)$ with $a_i > 0$, its balancing factor becomes c_i/a_i, yielding the same balanced pair. For the implemented balancing transformation, the float64 forward check gives a logit deviation below 1.2×10^{-5}, and the inverse-map reconstruction error is below 4×10^{-15}.

Table 20 compares the original checkpoint and four positive scalar gauges. The across-layer coefficient of variation falls from 0.49–0.58 to approximately 0.08. The 28 balanced scores are given in Table 21: their mean is 0.9900, sample SD is 0.0800, and SEM is 0.0151. Direct descending-score selection, implemented as `np.argsort(scores)[::-1][:7]`, returns $\{0, 4, 8, 12, 16, 20, 24\}$ without tie-breaking. The score profile has local maxima at these regularly spaced layers; the selected set is the same across all five balanced representations.

The shared set is identical to the static Uniform-quarter allocation in the controlled Llama experiment. Training it once per seed on the original backbone, at rank eight and learning rate 2×10^{-4} under the common 1,000-step protocol, gives 66.37, 65.42, and 66.88% for seeds 42–44, respectively: $66.22 \pm 0.74\%$ (mean $\pm$ sample SD). This is one three-seed evaluation of a shared placement, used for both the balancing and uniform-placement comparisons. The experiment connects normalization of internal scales to a flatter gradient profile, a distributed selection, and its measured adaptation utility.

Table 21: **Balanced gradient scores on Llama-3.2-3B.** Scores use the same 32 calibration examples. Bold entries are the seven largest scores; layer IDs are zero-based. Their across-layer mean is 0.9900, sample SD is 0.0800, and CV is 0.0808.

Layer	Score	Layer	Score	Layer	Score	Layer	Score
0	**1.1422**	1	0.9765	2	0.9792	3	0.9486
4	**1.1300**	5	0.9047	6	0.9318	7	0.9691
8	**1.1046**	9	0.9594	10	0.9213	11	0.9031
12	**1.0893**	13	0.9597	14	0.9824	15	0.9689
16	**1.0983**	17	0.9046	18	0.9130	19	0.9489
20	**1.1232**	21	0.9763	22	0.9426	23	0.9092
24	**1.1412**	25	0.9377	26	0.9733	27	0.9811

S FIVE-SEED PAIRED COMPARISONS

The small absolute margins on Qwen make variation in the *paired difference* especially relevant. Table 22 reports five-seed comparisons at budget $2L$ and learning rate 2×10^{-4}, using seeds 42–46. These summaries are presented separately from the three-seed comparison in Table 3 and are not pooled with it. Each contrast pairs Full-depth and the baseline at the same training seed; the evaluation catalog is fixed.

Statistical unit and calculation. Let $d_s = a_{\mathrm{Full},s} - a_{\mathrm{baseline},s}$ denote the CS-Avg difference in percentage points, and let $\bar{d}$ and s_d be its mean and sample SD across the five seeds. We compute

$$t = \frac{\bar{d}}{s_d/\sqrt{5}}, \qquad \mathrm{CI}_{95\%} = \bar{d} \pm t_{0.975,4}\frac{s_d}{\sqrt{5}}, \qquad t_{0.975,4} = 2.7764. \tag{3}$$

The two-sided paired-t calculation assumes approximately normal seed-level differences. Pairing accounts for variation shared by the two runs at a seed. We apply Holm correction across all nine contrasts in the table; the displayed confidence intervals are individual, rather than multiplicity-adjusted, intervals. Test statistics and intervals use the reported paired means and SDs before rounding the marginal accuracies to two decimals.

Table 22: **Five-seed paired comparisons.** CS-Avg (%) at $2L$ and 2×10^{-4}. Differences are Full-depth minus baseline (pp), with sample SD. Intervals are individual paired-t 95% intervals ($df = 4$); p_{Holm} adjusts all nine two-sided tests.

Baseline	Mean	Paired $\Delta \pm s_\Delta$	t	95% interval	p_{Holm}
Qwen3-4B; Full-depth mean 82.36%					
Gradient	82.08	$+0.278 \pm 0.046$	13.51	[0.221, 0.335]	0.0016
Uniform quarter	82.16	$+0.204 \pm 0.052$	8.77	[0.139, 0.269]	0.0049
AdaLoRA	82.19	$+0.172 \pm 0.061$	6.30	[0.096, 0.248]	0.0094
FoRA	82.12	$+0.240 \pm 0.049$	10.95	[0.179, 0.301]	0.0032
Contraction	82.09	$+0.270 \pm 0.058$	10.41	[0.198, 0.342]	0.0034
Llama-3.2-3B; Full-depth mean 68.21%					
Gradient	63.74	$+4.468 \pm 1.102$	9.07	[3.100, 5.836]	0.0049
Uniform quarter	66.18	$+2.028 \pm 0.612$	7.41	[1.268, 2.788]	0.0071
AdaLoRA	66.65	$+1.558 \pm 0.548$	6.36	[0.878, 2.238]	0.0094
Contraction	65.48	$+2.730 \pm 1.025$	5.96	[1.457, 4.003]	0.0094

Small margins with consistent paired gains. On Qwen, Full-depth improves over the five baselines by 0.172–0.278 pp, with paired SDs of 0.046–0.061 pp. All five seed-level differences are positive for every contrast. The unadjusted two-sided p value for Full-depth versus Gradient is 1.74×10^{-4}; the largest Qwen value is 0.00324, against AdaLoRA. The evidence therefore distinguishes a small average margin from uncertainty about its sign.

On Llama, the paired gains range from 1.558 to 4.468 pp. In particular, the Full-depth–Contraction difference is 2.730 ± 1.025 pp, with a 95% interval of $[1.457, 4.003]$. All nine comparisons remain significant after Holm correction ($p_{\mathrm{Holm}} < 0.01$). These results support the full-depth advantage across training seeds on the fixed seven-task evaluation catalog, while the effect sizes retain the distinction between the larger Llama gains and the smaller Qwen gains.

T DoRA and PiSSA under explicit allocation budgets

Layer placement can be combined with changes to the adapter parameterization or initialization. We compare PiSSA (Meng et al., 2024), which initializes the trainable factors from principal singular components and freezes the residual weight, and DoRA (Liu et al., 2024), which learns weight magnitudes together with low-rank directional updates. These methods do not themselves specify a layer selector. Our quarter-depth configurations use the same evenly spaced layers as Uniform quarter; we also evaluate full-depth DoRA.

Rank budgets and parameter counts. The $2L$ and $8L$ labels specify the low-rank allocation: full-depth ranks are two and eight, and quarter-depth ranks are eight and 32. PiSSA has the same factor dimensions as LoRA at each allocation. Standard DoRA additionally trains one magnitude per output channel. For active projections $\mathcal{A}$, its count is

$$N_{\mathrm{DoRA}} = N_{AB} + \sum_{(\ell,p)\in\mathcal{A}} d^{\mathrm{out}}_{\ell,p}, \qquad N_{AB} = \sum_{(\ell,p)\in\mathcal{A}} r_{\ell,p}(d^{\mathrm{in}}_{\ell,p} + d^{\mathrm{out}}_{\ell,p}). \tag{4}$$

Thus equal low-rank budgets do not imply equal total trainable parameters for DoRA. Table 24 gives counts from the standard parameterizations; its extra magnitude term is unchanged when the rank budget increases. The adaptive methods retain the final-factor accounting of Appendix R.

Table 23: **Adapter variants and placement.** CS-Avg (%). Fixed-rate columns use $2L$, 2×10^{-4}, and mean $\pm$ sample SD. Tuned columns report final evaluations at 4×10^{-4}. Budgets match low-rank factors; DoRA additionally trains magnitudes. Bold marks the highest value in each column.

	Llama-3.2-3B			Qwen3-4B		
Method / placement	Fixed $2L$	Tuned $2L$	$8L$	Fixed $2L$	Tuned $2L$	$8L$
Gradient top-k	63.83 ± 0.97	68.80	72.94	82.10 ± 0.08	82.76	83.29
Uniform quarter	66.22 ± 0.74	70.45	73.08	82.16 ± 0.07	82.87	83.69
AdaLoRA	66.71 ± 0.82	70.82	73.40	82.20 ± 0.09	82.95	83.78
FoRA	65.12 ± 0.95	69.45	73.05	82.13 ± 0.08	82.80	83.42
PiSSA quarter	66.88 ± 0.81	71.18	73.32	82.22 ± 0.08	83.02	83.82
DoRA quarter	67.15 ± 0.76	71.42	73.55	82.25 ± 0.07	83.10	83.90
LoRA full-depth	68.32 ± 1.07	72.31	74.25	82.37 ± 0.12	83.55	84.37
DoRA full-depth	$\mathbf{68.95 \pm 0.88}$	**72.84**	**74.68**	$\mathbf{82.48 \pm 0.10}$	**83.72**	**84.55**

Factorization gains complement distributed placement. At fixed learning rate, quarter-depth PiSSA and DoRA improve on Uniform quarter on both models. Full-depth LoRA nevertheless exceeds quarter-depth DoRA by 1.17 pp on Llama and 0.12 pp on Qwen, even though the latter trains about 6.4% and 6.7% more parameters. It also leads both quarter-depth variants at each tuned budget. Full-depth DoRA attains the highest means in this expanded comparison, including 68.95% on Llama and 82.48% on Qwen at fixed rate. The combined configuration strengthens the practical case for distributed placement; its additional magnitude parameters are included in the resource comparison.

Table 24: **Standard parameter counts at $2L$.** LoRA/PiSSA counts cover the A/B factors; DoRA totals additionally include its trainable output-channel magnitudes.

Model	LoRA / PiSSA	DoRA quarter	DoRA full-depth
Llama-3.2-3B	3,039,232	3,232,768	3,813,376
Qwen3-4B	4,128,768	4,405,248	5,234,688

U SCALING THE PLACEMENT COMPARISON TO LLAMA-3.1-8B

We extend the downstream placement comparison to Llama-3.1-8B, with 32 transformer layers, hidden width 4,096, and intermediate width 14,336. Its grouped-query attention uses 32 query heads and eight key/value heads, each of width 128. The experiment compares static selections, uniform spacing, adaptive rank allocation, and quarter-depth DoRA under the same seven-task CS-Avg metric.

Architecture-specific allocation budget. For the seven target projections, the sum of input and output widths in one layer is

$$S_{8\mathrm{B}} = \underbrace{8{,}192}_{q} + \underbrace{2(5{,}120)}_{k,v} + \underbrace{8{,}192}_{o} + \underbrace{3(18{,}432)}_{\text{gate,up,down}} = 81{,}920, \tag{5}$$

$$N_{AB}^{2L} = 32 \times 2 \times 81{,}920 = 8 \times 8 \times 81{,}920 = 5{,}242{,}880. \tag{6}$$

The full-depth and quarter-depth LoRA allocations therefore have exactly equal factor counts. At budget $8L$, the count is $20{,}971{,}520$. For standard quarter-depth DoRA, the eight adapted layers add $8 \times 43{,}008 = 344{,}064$ magnitudes, giving $5{,}586{,}944$ and $21{,}315{,}584$ parameters at the two budgets. AdaLoRA uses all 32 layers with a target average rank of two at $2L$; its final budget is defined by the weighted factor count in Equation 1.

Table 25: **Llama-3.1-8B placement comparison.** CS-Avg (%). Fixed-rate results use $2L$ and 2×10^{-4} (three-seed mean $\pm$ sample SD); tuned evaluations use 4×10^{-4}. Static quarter-depth ranks are 8/32 and full-depth ranks are 2/8 at budgets $2L/8L$. DoRA's extra magnitudes are counted in the text.

Method / placement	Active layers	Fixed $2L$	Tuned $2L$	Tuned $8L$
Gradient top-k	8	73.45 ± 0.68	76.24	79.20
Uniform quarter	8	75.12 ± 0.51	77.82	79.85
Contraction top-k	8	74.68 ± 0.85	77.40	79.62
AdaLoRA	32	75.38 ± 0.58	78.04	80.02
DoRA quarter	8	75.55 ± 0.52	78.36	80.18
LoRA full-depth	32	$\mathbf{76.84 \pm 0.62}$	**79.15**	**80.62**

Placement ordering at larger scale. At fixed learning rate, Uniform quarter exceeds Gradient by 1.67 pp while keeping both the number of adapted layers and the rank per layer fixed. Full-depth LoRA reaches $76.84 \pm 0.62\%$, ahead of Gradient by 3.39 pp, AdaLoRA by 1.46 pp, and quarter-depth DoRA by 1.29 pp. Full-depth has the highest mean among the six evaluated configurations. The reported Gradient selection is concentrated in the first 12 layers, whereas the uniform set spans the model at stride four.

The tuned comparison retains the same Full-depth ordering at both budgets. Its margin over Uniform quarter decreases from 1.33 pp at $2L$ (79.15 versus 77.82%) to 0.77 pp at $8L$ (80.62 versus 79.85%). This reproduces the smaller distributed-placement advantage at the larger budget observed on Llama-3.2-3B. Together with the smaller-model comparisons, the 8B results extend the empirical support for distributed low-rank allocation across model scales and adapter choices.