DO PRETRAINED SPECTRAL SUBSPACES MATTER FOR ADAPTATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Do accuracy gains require updates along principal pretrained directions, and what does interaction between spectral regions change? We use fixed-band adapters to separate spectral location from within-band flexibility, then study leading–tail interaction during training and at trained checkpoints. Under fixed recipes, leading, middle, tail and random supports all improve accuracy on RTE and CommonsenseQA. These supports are not interchangeable: tail gives the lowest decoder negative log-likelihood (NLL) within each adapter family. On encoder tasks, an interaction penalty sharply reduces cross-band energy share relative to a roughly norm-matched control. It lowers uncalibrated NLL without a resolved accuracy gain; temperature scaling removes most of the NLL advantage. In a separate decoder study, removing learned cross-region components raises choice NLL relative to shrinking each matrix's update to the same remaining norm. Thus, accuracy gains do not require principal support, while the predictive effects of cross-region interaction depend on the metric and the intervention.

1 INTRODUCTION

Where should a parameter-efficient update act in a pretrained weight matrix? Spectral adaptation methods make different choices. PiSSA initializes trainable factors from principal singular directions, whereas MiLoRA uses minor directions (Meng et al., 2024; Wang et al., 2025b). SVDiff and SVFT instead optimize coefficients or selected combinations in a fixed singular basis (Han et al., 2023; Lingam et al., 2024). Comparisons among these methods can change both *where* an update acts and *how freely* directions combine.

Initialization is not confinement. Trainable factors may leave their initial subspaces; a fixed-support adapter remains restricted throughout training. Within that support, changing paired directions independently also differs from allowing them to combine. We therefore ask: **which spectral regions support accuracy gains under confinement**, and **what changes when updates connect leading and tail regions**?

Our adapters make these choices testable. Coefficient adapters change paired directions, while rotated adapters allow additional combinations within the same support. We compare equal-width leading, middle, tail and random supports. Parameter counts match across supports within each family; rotations add parameters. A separate construction permits leading–tail interaction, which we suppress during training with interaction regularization. Checkpoint interventions test removal of learned cross-region components against shrinking each matrix's update to the same remaining norm. The adapters serve as experimental instruments for studying spectral restrictions.

Our contributions are:

- **Separating support from interaction freedom.** Our analysis separates two sources of approximation error: unavailable directions and restricted combinations within the available support (Proposition 1). It also explains how scaling creates cross-region components and why suppressing them need not freeze leading adaptation (Section 3).
- **Accuracy gains are not confined to principal directions.** Leading, middle, tail and random supports all improve accuracy on RTE and CommonsenseQA. Location nevertheless affects performance: tail gives the lowest decoder NLL within each adapter family, while accuracy effects depend on the task and adapter flexibility (Section 5.1).

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.

Change band strengths	**Combine band directions**	**Allow interaction**	**Penalize interaction**
$\begin{pmatrix} 0 & 0 \\ 0 & \mathrm{diag}(h) \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & H \end{pmatrix}$	$\begin{pmatrix} C_{LL} & C_{LT} \\ C_{TL} & C_{TT} \end{pmatrix}$	$\begin{pmatrix} C_{LL} & 0 \\ 0 & C_{TT} \end{pmatrix}$
Fixed directions	Rotations within support	Input/output scaling	Leading changes can remain

Figure 1: Update structure in the fixed pretrained basis, illustrated with tail support. Coefficients and rotations keep updates within the chosen band. Scaling permits changes within and between leading and tail regions; its zero-interaction limit retains both within-region blocks.

- **Testing whether reducing interaction improves predictions.** On encoder tasks, interaction regularization changes cross-band structure relative to a roughly norm-matched control without establishing an accuracy gain. At trained decoder checkpoints, removing learned cross-region components raises choice NLL relative to per-matrix norm-matched shrinking (Section 5.2).

2 Separating spectral location and flexibility

2.1 Confine the update to fixed directions

A fixed spectral band makes location an explicit constraint. For a square pretrained matrix $W_0 \in \mathbb{R}^{d\times d}$, write

$$W_0 = U\Sigma V^\top, \tag{1}$$

with singular values in descending order. Choose an index set B of width $r = |B|$ and hold its input and output bases fixed. An adapted layer $W_{\text{eff}} = W_0 + \Delta_B$ uses

$$\Delta_B = U_B H V_B^\top. \tag{2}$$

Learning the $r \times r$ core H changes the update without leaving these spans. We can therefore vary the freedom inside H while holding its location fixed. Leading, middle and tail bands use the same construction; Appendix A.4 covers rectangular weights.

For the random-frame control, we replace U_B, V_B with independent random orthonormal bases of the same width, drawn once per seed and matrix and shared by both adapter forms. The update stays in that frame but can connect leading and tail directions in the pretrained basis.

2.2 Change coefficients, then allow rotations

The coefficient adapter (C) sets $H = \mathrm{diag}(h)$, with $h \in \mathbb{R}^r$ trainable. It changes each paired input/output direction independently. Fixed-basis coefficient tuning is established by SVDiff and SVFT (Han et al., 2023; Lingam et al., 2024); here it provides a restricted reference.

The rotated adapter (R) allows two learned orthogonal rotations over the last $q \le r$ directions of the selected band. Split the band into $r - q$ fixed directions (M) and q rotatable directions (S), and set

$$H = \begin{bmatrix} \mathrm{diag}(h_M) & 0 \\ 0 & R_U\,\mathrm{diag}(h_S)R_V^\top \end{bmatrix}, \qquad R_U, R_V \in \mathbb{O}(q). \tag{3}$$

At $q = 0$, this reduces to fixed coefficients. At $q = r$, arbitrary coefficients and the inner SVD can represent any real core H. Intermediate values allow combinations within a smaller block. We enforce orthogonality through a matrix exponential (Appendix B.7). The implementation names are LinLoRA for C and OrthoLoRA for R.

Both families use the same support, but rotations add trainable parameters. Their comparison tests added flexibility within a region, without isolating rotation at equal capacity.

2.3 Separate location from flexibility

A restricted adapter can miss a target update for two reasons: the required directions lie outside its support, or the required combinations are forbidden inside that support. Rotations relax the second restriction while preserving the first. Let G be a target weight update; $\|\cdot\|_F$ denotes the Frobenius norm, the square root of the sum of squared entries.

Proposition 1 (Two distinct approximation bottlenecks). *For an arbitrary target matrix $G \in \mathbb{R}^{d\times d}$, let $G_{BB} = U_B^\top G V_B$. Then*

$$\min_h \left\| G - U_B \operatorname{diag}(h) V_B^\top \right\|_F^2 = \|G\|_F^2 - \sum_{i=1}^{r} |(G_{BB})_{ii}|^2, \tag{4}$$

$$\min_H \left\| G - U_B H V_B^\top \right\|_F^2 = \|G\|_F^2 - \|G_{BB}\|_F^2 . \tag{5}$$

The improvement in the best attainable squared error equals the off-diagonal energy of G_{BB}. Both families retain the same unavoidable outside-support residual.

The free core can express every input/output combination inside the band; diagonal coefficients keep only paired directions. The proof follows by orthogonal projection (Appendix A.1). This is an approximation result in weight space, not a guarantee about trained prediction loss. Our experiments rotate only part of each band, rather than using the full-core limit.

For interaction analysis, we now specialize to a tail band T of width r and its leading complement L of width $k = d - r$. Thus $W_0 = U_L \Sigma_L V_L^\top + U_T \Sigma_T V_T^\top$. The leading complement includes the middle experimental band; its width depends on the backbone.

Local tail confinement preserves leading outputs for the same layer input: $U_L^\top (W_0 + \Delta_T)x = U_L^\top W_0 x$. The pretrained leading components remain in the model, and earlier layers can still change their inputs. Confinement also need not preserve the ranking of singular directions. The original leading directions remain top-ranked when

$$\sigma_k(W_0) > \|\Sigma_T + H\|_2. \tag{6}$$

Appendix A.3 gives a counterexample when this separation fails.

3 Controlling leading–tail interaction

Cross-band interaction becomes possible when we relax confinement. Diagonal output and input scalers E, D act in the layer's original coordinates rather than its singular basis, so $EU_T H V_T^\top D$ can connect directions outside the selected tail.

The scaled adapter used in our interaction experiments also includes a leading core. We write both choices in one form:

$$\Delta = E\big(U_L S_L V_L^\top + U_T H V_T^\top\big) D, \quad S_L = \begin{cases} 0 & \text{tail core only,} \\ I & \text{leading-plus-tail core.} \end{cases} \tag{7}$$

The expression is added to W_0. With $S_L = I$, learning the scalers changes both the leading identity core and the tail contribution. We call this the scaled leading-plus-tail adapter: its implementation is UILinLoRA with diagonal H, or UIOrthoLoRA with rotations. The UNREG/MIX/NORM comparisons use the coefficient form.

3.1 Identify the interactions to control

We inspect the effective update in the original pretrained basis:

$$C = U^\top \Delta V = \begin{bmatrix} C_{LL} & C_{LT} \\ C_{TL} & C_{TT} \end{bmatrix}. \tag{8}$$

C_{TT} records tail-to-tail changes, including any within-tail combinations. C_{LL} records changes among leading directions. The two cross blocks link the leading and tail groups. Strict tail confinement leaves only C_{TT}. Decoupling the two groups instead leaves both within-group blocks, C_{LL} and C_{TT}.

Let $A = U^\top E U$, $B = V^\top D V$, and define the two scaler-mixing operators

$$M_E = U_L^\top E U_T, \qquad M_D = V_L^\top D V_T, \qquad \mu = \|M_E\|_2, \quad \nu = \|M_D\|_2. \tag{9}$$

The norm $\|\cdot\|_2$ is the operator norm. Direct multiplication gives

$$C_{ab} = A_{aL} S_L B_{Lb} + A_{aT} H B_{Tb}, \qquad a, b \in \{L, T\}. \tag{10}$$

The two operators measure how the scalers connect the groups. When both are zero, the cross blocks vanish. A tail-only core then recovers strict confinement. The leading-plus-tail core retains its leading block: for example, $E = D = I, H = 0$ gives $C_{LL} = I$. More generally, the block bounds imply

$$\max\{\|C_{LT}\|_2, \|C_{TL}\|_2\} \leq (\|S_L\|_2 + \|H\|_2) \max\{\|E\|_2 \nu, \|D\|_2 \mu\}. \tag{11}$$

Small mixing operators control interaction only when the other factor norms stay bounded; rescaling can otherwise keep cross blocks large despite a small factor penalty. We therefore measure the attained update directly (bounds: Appendix A.2; counterexample: Appendix A.9).

3.2 How the interaction penalty works

We call the interaction penalty MIX and train with

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \sum_{\ell \in \mathcal{M}} \left(\lambda_E \|M_{E,\ell}\|_F^2 + \lambda_D \|M_{D,\ell}\|_F^2 \right), \tag{12}$$

where $\mathcal{M}$ contains the adapted matrices. The two coefficients control the pressure on output-side and input-side interaction.

The penalty's mechanism is explicit. Write $P = U_L U_L^\top$ and $E = \text{diag}(\epsilon)$. Then

$$\|M_E\|_F^2 = \tfrac{1}{2}\|EP - PE\|_F^2 = \sum_{i<j} P_{ij}^2 (\epsilon_i - \epsilon_j)^2. \tag{13}$$

MIX penalizes differences between coordinate scalers, weighted by the pretrained projector. If the graphs with edges $P_{ij} \neq 0$ and $(V_L V_L^\top)_{ij} \neq 0$ are connected, $M_E = M_D = 0$ implies uniform scalers $E = aI, D = bI$. The leading-plus-tail core then has $C_{LL} = abI$, $C_{TT} = abH$ and no cross blocks. Decoupling the regions therefore need not freeze the leading update (Appendix A.7). Consistent with this limit, MIX retains a predominantly identity-aligned learned leading update on both encoder tasks (Appendix C.4). Training need not reach the limit, so we measure attained geometry and predictions directly.

For the training-time comparison, an update-size penalty controls for shrinking the update: it penalizes total displacement from the pretrained weights rather than scaler interaction. We compare attained cross blocks and predictions, measuring the effective update to account for reciprocal factor rescaling (Appendices A.7–A.10).

A separate post hoc test starts from trained random-frame or LoRA checkpoints. It zeroes C_{LT} and C_{TL}, then compares the result with scaling each matrix's entire update to the same remaining Frobenius norm; no adapter is retrained (Appendices B.17–B.20). These are different adapter families from the training-time study: removal tests endpoint sensitivity, not whether training needs interaction.

4 Measuring adaptation

The same saved checkpoint supplies both outcomes: what the model predicts and which spectral directions its update connects.

Experiment map. The location study uses RTE, CommonsenseQA (CSQA) and GSM8K, adapting the listed attention projections in every layer. Q/K/V/O denote query, key, value and output projections; r/d is the fraction of spectral directions available to each band, and q is the rotation-block width.

Backbone	Task	Projections	Band r/d	Rotation q
RoBERTa-base	RTE	Q/K/V/O	256/768	64
Qwen2.5-1.5B	CSQA, GSM8K	Q/O	512/1536	128
Qwen2.5-3B	CSQA	Q/O	512/2048	128
Llama-3.1-8B	CSQA	Q/O	512/4096	128

All decoders use their instruction-tuned checkpoints. Bands have equal widths within each backbone, but cover different fractions across backbones; R rotates only the last quarter of each band. Training-time interaction comparisons use RoBERTa on RTE/MRPC and Qwen1.5B on CSQA, with the same

projections and a leading two-thirds/tail third partition. Checkpoint interventions use each decoder's tail width above and its leading complement. GSM8K uses an exploratory revised recipe; the Llama middle band is a post hoc addition, and Qwen3B uses a post hoc admission rule (Appendix B).

Prediction loss and accuracy. Accuracy counts correct labels on RTE/MRPC, correct choices on CommonsenseQA and the final number of a generated GSM8K answer. Loss is negative log likelihood (NLL) per label or choice, or per reference-completion token on GSM8K. Location and training-time tables use uncalibrated NLL from each cohort's native evaluation. Within each checkpoint-intervention cohort, all variants use one float32 path, separate from native endpoint scores. We do not pool loss units across tasks. Primary comparisons pair three training seeds at fixed endpoints; seed and evaluation-example uncertainty are reported separately. Appendices C.10 and C.12 examine selection-fitted temperature scaling, which leaves classification predictions unchanged when the temperature is positive.

Interaction between leading and tail directions. The four blocks in (8) partition the squared update norm:

$$\|\Delta\|_F^2 = \sum_{a,b\in\{L,T\}} \|C_{ab}\|_F^2. \tag{14}$$

For a nonzero update, $p_{ab} = \|C_{ab}\|_F^2/\|\Delta\|_F^2$ is the fraction of energy in a block. We call $p_{\text{cross}} = p_{LT} + p_{TL}$ the cross share: it measures the share connecting leading and tail directions. The total share outside the tail is $p_{\text{off}} = p_{LL} + p_{\text{cross}}$. Off-diagonal entries *inside* C_{TT} instead describe combinations within the tail. Fractions at an exactly zero update are undefined.

Overall size and distribution across the network. Relative update size is $\rho_F = \|\Delta\|_F/\|W_0\|_F$; it compares interaction control with simply reducing weight change. Network-wide summaries sum squared energies across adapted matrices before taking ratios. Because the scaled leading-plus-tail adapter has a nonzero update at insertion, we distinguish total displacement from W_0 from learned displacement since insertion (protocol and probe in Appendix B; layer profiles and learned displacement in Appendix C).

5 Experiments: location and interaction

We test where confined updates can adapt and what leading–tail interaction changes. Each comparison uses a common recipe and three paired training seeds. We evaluate accuracy and prediction loss on held-out task examples; protocols, data splits and full seed results are in the appendices.

5.1 Does adaptation require particular spectral directions?

We compare equal-width leading, middle, tail and random supports using the coefficient (C) and rotated (R) adapters from Section 2. Parameter counts match across supports within each family; rotations add parameters. Random frames can connect pretrained regions, so their comparison tests orientation as a whole. Protocols and follow-ups are in Appendix B.

Adaptation is not confined to principal directions. Leading, middle and tail updates all improve accuracy over the reference on RTE and CommonsenseQA, as do random frames (Table 1). Location still affects performance: tail gives the lowest decoder NLL within both families, and tail rotations exceed leading rotations in accuracy on Qwen3B and Llama (Table 2). On RTE, the accuracy gains come with higher NLL than the head-only reference. On GSM8K, every arm lowers completion NLL, but none has a resolved accuracy gain over the frozen reference; tail rotations have a resolved decline (Appendix B.10).

Within-band flexibility improves likelihood more reliably than accuracy. Adding rotations lowers decoder NLL within every tested support. Mean accuracy increases throughout CommonsenseQA but decreases throughout the exploratory GSM8K follow-up, including a resolved decline at tail support (Table 2; Appendix B.10). Better likelihood therefore does not guarantee better task decisions. On Qwen3B, tail rotations average 0.49 accuracy points below LoRA, with an unresolved paired difference and lower choice NLL (Appendix B.16).

Table 1: Adaptation within fixed spectral support. Cells report accuracy (%, ↑) / NLL (↓). C learns coefficients; R adds rotations within the same support. Leading, middle and tail bands have equal widths within each model; Random uses a fixed random frame. Adapted entries are three-seed means. References are head-only RoBERTa or frozen decoders. NLL scores RTE class labels, CSQA answer choices or GSM8K reference-completion tokens (protocols: Appendix B).

Support	Form	RoBERTa RTE	Qwen 1.5B GSM8K	Qwen 1.5B CSQA	Qwen 3B CSQA	Llama 8B CSQA
Leading	C	72.68/1.783	59.11/0.5439	75.51/0.6651	78.27/0.6170	77.31/0.6028
	R	74.61/1.932	58.51/0.5285	76.71/0.6258	80.75/0.5402	79.01/0.5567
Middle	C	69.92/1.726	59.20/0.5405	75.70/0.6623	78.38/0.6195	77.42/0.6026
	R	71.72/2.047	57.73/0.5233	76.69/0.6232	80.89/0.5419	79.14/0.5547
Tail	C	69.19/2.007	58.94/0.5297	76.03/0.6493	79.77/0.5719	78.13/0.5720
	R	68.35/2.250	56.86/0.5093	77.15/0.6128	82.23/0.5135	80.48/0.5303
Random	C	71.96/1.684	59.98/0.5365	75.68/0.6562	78.65/0.5969	78.05/0.5938
	R	74.13/1.733	56.68/0.5198	77.40/0.6176	81.22/0.5325	79.80/0.5456
Reference	–	58.12/0.680	59.11/0.5934	70.60/1.9129	76.99/2.9985	76.66/0.9739

Table 2: Paired effects of spectral location and added rotations. Each row reports the first configuration minus the second. Positive accuracy differences favor the first; negative NLL differences favor the first. Accuracy is in percentage points; NLL scores CSQA choices or GSM8K completion tokens. Brackets give nominal 95% intervals over three paired training seeds; intervals containing zero are unresolved (Appendix B).

Model / task	Contrast	Δ accuracy (pp)	Δ NLL
Qwen1.5B / CSQA	Tail R − Leading R	$+0.44$ $[-0.08, +0.95]$	-0.0130 $[-0.0158, -0.0101]$
	Tail R − Tail C	$+1.12$ $[+0.81, +1.43]$	-0.0365 $[-0.0389, -0.0341]$
Qwen1.5B / GSM8K	Tail R − Leading R	-1.65 $[-4.28, +0.99]$	-0.0191 $[-0.0193, -0.0189]$
	Tail R − Tail C	-2.08 $[-4.05, -0.12]$	-0.0204 $[-0.0207, -0.0200]$
Qwen3B / CSQA	Tail R − Leading R	$+1.47$ $[+1.12, +1.83]$	-0.0266 $[-0.0269, -0.0263]$
	Tail R − Middle R	$+1.34$ $[+1.03, +1.65]$	-0.0284 $[-0.0324, -0.0244]$
	Tail R − Random R	$+1.01$ $[+0.42, +1.60]$	-0.0190 $[-0.0230, -0.0150]$
	Tail R − Tail C	$+2.46$ $[+2.10, +2.81]$	-0.0583 $[-0.0636, -0.0530]$
Llama8B / CSQA	Tail R − Leading R	$+1.47$ $[+0.05, +2.90]$	-0.0264 $[-0.0298, -0.0229]$
	Tail R − Tail C	$+2.35$ $[+0.82, +3.87]$	-0.0417 $[-0.0490, -0.0343]$

5.2 Should spectral regions interact?

Training compares the scaled leading-plus-tail coefficient adapter (Section 3) on RoBERTa-base (RTE/MRPC) and Qwen2.5-1.5B (CSQA), with leading two-thirds and tail third connected by input/output scalers. We apply no penalty (UNREG), an interaction penalty (MIX) or an update-size penalty (NORM), calibrated using norms alone (Appendix B.4). Separately, we remove cross-region components from trained random-frame and LoRA checkpoints.

Table 3: Training-time control of leading–tail interaction. UNREG has no penalty; MIX penalizes interaction; NORM penalizes update size. The scaled leading-plus-tail coefficient adapter uses RoBERTa-base on RTE/MRPC and Qwen2.5-1.5B on CSQA. Values are three-seed means. ρ_F is relative update norm; Cross is the percentage of update energy connecting the two regions. Accuracy is percent (↑); uncalibrated NLL (↓) scores class labels or CSQA answer choices. The CSQA size control is imperfect.

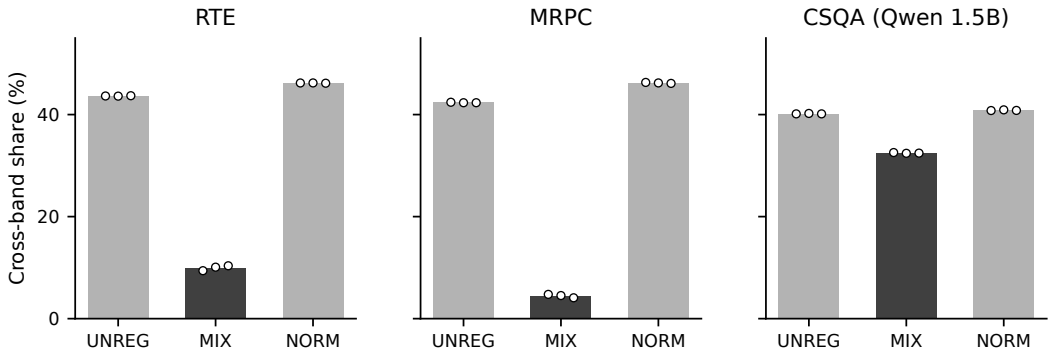
Task	Condition	ρ_F	Cross (%)	Accuracy	NLL
RTE	UNREG	0.2751	43.7	71.72	2.429
RTE	MIX	0.0736	10.0	71.36	1.229
RTE	NORM	0.0759	46.2	71.24	2.193
MRPC	UNREG	0.1802	42.4	86.19	0.871
MRPC	MIX	0.0407	4.5	85.46	0.395
MRPC	NORM	0.0423	46.2	85.78	0.755
CSQA	UNREG	0.0060	40.2	74.67	0.675
CSQA	MIX	0.0031	32.5	74.50	0.689
CSQA	NORM	0.0040	40.8	74.69	0.679

Suppressing interaction changes geometry without a resolved accuracy gain. At similar encoder update norms, MIX retains about one-fifth of NORM's cross-band share on RTE and one-tenth on MRPC, while leading adaptation remains (Table 3; Figure 2). MIX lowers uncalibrated encoder NLL, but temperature scaling removes most of this advantage; the remaining NLL contrasts and the accuracy differences are unresolved (Appendix C.10). On CommonsenseQA, MIX reduces cross share only modestly and raises choice NLL relative to both controls, with unresolved accuracy differences and a failed size match (Appendix B.3).

Learned cross-region structure matters at the endpoint. Removing cross-region components raises choice NLL relative to per-matrix norm-matched shrinking in all seven tested comparisons; three also show resolved accuracy declines (Table 4). Equal remaining norms therefore do not imply equal predictive performance. Learned cross-region structure contributes to endpoint predictions, although this does not establish that interaction is necessary during training (Appendices B.17–B.20).

Table 4: Endpoint effects of removing cross-region components on CommonsenseQA. Rows report removal minus uniform shrinking of the same saved updates, matched in remaining norm separately for each matrix, without retraining. Negative accuracy differences and positive choice-NLL differences mean removal is more damaging. Accuracy differences are percentage points; brackets give nominal 95% intervals over three paired training seeds. Random C/R use coefficients/rotations within a fixed random frame; LoRA r8 has rank eight.

Model	Family	Accuracy difference (pp) [95%]	Choice NLL difference [95%]
Qwen2.5-1.5B	Random R (primary)	-1.28 $[-2.82, +0.26]$	$+0.1403$ $[+0.1359, +0.1447]$
Qwen2.5-1.5B	Random C	-0.90 $[-1.71, -0.09]$	$+0.1720$ $[+0.1646, +0.1795]$
Llama-3.1-8B	Random R (primary)	-0.60 $[-1.54, +0.34]$	$+0.0273$ $[+0.0245, +0.0301]$
Llama-3.1-8B	Random C	-0.30 $[-0.72, +0.12]$	$+0.0295$ $[+0.0264, +0.0327]$
Qwen2.5-3B	Random R (primary)	-0.85 $[-1.16, -0.54]$	$+0.3222$ $[+0.2866, +0.3578]$
Qwen2.5-3B	Random C	-0.41 $[-0.61, -0.21]$	$+0.3980$ $[+0.3575, +0.4384]$
Qwen2.5-3B	LoRA r8	$+0.00$ $[-1.42, +1.42]$	$+0.0256$ $[+0.0123, +0.0389]$



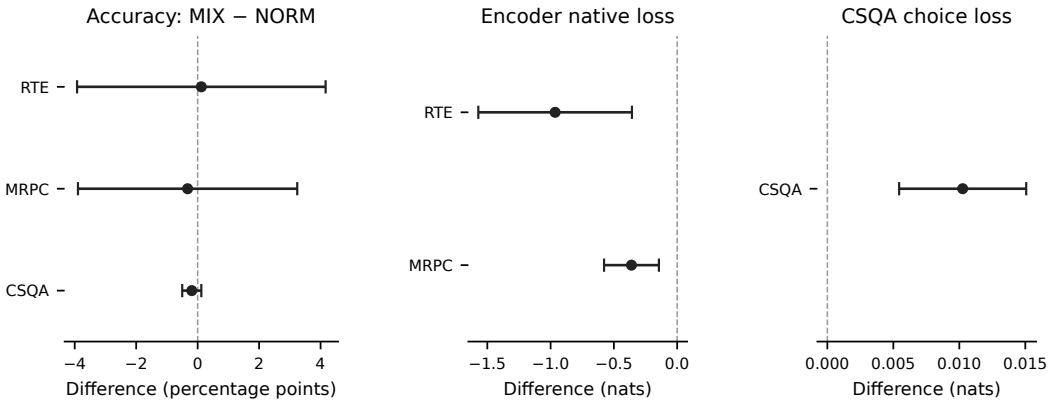


Figure 2: Geometry and prediction changes under MIX. Top: mean cross-band share with individual training seeds. Bottom: paired MIX−NORM differences in accuracy (left) and NLL (middle, right), with nominal 95% intervals over three seeds. Positive accuracy differences favor MIX; negative NLL differences favor MIX. Encoder NLL is computed over class labels and CSQA NLL over answer choices. The CSQA size control is imperfect.

6 Relation to spectral adaptation methods

Choosing spectral location. LoRA constrains update rank (Hu et al., 2022); AdaLoRA allocates its singular-value budget (Zhang et al., 2023). PiSSA and MiLoRA initialize free factors from leading and minor pretrained directions (Meng et al., 2024; Wang et al., 2025b), while KaSA and CorDA also use decompositions to guide adaptation (Wang et al., 2025a; Yang et al., 2024). Our equal-width bands enforce location throughout training, separating confinement from initialization. VeRA and RandLoRA already adapt with fixed random matrices (Kopiczko et al., 2024; Albert et al., 2025); our random frames match band width and core freedom to test orientation, including the resulting cross-region structure.

Allowing combinations within fixed support. SVDiff trains spectral shifts, SVFT learns selected combinations of fixed singular vectors, and LoRA-XS trains a small core on fixed support (Han et al., 2023; Lingam et al., 2024; Bałazy et al., 2025). Spectral Adapter rotates a selected leading core, and PSOFT relaxes orthogonality inside principal support (Zhang & Pilanci, 2024; Wu et al., 2026). We vary support, within-band flexibility and cross-region interaction separately; Appendix A.5 gives the corresponding coordinate forms.

Other restrictions on adaptation. OFT/BOFT and HRA constrain orthogonal transformations (Qiu et al., 2023; Liu et al., 2024b; Yuan et al., 2024); DoRA, SSF and IA3 restrict factors or scalers (Liu et al., 2024a; Lian et al., 2022; Liu et al., 2022); FourierFT tunes Fourier coefficients (Gao et al., 2024). Restricted-parameter studies more broadly ask which adaptation freedoms are needed (Li et al., 2018; Aghajanyan et al., 2021). We examine that question through location and interaction in the pretrained singular basis.

Geometry of learned updates. Prior work studies the rank and forgetting effects of full fine-tuning versus LoRA (Biderman et al., 2024), new high-ranking singular directions introduced by LoRA (Shuttleworth et al., 2025), and differences in principal alignment between reinforcement and supervised learning (Zhu et al., 2025). Our measurements keep the pretrained frame fixed. Energy in that frame is distinct from movement of the adapted model's leading singular subspace (Appendix A.3; Wedin, 1972).

7 Scope and practical considerations

We compare one encoder and three decoders under fixed recipes; Qwen 3B and Llama cover CommonsenseQA only. Bands share a learning rate within each task, although spectral initializations can favor different rates (Lee et al., 2026). We therefore do not identify each band's best tuned performance. Qwen 3B proceeded under a disclosed post hoc validity rule after its registered pilot accuracy gate failed (Appendix B.14). Its post hoc LoRA comparison uses the same recipe without a separate hyperparameter search (Appendix B.16). The GSM8K recipe is also post hoc: it reuses an inspected test split and generates on a fixed 768-example subset.

Intervals are descriptive and unadjusted; unresolved differences do not establish equivalence. Encoder MIX–NORM effects near 5 points are needed for 80% power (Appendix C.8).

The penalties also change module allocation. Size matching is approximate on the encoder tasks and fails on CommonsenseQA, so these comparisons do not isolate interaction at equal update size. Checkpoint interventions are post hoc and do not isolate rank or distance from the trained update. Temperature scaling tests one form of confidence correction.

The scaled leading-plus-tail construction has substantial costs before merging. Its RoBERTa-base branch uses about 192 times rank-4 LoRA's multiply–accumulates; full pretrained bases also require quadratic storage per matrix. The arithmetic ratio is not measured throughput, and a small trainable parameter count does not imply a small training footprint (Appendix D).

8 Conclusion

Updating principal pretrained directions is not required for accuracy gains in the tested RTE and CommonsenseQA settings. Leading, middle, tail and random supports all permit adaptation, but their performance differs: tail gives the lowest decoder NLL within each adapter family, while accuracy depends on the task and within-band flexibility.

Cross-region interaction is a separate property. On encoder tasks, MIX suppresses it while retaining leading changes, without establishing an accuracy advantage; its native NLL benefit is largely removed by temperature scaling. On CommonsenseQA, MIX raises choice NLL with a modest cross-share reduction and an imperfect size control. At trained decoder checkpoints, selective cross-region removal produces higher choice NLL than per-matrix norm-matched shrinking.

Together, these results separate *where* an update can act from *how* spectral regions interact: successful adaptation does not require updates along principal directions, and the predictive effects of cross-region structure depend on the metric and intervention.

REPRODUCIBILITY STATEMENT

Proofs, protocols and costs appear in Appendices A–D. The supplementary analysis archive contains the displayed outcomes, retained encoder logits, decoder per-example losses, source hashes and scripts that regenerate the tables and contrasts. Original and post hoc cohorts remain distinct. The separate code archive supplies training, evaluation and checkpoint-intervention implementations, including the recorded source versions. Neither archive includes model weights, prepared inputs or checkpoints. The analysis archive reproduces analysis of exports, not training or checkpoint inference; the post hoc prediction-disagreement analysis (Appendix C.14) is a separate CPU script over the same saved predictions. Historical comparisons use separate protocols.

AI USE STATEMENT

Generative AI tools assisted with conceptual framing, mathematical formulation and proof drafting, literature lookup, experiment design, implementation of experiment and analysis code, interpretation of results, figure and table generation, and manuscript drafting and editing. The reported model measurements come from experimental artifacts; synthetic-matrix checks are identified separately. Automated algebra, data-reconstruction, and document checks support verification but do not replace author review of the scientific claims. The authors are responsible for the final text, code, and reported results.

ETHICS STATEMENT

The study analyzes adaptation of existing language models using existing benchmarks. Weight-space stability does not certify factual correctness, fairness, privacy, or safe behavior. Efficient adaptation can reduce training costs but can also facilitate harmful specialization or preserve biases present in a base model. Our structural claims should not be interpreted as deployment guarantees.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7319–7328, 2021. URL https://aclanthology.org/2021.acl-long.568.

Paul Albert, Frederic Z Zhang, Hemanth Saratchandran, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. RandLoRA: Full-rank parameter-efficient fine-tuning of large models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/382b95a6580cfb0b5ca33c74b4e0e770-Abstract-Conference.html.

Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. LoRA-XS: Low-rank adaptation with extremely small number of parameters. In *European Conference on Artificial Intelligence*, 2025. URL https://arxiv.org/abs/2405.17604.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.09673.

Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. Parameter-efficient fine-tuning with discrete Fourier transform. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 14884–14901, 2024. URL https://proceedings.mlr.press/v235/gao24o.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. SVDiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7323–7334, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Dawid J Kopiczko, Tijmen Blankevoort, and Yuki M Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, 2024.

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

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryup8-WCW.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2210.08823.

Vijay Lingam, Atula Tejaswi, Aditya Vavre, Aneesh Shetty, Gautham K Gudur, Joydeep Ghosh, Alex Dimakis, Eunsol Choi, Aleksandar Bojchevski, and Sujay Sanghavi. SVFT: Parameter-efficient fine-tuning with singular vectors. *Advances in Neural Information Processing Systems*, 37:41425–41446, 2024.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2205.05638.

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 *Forty-first International Conference on Machine Learning*, 2024a.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and Bernhard Schölkopf. Parameter-efficient orthogonal finetuning via butterfly factorization. In *International Conference on Learning Representations*, 2024b. URL https://arxiv.org/abs/2311.06243.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Xinyu Ma, Xu Chu, Zhibang Yang, Yang Lin, Xin Gao, and Junfeng Zhao. Parameter efficient quasi-orthogonal fine-tuning via givens rotation. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2404.04316.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 201–206, 2017. URL https://aclanthology.org/W17-5525.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.07280.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2410.21228v3.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

Fan Wang, Juyong Jiang, Chansung Park, Sunghun Kim, and Jing Tang. KaSA: Knowledge-aware singular-value adaptation of large language models. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2412.06071.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. MiLoRA: Harnessing minor singular components for parameter-efficient LLM finetuning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4823–4836, 2025b.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972. doi: 10.1007/BF01932678.

Fei Wu, Jia Hu, Geyong Min, and Shiqiang Wang. Efficient orthogonal fine-tuning with principal subspace adaptation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.11235.

Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. CorDA: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/83f95bb0ac5046338ea2afe3390e9f4b-Abstract-Conference.html.

Shen Yuan, Haotian Liu, and Hongteng Xu. Bridging the gap between low-rank and orthogonal adaptation via householder reflection adaptation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/cdd0640218a27e9e2c0e52e324e25db0-Abstract-Conference.html.

Fangzhao Zhang and Mert Pilanci. Spectral adapter: Fine-tuning in spectral space. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.13952.

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

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143, 2024. URL https://proceedings.mlr.press/v235/zhao24s.html.

Hanqing Zhu, Zhenyu Zhang, Hanxian Huang, DiJia Su, Zechun Liu, Jiawei Zhao, Igor Fedorov, Hamed Pirsiavash, Zhizhou Sha, Jinwon Lee, David Z. Pan, Zhangyang Wang, Yuandong Tian, and Kai Sheng Tai. The path not taken: RLVR provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

NOTATION GUIDE

Symbol	Meaning
W_0, W_{eff}	Pretrained and adapted effective weight matrices.
U, Σ, V	Fixed pretrained SVD; L denotes leading and T its tail complement.
B, r, q	Selected band, its width $r = \lvert B\rvert$, and rotation-block width $q \le r$.
d, k	Square weight width and leading cutoff; $k = d - r$ for the tail partition.
Δ, C	Effective update $W_{\text{eff}} - W_0$ and coordinates $U^\top \Delta V$.
C_{ab}	Block with output subspace a and input subspace b, $a, b \in \{L, T\}$.
Δ_{learned}	Displacement from the inserted model, $W_{\text{eff}} - W_{\text{eff}}(0)$.
S_L, H	Leading and tail cores in the unified additive update; practical $S_L = I$.
E, D	Ambient diagonal output/input scalers.
A, B	In the bounds: $U^\top EU, V^\top DV$; prior-method factor conventions are stated locally.
M_E, M_D	Scaler mixing operators $U_L^\top EU_T, V_L^\top DV_T$.
μ, ν	Operator norms of M_E, M_D.
e, d_s, s, h_s	Operator norms of E, D, S_L, H, respectively.
p_{ab}, p_{cross}	Normalized block energies and $p_{LT} + p_{TL}$.
$p_{\text{off}}, q_{\text{off}}$	Off-tail energy fraction and its square root (norm ratio).
ρ_F	Relative Frobenius update norm; pooling convention specified per table.
η_I	Fraction of leading-block energy in its fitted scalar-identity component.
P, Q	Left/right leading projectors; context specifies any random projector.
d_{ch}, o	Normalized chordal distance and leading-subspace overlap.
λ_E, λ_D	Coefficients of the summed scaler-mixing penalties.
β	Coefficient of the mean module-relative squared-norm penalty.
δ_s, e_s	Signed NORM/MIX norm error and its absolute value for seed s.
$k_{\text{val}}, k_{\text{vec}}$	Implementation settings for coefficient and rotation dimensions.

Evidence map. Appendix B contains the encoder and decoder cohorts, including location studies, training-time interaction comparisons, checkpoint interventions and separately labelled seed extensions. Appendix C contains calibration, geometry and uncertainty analyses; Appendix D gives costs. Historical comparisons are in Appendices E–G.

A PROOFS AND BOUNDARY CASES

A.1 FIXED COEFFICIENTS AND WITHIN-BAND FLEXIBILITY

Let $G_{BB} = U_B^\top G V_B$. Orthogonal projection gives

$$\left\|G - U_B H V_B^\top\right\|_F^2 = \|G\|_F^2 - \|G_{BB}\|_F^2 + \|G_{BB} - H\|_F^2 . \tag{15}$$

For a free core the last term is minimized at $H = G_{BB}$. For a diagonal core it is minimized by $h_i = (G_{BB})_{ii}$, leaving the off-diagonal entries unchanged. Substitution proves Proposition 1. An SVD of the minimizing free core gives the representation (3) with $q = r$. Thus full within-band rotations can attain the free-core approximation optimum.

For a separate family whose free core has rank at most $q \le r$, rather than the partial-rotation core of (3), truncated SVD of G_{BB} gives minimum squared error

$$\|G\|_F^2 - \sum_{j=1}^{q} \sigma_j(G_{BB})^2. \tag{16}$$

This expression keeps the outside-support residual explicit. The unrestricted rank-q approximation of G can use directions elsewhere.

Random frames. The same projection argument covers a band in a random basis. Let $Q_U, Q_V \in \mathbb{R}^{d\times r}$ be independent Haar-distributed orthonormal frames, let G be fixed independently of them, and let H range over the partial-rotation family of (3): $r - q$ diagonal entries and a free $q \times q$ block, with allowed-entry mask S_q of size $m = r - q + q^2$. The rank-one matrices $q_{U,i} q_{V,j}^\top$ are orthonormal in the Frobenius inner product, so

$$\min_H \left\|G - Q_U H Q_V^\top\right\|_F^2 = \|G\|_F^2 - \left\|P_{S_q}(Q_U^\top G Q_V)\right\|_F^2 , \tag{17}$$

where P_{S_q} keeps the allowed entries; the SVD of the free block supplies its two rotations and signed coefficients. Each frame column has second moment I/d and the two frames are independent, so $\mathbb{E}[(q_{U,i}^\top G q_{V,j})^2] = \|G\|_F^2 / d^2$ for every allowed entry and

$$\mathbb{E} \min_H \left\|G - Q_U H Q_V^\top\right\|_F^2 = \left(1 - m/d^2\right) \|G\|_F^2 . \tag{18}$$

At $q = 0$ the expected captured fraction is r/d^2; at $q = r$ it is r^2/d^2. For the tested rotated forms, $m = 4{,}288$ per encoder matrix and $16{,}768$ per decoder matrix, against $8{,}448$ and $33{,}280$ stored trainable entries. An equal-width random frame is therefore a dimension-only reference for the spectral bands: any advantage of a pretrained band over it cannot come from a larger expressible family; it must come from which directions the frame spans or from optimization within it. The identity holds for a target fixed before the frame is drawn, not for one chosen after training in that frame, and it says nothing about optimization dynamics; at $h = 0$ the rotation gradients vanish, so the initial tangent family is smaller still.

A.2 PROOF OF THE UNIFIED MIXING BOUNDS

Proposition 2 (Interaction bounds with and without a leading core)**.** *Write* $e = \|E\|_2$, $d_s = \|D\|_2$, $s = \|S_L\|_2$, *and* $h_s = \|H\|_2$. *For* (7)*,*

$$\|C_{LL}\|_2 \le e s d_s + \mu h_s \nu, \qquad \|C_{LT}\|_2 \le e s \nu + \mu h_s d_s, \tag{19}$$

$$\|C_{TL}\|_2 \le \mu s d_s + e h_s \nu, \qquad \|C_{TT}\|_2 \le \mu s \nu + e h_s d_s. \tag{20}$$

If $\mu = \nu = 0$*, the cross blocks vanish. In addition, the leading block vanishes when* $S_L = 0$.

The proof applies submultiplicativity to (10). The additional leading-core terms are the difference between the idealized and practical analyses. With $S_L = 0$, leading–leading leakage requires mixing on both sides, while each cross block can be first order in one mixing quantity. With $S_L \neq 0$, direct leading adaptation remains even when the cross terms vanish.

Corollary 3 (Tail confinement and block decoupling)**.** *For $S_L = 0$, define $C_{\text{tail}} = \operatorname{diag}(0, C_{TT})$. Then*

$$\|C - C_{\text{tail}}\|_2 \le h_s \sqrt{(\mu\nu)^2 + (\mu d_s)^2 + (e\nu)^2}. \tag{21}$$

Consequently, $\mu, \nu \to 0$ implies confinement in absolute block norm if e, d_s, h_s stay bounded. If e, d_s, h_s, s stay bounded for a general leading core, the same limit instead makes C block-diagonal; it need not make C_{LL} small.

Because U, V are orthogonal,

$$C = A \begin{bmatrix} S_L & 0 \\ 0 & H \end{bmatrix} B. \tag{22}$$

Direct multiplication gives

$$C_{LL} = A_{LL} S_L B_{LL} + A_{LT} H B_{TL}, \tag{23}$$
$$C_{LT} = A_{LL} S_L B_{LT} + A_{LT} H B_{TT}, \tag{24}$$
$$C_{TL} = A_{TL} S_L B_{LL} + A_{TT} H B_{TL}, \tag{25}$$
$$C_{TT} = A_{TL} S_L B_{LT} + A_{TT} H B_{TT}. \tag{26}$$

Symmetry yields $\|A_{TL}\|_2 = \|A_{LT}\|_2 = \mu$ and $\|B_{TL}\|_2 = \|B_{LT}\|_2 = \nu$. Orthogonal compression bounds the diagonal blocks of A by e and those of B by d_s. Apply the triangle inequality and submultiplicativity to each product to obtain (19)–(20).

For the tail-only core, $s = 0$ and

$$C - C_{\text{tail}} = \begin{bmatrix} C_{LL} & C_{LT} \\ C_{TL} & 0 \end{bmatrix}. \tag{27}$$

For any block vector (x, y), the norms of the two output blocks are bounded componentwise by

$$\begin{bmatrix} \|C_{LL}\|_2 & \|C_{LT}\|_2 \\ \|C_{TL}\|_2 & 0 \end{bmatrix} \begin{bmatrix} \|x\|_2 \\ \|y\|_2 \end{bmatrix}. \tag{28}$$

Taking a supremum and bounding the norm of this nonnegative 2×2 matrix by its Frobenius norm proves

$$\|C - C_{\text{tail}}\|_2 \le \sqrt{\|C_{LL}\|_2^2 + \|C_{LT}\|_2^2 + \|C_{TL}\|_2^2}. \tag{29}$$

Substitution gives (21). If the stated factor norms remain bounded, its right-hand side goes to zero as $\mu, \nu \to 0$.

For the general core, let $C_{\text{diag}} = \operatorname{diag}(C_{LL}, C_{TT})$. The off-diagonal block matrix satisfies the exact identity

$$\|C - C_{\text{diag}}\|_2 = \max\{\|C_{LT}\|_2, \|C_{TL}\|_2\}. \tag{30}$$

The cross-block bounds therefore imply convergence to the set of block-diagonal updates when e, d_s, h_s, s are bounded and $\mu, \nu \to 0$. The term $A_{LL} S_L B_{LL}$ need not vanish. This proves Corollary 3 and states precisely the different limits in the two cases.

A.3 Confinement does not prevent singular-value crossings

For a tail-only update without ambient scaling,

$$U^\top (W_0 + \Delta) V = \begin{bmatrix} \Sigma_L & 0 \\ 0 & \Sigma_T + H \end{bmatrix}. \tag{31}$$

The singular values are the union of those of the two diagonal blocks, which proves the sufficient condition (6). Consider $W_0 = \operatorname{diag}(2, 1)$, cutoff $k = 1$, and $\Delta = \operatorname{diag}(0, 2)$. The update is strictly tail-supported with no cross mixing and no leading-block modification, but $W_0 + \Delta = \operatorname{diag}(2, 3)$ has the other coordinate as its top singular direction. Its largest-principal-angle sine relative to the original leading direction is one. Thus block decoupling and stability of the top-ranked subspace are distinct even in a two-dimensional example.

A.4 Rectangular weights and complete energy accounting

For $W_0 \in \mathbb{R}^{m\times n}$, select $k \le \min(m,n)$ leading singular vectors and define

$$P_L = U_L U_L^\top, \quad P_T = I_m - P_L, \qquad Q_L = V_L V_L^\top, \quad Q_T = I_n - Q_L. \tag{32}$$

Then $\Delta_{ab} = P_a \Delta Q_b$ gives

$$\Delta = \sum_{a,b} \Delta_{ab}, \qquad \|\Delta\|_F^2 = \sum_{a,b} \|\Delta_{ab}\|_F^2 . \tag{33}$$

Pairwise Frobenius inner products vanish because complementary projectors multiply to zero. The tail here is the full complement and includes any unmatched null directions. A chosen adapter may span only a subset of it; that available subspace must not be confused with the diagnostic complement. Complete orthogonal bases recover the corresponding coordinate blocks. The square attention-matrix experiment avoids this issue, but an economy-SVD diagnostic is not automatically exhaustive for other layers.

A.5 Representative methods in the fixed coordinate frame

These are algebraic consequences of update forms, not measurements on prior methods; Table 5 lists the placements. Let Q_J select the declared pretrained singular indices J, and write $A_a = U_a^\top A$, $B_b = BV_b$. Any additive update AB has $C_{ab} = A_a B_b$; low rank alone forces no LL/LT/TL/TT block to vanish. A learned sum of low-rank terms has the sum of these blocks. A fixed-basis core instead has $\Delta = UQ_J K Q_J^\top V^\top$ and $C = Q_J K Q_J^\top$, so only its selected support is reachable.

Table 5: Representative coordinate placements, relative to the original W_0. "All" means no block is structurally forced to zero, not that every fitted update populates every block. The formulas absorb any fixed adapter scaling into the coefficients.

Method/form	Effective update	Coordinates C	Permitted support
LoRA / AdaLoRA	AB / APB	$U^\top ABV$ / $U^\top APBV$	All
PiSSA / MiLoRA	$AB - A_0B_0$	$U^\top(AB - A_0B_0)V$	All after free-factor training
SVDiff	$U\,\mathrm{diag}(h)V^\top$	$\mathrm{diag}(h)$	LL, TT diagonal
SVFT	$UMV^\top$	M	Declared sparsity pattern
LoRA-XS	$U_J\Sigma_J R V_J^\top$	$Q_J\Sigma_J R Q_J^\top$	$J \times J$
Spectral Adapter (rotation)	$U_L(R_U\Sigma_L R_V^\top - \Sigma_L)V_L^\top$	leading core difference	LL
PSOFT	$A'(K_P - I)B'$	$Q_J(K_P - I)\Sigma_J Q_J^\top$	$J \times J$ (principal)
OFT / BOFT (pure left map)	$(R - I)W_0$	$(U^\top RU - I)\Sigma$	All
HRA (pure right map)	$W_0(R - I)$	$\Sigma(V^\top RV - I)$	All

For LoRA and AdaLoRA the low-rank factors are trainable (Hu et al., 2022; Zhang et al., 2023). PiSSA/MiLoRA's retained residual means displacement from the original checkpoint subtracts the initial product; the leading/minor initialization does not freeze the factor spans (Meng et al., 2024; Wang et al., 2025b). SVDiff changes singular values (Han et al., 2023); SVFT explicitly permits off-diagonal entries (Lingam et al., 2024). In LoRA-XS the published initialization includes Σ_J in the left fixed factor, hence $C_{JJ} = \Sigma_J R$, not literally R unless it is reparameterized (Bałazy et al., 2025). If J lies inside the declared leading set, every other diagnostic block is zero.

The Spectral Adapter entry describes its rotational leading-core variant (Zhang & Pilanci, 2024). OFT/BOFT and HRA use orthogonal R in the pure forms shown (Qiu et al., 2023; Liu et al., 2024b; Yuan et al., 2024); singular-value preservation is compatible with nonzero cross blocks. Left/right placement follows the stated weight convention. Additional learned magnitude scales, such as rescaled OFT, invalidate the pure orthogonal spectrum-preservation statement. These examples distinguish fixed support, a coefficient pattern, and an ambient isometry without equating any of them with preserved behavior. They do not establish cross-method empirical effects.

For PSOFT, $A' = U_J$, $B' = \Sigma_J V_J^\top$, and $K_P = \mathrm{diag}(\alpha) R\,\mathrm{diag}(\beta)$ act inside a fixed principal core; the residual of W_0 stays frozen (Wu et al., 2026). Thus $C_{JJ} = (K_P - I)\Sigma_J$, with zero support outside $J \times J$ even after vector relaxation. Its geometry condition $R^\top A^\top AR = A^\top A$ preserves column Gram geometry; choosing $A' = U_J$ reduces it to orthogonality. Our Corollary 3 instead

bounds support outside specified blocks. Neither property alone implies the other: a nonorthogonal core remains confined, while an ambient orthogonal map can mix blocks. Principal versus tail selection is a further independent choice.

GaLore projects gradients into low-rank subspaces that can change during training (Zhao et al., 2024); the sum of such updates need not remain inside one fixed pretrained span. GOFT/qGOFT use Givens rotations and relaxed angular/norm constraints (Ma et al., 2024), complementing the ambient-transform examples above. These differ from a fixed update-support constraint.

A.6 A DIMENSION-ONLY REFERENCE FOR ALLOCATION

Let the vectorized nonzero update have a rotationally invariant direction in the d^2-dimensional matrix space. By symmetry each squared coordinate has expected fraction $1/d^2$ of the total. The leading, cross, and tail blocks have k^2, $2kr$, and r^2 coordinates, respectively. Their expected energy fractions are therefore k^2/d^2, $2kr/d^2$, and r^2/d^2. For $r/d = 1/3$, these are $4/9$, $4/9$, and $1/9$, with off-tail expectation $8/9$. The square root $\sqrt{8/9} \simeq 0.9428$ is a reference on the norm-ratio scale, not the expectation $\mathbb{E}[q_{\text{off}}]$. These calculations are a null reference, not a fitted model or a statistical test of the observed updates.

A.7 THE SCALER PENALTY AS A PROJECTOR COMMUTATOR

In the complete U basis, P is $\operatorname{diag}(I_k, 0)$. Writing $A = U^\top E U$ gives

$$U^\top (EP - PE)U = \begin{bmatrix} 0 & -A_{LT} \\ A_{TL} & 0 \end{bmatrix}. \tag{34}$$

Because A is symmetric, orthogonal invariance yields $\|EP - PE\|_F^2 = 2\|A_{LT}\|_F^2$. In ambient coordinates, the (i, j) entry of the commutator is $(\epsilon_i - \epsilon_j)P_{ij}$. Squaring and combining its symmetric pairs proves (13). Equivalently, the penalty is $\epsilon^\top L_P \epsilon$, where L_P is the graph Laplacian with off-diagonal edge weights P_{ij}^2. Every summand is nonnegative. Zero penalty therefore means equal scales across every nonzero edge, or constant scales on each connected component. A connected graph has only globally constant scales in this nullspace. A disconnected projector permits different constants on disconnected components; for a coordinate-aligned projector, all off-diagonal weights vanish and every ambient diagonal scaler commutes with it. The result is conditional on the fixed pretrained projector, not a measured connectivity claim for the experiment.

A.8 RANDOM-PROJECTOR REFERENCE AND SCALER HOMOGENIZATION

Write $J_P(\epsilon) = \epsilon^\top L_P \epsilon$ for (13). Since $L_P \mathbf{1} = 0$, adding a constant to every scale does not change the penalty. Under the penalty alone, $\dot{\epsilon} = -2\lambda L_P \epsilon$ preserves $\bar{\epsilon}$ and damps each positive-eigenvalue component of L_P. On a connected graph it converges to $\bar{\epsilon}\mathbf{1}$. This statement concerns ordinary gradient flow of this isolated quadratic, not AdamW with a task loss; neither shrinkage of the mean nor convergence of the trained model follows from it.

For $d > 1$, a uniformly random rank-k orthogonal projector P in $\mathbb{R}^d$, and a fixed scaler vector independent of P,

$$\mathbb{E}_P[J_P(\epsilon)] = \frac{k(d-k)}{(d-1)(d+2)} \|\epsilon - \bar{\epsilon}\mathbf{1}\|_2^2. \tag{35}$$

To verify the constant, write $P = QQ^\top$ for the first k columns of a Haar orthogonal matrix. The spherical fourth moments give $\mathbb{E}[P_{ii}^2] = k(k+2)/(d(d+2))$. Since $\sum_{ij} P_{ij}^2 = k$, symmetry gives $\mathbb{E}[P_{ij}^2] = k(d-k)/(d(d-1)(d+2))$ for $i \neq j$. Substitute into (13) and use $\sum_{i<j}(\epsilon_i - \epsilon_j)^2 = d\|\epsilon - \bar{\epsilon}\mathbf{1}\|_2^2$.

This expectation motivates a centered-scaler control. It does not equate one fixed random projector with the pretrained projector: their positive-eigenvalue directions and strengths generally differ. It also cannot be substituted into training when the learned scalers depend on the sampled projector. A shared scalar nullspace alone does not imply identical trajectories or endpoints.

A.9 FACTOR-SCALE IDENTITIES

Write the inner matrix in (7) as K. For any $a > 0$, $(aE)K(D/a) = EKD$. Because each mixing operator is linear in its respective scaler, this transformation multiplies its two magnitudes by a and $1/a$. For $S_L = 0$, the additional identity

$$(aE)U_T(H/a^2)V_T^\top(aD) = EU_THV_T^\top D \tag{36}$$

shrinks both squared Frobenius penalties by a^2. It can make a raw factor penalty arbitrarily small with a fixed effective update if the core is allowed to diverge. This rescaling illustrates why the confinement bound assumes bounded factors. Measuring the effective update and the factor norms checks both sides of that statement.

Even bounded factors do not turn an absolute cross-block bound into a bound on normalized cross share without a lower bound on the total update norm. For an explicit practical-core example, let U rotate by $\pi/4$, $V = I$, $S_L = 1$, $H = 0$, $E = a\,\mathrm{diag}(2, 1)$ and $D = aI$, for $a > 0$. Then

$$C = a^2\begin{bmatrix} 3/2 & 0 \\ -1/2 & 0 \end{bmatrix}, \qquad \|M_E\|_F^2 + \|M_D\|_F^2 = a^2/4, \qquad p_{\mathrm{cross}} = 1/10. \tag{37}$$

As $a \to 0$, the penalty and absolute cross energy vanish while the share stays fixed. This is a family of parameter states, not a trajectory of the isolated mean-preserving flow in Appendix A.8. It explains why the normalized outcome under joint task training must be measured.

A.10 MAGNITUDE FLOW IN EFFECTIVE AND FACTOR COORDINATES

For a freely optimized effective update $x = \mathrm{vec}(\Delta)$, isolated gradient flow of $\beta\|x\|_2^2$ is radial, $\dot{x} = -2\beta x$, and preserves block fractions. If $x = f(\theta)$ is parameterized by adapter factors with Jacobian J_θ, ordinary parameter gradient flow instead induces

$$\dot{x} = -2\beta J_\theta J_\theta^\top x. \tag{38}$$

This need not be radial. A task objective or adaptive optimizer introduces further changes. For an explicit penalty-only example, fix $E = \mathrm{diag}(2, 1)$, $D = V = I$, let U rotate by $\pi/6$, fix $S_L = 1$ and train only $H = h$. Then

$$C = \begin{bmatrix} 7/4 & -\sqrt{3}h/4 \\ -\sqrt{3}/4 & 5h/4 \end{bmatrix}, \quad \|C\|_F^2 = (13 + 7h^2)/4, \quad p_{\mathrm{cross}} = \frac{3(1 + h^2)}{4(13 + 7h^2)}. \tag{39}$$

Magnitude-only flow drives h toward zero and changes p_{cross} (0.075 at $h = 1$, approaching $3/52$ at zero). Thus orthogonal invariance of the norm objective does not force a constant trained cross share. On the MIX side, connected-projector, penalty-only homogenization predicts a scalar limit; finite task-trained cross shares and identity residuals remain empirical quantities.

B PRIMARY COHORTS AND EVALUATION PROTOCOLS

This section collects the protocols and outcomes for the encoder and decoder cohorts, with their separately labelled post hoc additions. The RTE band/head study's later held-out evaluation uses the same frozen endpoints and remains distinct from its inner-selection scores. Appendix C groups calibration, geometry and uncertainty checks, including the earlier Qwen1.5B seed extension. Appendix D reports computational costs. Historical protocols and comparisons are collected last in Appendices E–G; they do not enter the primary comparisons.

B.1 METRIC DEFINITIONS AND AGGREGATION

For each adapted matrix, the reported metrics are

$$\rho_F = \|\Delta\|_F \,/\, \|W_0\|_F\,, \qquad q_{\mathrm{off}} = \|C_{\mathrm{off}}\|_F \,/\, \|C\|_F\,, \tag{40}$$

$$\mathrm{Leak}_{LL} = \|C_{LL}\|_2 \,/\, \|\Delta\|_2\,, \qquad C_{\mathrm{off}} = \begin{bmatrix} C_{LL} & C_{LT} \\ C_{TL} & 0 \end{bmatrix}. \tag{41}$$

The mixing magnitudes are operator norms, while their penalties use squared Frobenius norms. Leading left and right drift use the largest-principal-angle sine between the leading subspaces of W_0 and $W_0 + \Delta$. The archived tables average each metric across the 48 modules within a run, then across seeds; they do not pool module energies before normalization.

The new fixed-endpoint study instead reports

$$\rho_{F,\mathrm{pool}} = \sqrt{\frac{\sum_\ell \|\Delta_\ell\|_F^2}{\sum_\ell \|W_{0,\ell}\|_F^2}}, \qquad p_{ab,\mathrm{pool}} = \frac{\sum_\ell \|C_{ab,\ell}\|_F^2}{\sum_\ell \|\Delta_\ell\|_F^2}. \tag{42}$$

Its equal-module comparison averages the individual $p_{ab,\ell}$ instead. A pooled fraction weights modules by update energy; an equal-module fraction does not. NORM's training objective averages squared module-relative norms, whereas calibration targets $\rho_{F,\mathrm{pool}}$. Neither matching a pooled norm nor rescaling it matches each module. Zero-update fractions are undefined. In both protocols, seeds, not modules, are the replication units.

B.2 Executed fresh-protocol settings and input validation

The fresh study uses the pinned RoBERTa-base checkpoint `e2da8e2f811d1448a5b465c236feacd80ffbac7b` and GLUE dataset revision `bcdcba79d07bc864c1c254ccfcedcce55bcc9a8c`. A fixed 20% inner-selection split is drawn from the training data with seed 271828; the official RTE and MRPC validation sets (277 and 408 examples) were held aside throughout training and selection, then evaluated once after the 18 checkpoints were frozen (Appendix B.6). Inspecting tokenization parity on held-aside input text is not a task-score evaluation. No model or checkpoint was selected using those held-aside labels.

Executed manifests specify float32 training, microbatch 8 with four accumulation steps (effective batch 32), maximum sequence length 128, gradient-norm clipping at 1, and zero weight decay. AdamW uses separate adapter and task-head groups with linear learning-rate decay to zero and no warmup. The adapter rate is 10^{-2} on both tasks; head rates are 5×10^{-4} on RTE and 10^{-3} on MRPC. Fixed endpoints are 5,670 and 2,760 optimizer steps, respectively. The scaled leading-plus-tail adapter initializes both ambient scalers and tail coefficients at 0.01 for RTE and 0.1 for MRPC, with an identity leading block and no rotations. These initialization values belong to the training-time interaction study, not the strict-band branch. Task evaluation uses batches of 16; probe evaluation uses batches of 8. The secondary checkpoint maximizes inner-selection accuracy, excludes step 0, and breaks ties toward the earliest step. Candidates occur every 128 steps and at the registered fractions $\{0, .01, .05, .1, .25, .5, .75, 1\}$ of the budget. The fixed endpoint remains primary regardless of the selected score.

Before calibration, a 32-example, 200-step overfit gate required training accuracy at least 0.95; a 600-step unregularized learning gate required inner accuracy at least three points above the majority baseline. RTE passed at the adopted rates. On MRPC, the historical adapter rate 0.05 reached the majority baseline (0.6744) and failed the second gate; the adopted common rate 0.01 reached 0.8474. This change preceded the comparisons and applied to every fresh MRPC arm. It does not revise the historical results.

The tokenizer repair explicitly restores the BPE model type omitted from its upstream serialization. An independent byte-level BPE implementation exactly reproduced input IDs and attention masks for 181 examples across both tasks and all prepared splits, including 51 truncated examples. Both durable invalidation records are enforced during extraction: all 39 runs from the corrupted-input campaigns are excluded. Run manifests bind source files, inputs, recipes and validation reports; an exported validation hash identifies the report, while independent reaggregation here audits the export rather than rerunning training.

The frozen-head probe contains 256 sampled examples from WikiText-2-raw-v1's test split (Merity et al., 2017), corpus revision `b08601e04326c79dfdd32d625aee71d232d685c3`. Sampling seed 271828 and masking seed 161803 fix the examples and 15% masking procedure. Summing per-example loss numerators and dividing by the 4,001 scored tokens gives original-backbone cross-entropy 2.095626. This reference uses CPU float32; adapted probes use GPU float32. The corpus is held out from task adaptation, but is not guaranteed unseen during pretraining. The probe measures behavior under fixed examples and masks, not preserved knowledge in general.

B.3 TRAINING-TIME INTERACTION ON COMMONSENSEQA

MIX lowers cross share and raises choice loss. This post hoc Qwen2.5-1.5B-Instruct study compares UNREG, MIX and NORM in the scaled leading-plus-tail coefficient adapter. MIX's cross share is lower than both controls, while its normalized five-choice NLL is higher. All three between-arm accuracy intervals include zero (Tables 6 and 10). These unresolved differences do not establish equivalence. The size control misses its registered tolerance, so the contrast combines changes in update size, allocation and geometry.

All arms use one fixed recipe and numerical path. We adapt the same 56 query/output matrices of width 1536, with tail width 512 and leading identity. The scaled leading-plus-tail adapter starts with nonzero coefficients and scalers of 0.01; frozen and inserted-untrained references therefore remain distinct. Each arm uses seeds 17, 42 and 123, 842 AdamW updates, learning rate 10^{-3} with linear decay, effective batch size 16 (2 examples accumulated over 8 batches), maximum length 160, no weight decay or warmup, and gradient clipping at 1. Only the answer-label token contributes to full-vocabulary training cross-entropy. Evaluation uses the common unmerged native forward with bfloat16 autocast over float32 weights, batch size 8, and all 1,221 labeled public-validation examples. This is not the official unlabeled test split. The 974-example inner split supports checks; the fixed endpoint is reloaded before final scoring. This study was specified after earlier interaction results were examined and remains separate from the float32 saved-checkpoint interventions.

Calibration chooses the nearest dose without matching size. MIX fixes the coefficient at 10^{-3} per side for the summed projector/scaler commutator penalty; it does not directly penalize measured cross-block energy. NORM penalizes the module mean of $\|\Delta W\|_F^2/\|W_0\|_F^2$. Calibration uses seed 31415 and pooled update norm alone, with no task-score or cross-share selection. The finite procedure comprises one MIX target and five NORM doses, including two rule-based refinements (Table 7). None meets the 5% tolerance. The nearest coefficient, 1000, leaves NORM 27.5% larger in calibration and 27.2% larger on average in confirmation. We retain this failed match as an imperfect control, without another search.

Table 6: Practical CommonsenseQA interaction comparison: mean $\pm$ sample SD over three seeds; references are single fixed states. Label NLL uses the full vocabulary; label mass is the probability on A–E. ρ_F and Cross describe total updates including insertion. Frozen cross share is undefined.

Arm	Accuracy (%)	Choice NLL	Label NLL	Label mass	ρ_F	Cross (%)
UNREG	74.67 ± 0.05	0.6745 ± 0.0011	0.6753 ± 0.0012	0.999196 ± 0.000010	0.006014 ± 0.000044	40.17 ± 0.05
MIX	74.50 ± 0.05	0.6889 ± 0.0005	0.6895 ± 0.0005	0.999342 ± 0.000008	0.003130 ± 0.000027	32.47 ± 0.08
NORM	74.69 ± 0.14	0.6786 ± 0.0015	0.6794 ± 0.0015	0.999264 ± 0.000018	0.003981 ± 0.000041	40.84 ± 0.07
Frozen	70.60	1.9129	1.9129	0.999989	0.000000	–
Inserted	70.76	1.9116	1.9116	0.999988	0.000073	0.00

Table 7: Complete six-run calibration frontier, seed 31415 and 842 updates. Signed mismatch is $100(\rho_F/\rho_F^{\mathrm{MIX}} - 1)$. NORM doses $0.01, 1, 100$ preceded refinements 1000 and 10000; the nearest dose is retained despite failed matching.

Arm	Coefficient	ρ_F	Mismatch (%)	Disposition
MIX	0.001	0.0031076	–	Target
NORM	0.01	0.0060571	+94.91	Outside tolerance
NORM	1	0.0060409	+94.39	Outside tolerance
NORM	100	0.0053988	+73.73	Outside tolerance
NORM	1000	0.0039622	+27.50	Selected; outside tolerance
NORM	10000	0.0019865	-36.08	Outside tolerance

Total and learned updates remain distinct. Table 9 reports each seed's total update $W_{\mathrm{final}} - W_0$ and change since insertion $W_{\mathrm{final}} - W_{\mathrm{inserted}}$. The initial pooled relative norm is 7.32849×10^{-5} and its cross share is numerically zero. Initial learned change is only numerical roundoff, so we do not interpret its block fractions. The accompanying `geometry.json` retains all 56 matrices' initial/final block energies, fractions, reconstruction checks and scaler summaries for every calibration

Table 8: All nine confirmation runs. Choice accuracy and normalized five-choice NLL are co-primary; full-vocabulary label NLL and label mass are retained supporting measurements.

Arm	Seed	Accuracy (%)	Choice NLL	Label NLL	Label mass
UNREG	17	74.6929	0.675295	0.676106	0.99918958
UNREG	42	74.6110	0.673225	0.674018	0.99920704
UNREG	123	74.6929	0.675114	0.675923	0.99919106
MIX	17	74.4472	0.689265	0.689929	0.99933666
MIX	42	74.5291	0.689061	0.689710	0.99935136
MIX	123	74.5291	0.688300	0.688962	0.99933867
NORM	17	74.6110	0.678254	0.679011	0.99924396
NORM	42	74.8567	0.677363	0.678084	0.99927992
NORM	123	74.6110	0.680249	0.680983	0.99926736

and confirmation state. Confirmation NORM/MIX size mismatches are 27.42%, 26.73% and 27.38% for seeds 17, 42 and 123.

Table 9: Pooled update geometry for every confirmation seed. Block fractions and Cross are percentages of squared update norm; Cross is LT+TL. Learned subtracts insertion, whereas total includes it. These describe attained geometry and do not establish a training mechanism.

Arm	Seed	Update	ρ_F	LL	LT	TL	TT	Cross
UNREG	17	Total	0.0060589	52.116	21.346	18.794	7.745	40.139
UNREG	17	Learned	0.0060528	52.019	21.389	18.832	7.761	40.220
UNREG	42	Total	0.0060143	51.973	21.415	18.809	7.803	40.224
UNREG	42	Learned	0.0060082	51.876	21.458	18.847	7.819	40.305
UNREG	123	Total	0.0059699	52.115	21.377	18.763	7.745	40.140
UNREG	123	Learned	0.0059638	52.016	21.421	18.802	7.761	40.223
MIX	17	Total	0.0031610	62.947	18.021	14.537	4.495	32.557
MIX	17	Learned	0.0031397	62.444	18.266	14.734	4.556	33.000
MIX	42	Total	0.0031190	63.105	17.901	14.509	4.485	32.410
MIX	42	Learned	0.0030975	62.590	18.151	14.711	4.548	32.862
MIX	123	Total	0.0031099	63.090	18.002	14.444	4.464	32.446
MIX	123	Learned	0.0030884	62.575	18.254	14.646	4.526	32.899
NORM	17	Total	0.0040276	51.073	22.477	18.314	8.136	40.791
NORM	17	Learned	0.0040215	50.923	22.546	18.371	8.161	40.916
NORM	42	Total	0.0039527	50.847	22.604	18.321	8.227	40.926
NORM	42	Learned	0.0039466	50.696	22.674	18.378	8.253	41.052
NORM	123	Total	0.0039615	51.036	22.531	18.285	8.149	40.815
NORM	123	Learned	0.0039553	50.882	22.601	18.342	8.175	40.943

Uncertainty is paired over training seeds. Tables 10–11 retain all 45 metric contrasts, including their three paired differences. Each interval is a nominal, unadjusted two-sided t interval with two degrees of freedom; examples are not treated as independent training replicates. Reference contrasts subtract one common frozen or inserted state from each trained seed. MIX–NORM is the primary treatment contrast; other treatment and reference contrasts are supporting. This comparison does not establish accuracy equivalence, matched-size causality, or compensation by retraining relative to the different saved-checkpoint adapter families.

B.4 Magnitude control: calibration and confirmation

The study of Section 5.2 follows registered protocols with file hashes. Each included run passed checkpoint reload and reproduction of task metrics, spectral diagnostics and the frozen-head probe. The common recipe uses sequence length 128, effective batch 32, an inner training/selection split, and fixed endpoints of 5,670 steps for RTE and 2,760 for MRPC. Calibration uses seed 31415; confirmation uses seeds 17, 42 and 123.

Table 10: All between-training-arm metric contrasts, first minus second. Accuracy and cross-share units are percentage points; scaled units are indicated. Nominal 95% paired intervals use three seeds ($df = 2$).

Contrast	Metric	Seed 17	Seed 42	Seed 123	Mean [95%]
MIX–NORM	Accuracy (pp)	-0.16380	-0.32760	-0.08190	$-0.19110\ [-0.50188, +0.11968]$
	Choice NLL	+0.01101	+0.01170	+0.00805	$+0.01025\ [+0.00544, +0.01507]$
	Label NLL	+0.01092	+0.01163	+0.00798	$+0.01017\ [+0.00537, +0.01498]$
	Label mass (10^{-4})	+0.92697	+0.71440	+0.71317	$+0.78485\ [+0.47909, +1.09060]$
	Total ρ_F (10^{-3})	-0.86662	-0.83365	-0.85151	$-0.85059\ [-0.89160, -0.80959]$
	Learned ρ_F (10^{-3})	-0.88173	-0.84913	-0.86681	$-0.86589\ [-0.90643, -0.82535]$
	Cross (pp)	-8.23402	-8.51583	-8.36954	$-8.37313\ [-8.72325, -8.02302]$
MIX–UNREG	Accuracy (pp)	-0.24570	-0.08190	-0.16380	$-0.16380\ [-0.36725, +0.03965]$
	Choice NLL	+0.01397	+0.01584	+0.01319	$+0.01433\ [+0.01095, +0.01771]$
	Label NLL	+0.01382	+0.01569	+0.01304	$+0.01418\ [+0.01080, +0.01757]$
	Label mass (10^{-4})	+1.47076	+1.44325	+1.47615	$+1.46339\ [+1.41955, +1.50722]$
	Total ρ_F (10^{-3})	-2.89788	-2.89525	-2.85997	$-2.88437\ [-2.93694, -2.83179]$
	Learned ρ_F (10^{-3})	-2.91303	-2.91075	-2.87533	$-2.89970\ [-2.95221, -2.84719]$
	Cross (pp)	-7.58184	-7.81397	-7.69420	$-7.69667\ [-7.98503, -7.40831]$
NORM–UNREG	Accuracy (pp)	-0.08190	+0.24570	-0.08190	$+0.02730\ [-0.44255, +0.49715]$
	Choice NLL	+0.00296	+0.00414	+0.00514	$+0.00408\ [+0.00137, +0.00678]$
	Label NLL	+0.00291	+0.00407	+0.00506	$+0.00401\ [+0.00133, +0.00669]$
	Label mass (10^{-4})	+0.54379	+0.72885	+0.76298	$+0.67854\ [+0.38556, +0.97152]$
	Total ρ_F (10^{-3})	-2.03125	-2.06160	-2.00846	$-2.03377\ [-2.10000, -1.96755]$
	Learned ρ_F (10^{-3})	-2.03130	-2.06163	-2.00852	$-2.03381\ [-2.10000, -1.96763]$
	Cross (pp)	+0.65218	+0.70186	+0.67535	$+0.67646\ [+0.61470, +0.73823]$

Table 11: All trained-minus-reference task contrasts. Each reference is a common fixed state. Accuracy differences are percentage points; label-mass differences are scaled by 10^4. Nominal intervals use the three training seeds ($df = 2$).

Contrast	Metric	Seed 17	Seed 42	Seed 123	Mean [95%]
UNREG–frozen	Accuracy (pp)	+4.09500	+4.01310	+4.09500	$+4.06770\ [+3.95024, +4.18517]$
	Choice NLL	-1.23764	-1.23971	-1.23782	$-1.23839\ [-1.24124, -1.23554]$
	Label NLL	-1.23684	-1.23893	-1.23702	$-1.23760\ [-1.24047, -1.23473]$
	Label mass (10^{-4})	-7.99199	-7.81749	-7.97726	$-7.92891\ [-8.16932, -7.68850]$
MIX–frozen	Accuracy (pp)	+3.84930	+3.93120	+3.93120	$+3.90390\ [+3.78644, +4.02137]$
	Choice NLL	-1.22367	-1.22387	-1.22464	$-1.22406\ [-1.22532, -1.22280]$
	Label NLL	-1.22302	-1.22324	-1.22399	$-1.22341\ [-1.22467, -1.22215]$
	Label mass (10^{-4})	-6.52123	-6.37423	-6.50111	$-6.46552\ [-6.66350, -6.26755]$
NORM–frozen	Accuracy (pp)	+4.01310	+4.25880	+4.01310	$+4.09500\ [+3.74262, +4.44739]$
	Choice NLL	-1.23468	-1.23557	-1.23269	$-1.23431\ [-1.23798, -1.23064]$
	Label NLL	-1.23394	-1.23486	-1.23196	$-1.23359\ [-1.23727, -1.22991]$
	Label mass (10^{-4})	-7.44820	-7.08864	-7.21428	$-7.25037\ [-7.70368, -6.79707]$
UNREG–inserted	Accuracy (pp)	+3.93120	+3.84930	+3.93120	$+3.90390\ [+3.78644, +4.02137]$
	Choice NLL	-1.23634	-1.23841	-1.23652	$-1.23709\ [-1.23994, -1.23424]$
	Label NLL	-1.23554	-1.23763	-1.23572	$-1.23630\ [-1.23917, -1.23343]$
	Label mass (10^{-4})	-7.98853	-7.81403	-7.97380	$-7.92546\ [-8.16587, -7.68505]$
MIX–inserted	Accuracy (pp)	+3.68550	+3.76740	+3.76740	$+3.74010\ [+3.62264, +3.85757]$
	Choice NLL	-1.22237	-1.22258	-1.22334	$-1.22276\ [-1.22402, -1.22150]$
	Label NLL	-1.22172	-1.22194	-1.22269	$-1.22211\ [-1.22337, -1.22085]$
	Label mass (10^{-4})	-6.51777	-6.37078	-6.49765	$-6.46207\ [-6.66005, -6.26409]$
NORM–inserted	Accuracy (pp)	+3.84930	+4.09500	+3.84930	$+3.93120\ [+3.57882, +4.28359]$
	Choice NLL	-1.23338	-1.23427	-1.23139	$-1.23301\ [-1.23669, -1.22934]$
	Label NLL	-1.23264	-1.23356	-1.23067	$-1.23229\ [-1.23597, -1.22861]$
	Label mass (10^{-4})	-7.44475	-7.08518	-7.21082	$-7.24692\ [-7.70022, -6.79361]$

Table 12: Confirmation of the calibration-selected norm control across seeds 17, 42, and 123 (mean±sample SD, $n = 3$), fixed-step endpoints, common protocol of Section 5.2. Cross P, LL, and TT are pooled energy percentages; Cross E weights modules equally. The probe is masked-token cross-entropy under the frozen original MLM head; accuracy is the inner-selection split at the fixed step, scaled by 100. Achieved errors $\delta_s = \rho_{\mathrm{NORM},s}/\rho_{\mathrm{MIX},s} - 1$ (signed; $e_s = |\delta_s|$) against the declared $\pm 5\%$ rule, in that seed order: RTE: $+8.0, -5.2, +7.5\%$; MRPC: $+5.5, +6.7, -0.1\%$. N is the dimension-only reference, not a run or test of isotropy.

Task	Condition	ρ_F	Cross P	Cross E	LL	TT	Probe CE	Acc
RTE	UNREG	0.2751 ± 0.0041	43.7 ± 0.0	43.5 ± 0.0	45.9 ± 0.1	10.4 ± 0.0	2.89 ± 0.17	72.4 ± 0.6
RTE	MIX	0.0736 ± 0.0046	10.0 ± 0.5	14.9 ± 0.9	45.2 ± 0.9	44.8 ± 1.4	2.94 ± 0.29	72.6 ± 3.1
RTE	NORM ($\beta = 7.5$)	0.0759 ± 0.0016	46.2 ± 0.0	46.3 ± 0.1	40.7 ± 0.0	13.1 ± 0.0	2.61 ± 0.08	74.6 ± 0.5
MRPC	UNREG	0.1802 ± 0.0009	42.4 ± 0.1	42.2 ± 0.0	48.0 ± 0.1	9.6 ± 0.1	2.75 ± 0.07	84.9 ± 0.2
MRPC	MIX	0.0407 ± 0.0017	4.5 ± 0.3	5.5 ± 0.1	69.8 ± 1.1	25.8 ± 1.4	2.84 ± 0.17	84.5 ± 0.5
MRPC	NORM ($\beta = 27$)	0.0423 ± 0.0005	46.2 ± 0.1	46.2 ± 0.1	40.6 ± 0.2	13.2 ± 0.1	2.57 ± 0.11	84.8 ± 0.2
—	N (dimension only)	—	44.4	44.4	44.4	11.1	—	—

The registered protocol fixes $\lambda_E = \lambda_D = 10^{-3}$ on both tasks, the same coefficients as the archived two-sided condition. The matching search varies only NORM's coefficient; it does not select MIX's dose from confirmation outcomes.

The dose-selection rule targets MIX's pooled relative Frobenius norm of the total effective update, including insertion. It selects the smallest absolute relative error, with ties toward the lower dose, using validated fixed-step norms. The target tolerance is $\pm 5\%$. The initial grid is $\beta \in \{0.01, 1, 100\}$ per task. A separately registered norms-only refinement adds bracketing geometric means and log-linear-fit-guided doses. The export retains every tried dose, immutable selection records and per-module norms. For each confirmation seed, the reported error is $e_s = |\rho_{\mathrm{NORM},s}/\rho_{\mathrm{MIX},s} - 1|$.

Table 13: Complete Frobenius-control dose frontier at the fixed-step endpoint, calibration seed 31415. Error is the absolute relative deviation of the pooled norm from the mixing arm's target (RTE 0.0792, MRPC 0.0413); Cross and TT are pooled energy percentages; the probe is masked-token cross-entropy. Every tried dose is shown; none is omitted.

Task	β	ρ_F	Error (%)	Cross (%)	TT (%)	Probe CE
RTE	0.01	0.2766	249.1	43.7	10.4	2.98
RTE	1	0.1621	104.6	44.4	11.1	2.62
RTE	4	0.0976	23.2	45.5	12.2	2.58
RTE	5.5	0.0856	8.0	45.7	12.6	2.54
RTE	7.5	0.0760	4.1	46.1	13.0	2.55
RTE	10	0.0654	17.4	46.5	13.4	2.67
RTE	100	0.0229	71.2	50.9	19.0	2.79
MRPC	0.01	0.1793	334.1	42.4	9.8	2.59
MRPC	1	0.1380	234.2	42.8	10.1	2.53
MRPC	10	0.0667	61.6	44.8	11.8	2.55
MRPC	14	0.0575	39.2	45.3	12.4	2.60
MRPC	19.5	0.0498	20.6	45.7	12.9	2.57
MRPC	27	0.0422	2.3	46.3	13.5	2.71
MRPC	100	0.0212	48.7	48.6	17.1	2.77

Table 13 lists every tried dose. Calibration and confirmation rows are stored separately. Table 14 reports all fixed and validation-selected confirmation outcomes on the inner-selection split. The original-backbone probe reference, masked-token cross-entropy 2.0956, uses the frozen original language-model head on a fixed 256-example set. The scaled leading-plus-tail adapter's inserted state is recorded separately.

Table 14: All confirmation task outcomes on the inner-selection split, scaled by 100. Fixed steps are 5,670 (RTE) and 2,760 (MRPC). Selected checkpoints are secondary, selection-conditioned outcomes on the same split; the main comparisons use fixed endpoints. F1 is shown for MRPC only.

Task	Arm	Seed	Fixed acc.	Fixed F1	Selected step	Selected acc.	Selected F1
RTE	UNREG	17	71.69	—	1920	73.49	—
RTE	UNREG	42	72.89	—	1280	75.70	—
RTE	UNREG	123	72.49	—	1152	76.10	—
RTE	MIX	17	75.90	—	4224	76.10	—
RTE	MIX	42	71.89	—	4253	73.49	—
RTE	MIX	123	69.88	—	2816	73.69	—
RTE	NORM	17	74.90	—	1418	75.50	—
RTE	NORM	42	74.90	—	4253	75.90	—
RTE	NORM	123	74.10	—	1408	76.51	—
MRPC	UNREG	17	84.74	89.04	2176	85.42	89.29
MRPC	UNREG	42	85.01	89.26	512	85.15	89.11
MRPC	UNREG	123	85.01	89.07	2304	85.42	89.42
MRPC	MIX	17	84.47	88.95	2760	84.47	88.95
MRPC	MIX	42	84.06	88.45	1280	85.29	89.07
MRPC	MIX	123	85.01	89.19	2070	85.69	89.51
MRPC	NORM	17	84.88	89.06	1792	85.83	89.60
MRPC	NORM	42	84.88	89.06	640	85.29	89.18
MRPC	NORM	123	84.60	88.78	1152	85.56	89.65

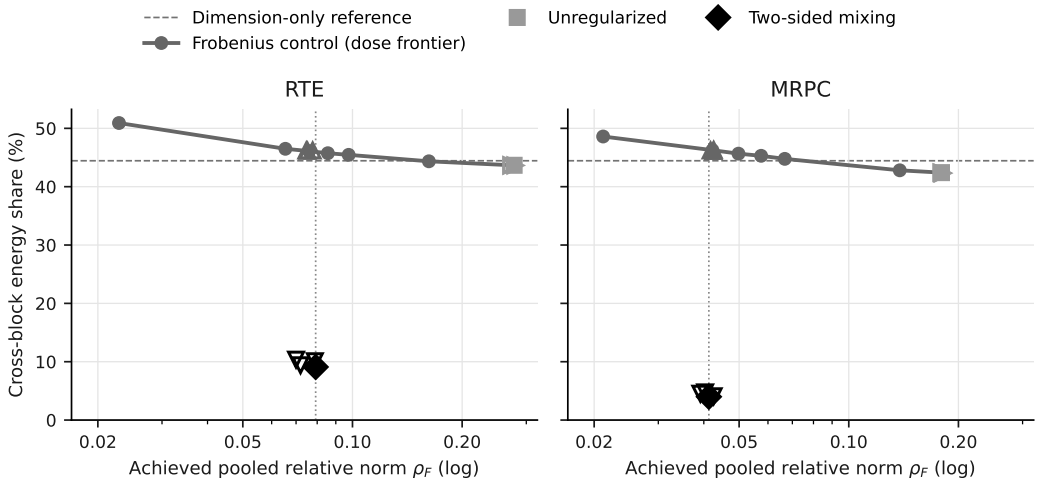


Figure 3: Interaction share across the tested size-penalty sweep. Filled circles show NORM calibration doses, squares UNREG and diamonds MIX. Open triangles show individual confirmation runs. The horizontal line is the $4/9$ dimension-only reference; the vertical line marks the MIX calibration size. Shrinking the NORM update leaves a large interaction share.

B.5 PAIRED CONFIRMATION CONTRASTS

Table 15 gives each seed difference, its mean and sample SD, and a nominal 95% paired Student-t interval ($n = 3$, two degrees of freedom). The intervals assume independent, approximately normal seed differences and describe variation on the fixed inner-selection split. They are unadjusted for multiple contrasts. For MIX–NORM accuracy, interval half-widths are 6.79 points on RTE and 1.55 on MRPC. Appendix C.8 separately calculates hypothetical power-based effect thresholds. Training seeds provide replication; modules and checkpoints describe each run.

Table 15: Paired fixed-endpoint differences (percentage points), left arm minus right. All three pairwise contrasts are reported for cross-block share and accuracy.

Task	Contrast	Metric	Seed 17	Seed 42	Seed 123	Mean±SD	Nominal 95% CI
RTE	MIX–NORM	Cross	-36.80	-36.11	-35.78	-36.23 ± 0.52	$[-37.51, -34.95]$
RTE	MIX–NORM	Accuracy	1.00	-3.01	-4.22	-2.07 ± 2.73	$[-8.87, 4.72]$
RTE	MIX–UNREG	Cross	-34.24	-33.52	-33.33	-33.70 ± 0.48	$[-34.89, -32.51]$
RTE	MIX–UNREG	Accuracy	4.22	-1.00	-2.61	0.20 ± 3.57	$[-8.67, 9.07]$
RTE	NORM–UNREG	Cross	2.55	2.59	2.45	2.53 ± 0.07	$[2.35, 2.71]$
RTE	NORM–UNREG	Accuracy	3.21	2.01	1.61	2.28 ± 0.84	$[0.20, 4.35]$
MRPC	MIX–NORM	Cross	-41.55	-41.68	-42.02	-41.75 ± 0.25	$[-42.36, -41.14]$
MRPC	MIX–NORM	Accuracy	-0.41	-0.82	0.41	-0.27 ± 0.62	$[-1.82, 1.28]$
MRPC	MIX–UNREG	Cross	-37.66	-37.80	-38.24	-37.90 ± 0.30	$[-38.65, -37.15]$
MRPC	MIX–UNREG	Accuracy	-0.27	-0.95	0.00	-0.41 ± 0.49	$[-1.63, 0.81]$
MRPC	NORM–UNREG	Cross	3.89	3.88	3.79	3.85 ± 0.06	$[3.71, 3.99]$
MRPC	NORM–UNREG	Accuracy	0.14	-0.14	-0.41	-0.14 ± 0.27	$[-0.81, 0.54]$

B.6 ONE-TIME EVALUATION ON HELD-ASIDE TASK SPLITS

After freezing all 18 fixed-endpoint confirmation checkpoints, we evaluated them once on the official RTE (277 examples) and MRPC (408 examples) validation sets. The frozen recipes, doses and checkpoints were retained throughout. CPU loading reproduced inner-selection accuracy and MRPC F1 exactly, with task loss agreeing within the recorded numerical tolerance. Exported per-example IDs, labels, predictions and losses reproduce the held-out metrics. Checkpoint hashes and validation-report identities link every row to the frozen population.

Held-out accuracies for UNREG/MIX/NORM are $71.72 \pm 1.50/71.36 \pm 1.63/71.24 \pm 1.63$ on RTE and $86.19 \pm 0.93/85.46 \pm 1.02/85.78 \pm 0.65$ on MRPC. MRPC F1 is $90.12 \pm 0.65/89.68 \pm 0.62/89.87 \pm 0.46$. Table 16 gives every endpoint; Table 17 gives every paired contrast. The original inner-selection outcomes are retained in Appendix B.22. Intervals use the three-seed assumptions in Appendix B.5 and are unadjusted for multiple contrasts.

Across the three confirmation seeds, MIX's norm coefficients of variation are about 6.2% on RTE and 4.2% on MRPC. This variation is comparable to the calibration rule's 5% tolerance. The confirmation report therefore retains every achieved error alongside the original doses and tolerance. Using several calibration seeds is a direct extension of this protocol.

B.7 STRICTLY CONFINED ADAPTATION: COMPLETE RTE RESULTS

Training and evaluation protocol. We use the pretrained RoBERTa-base encoder and a newly initialized two-layer classification head that maps its sentence-pair representation to two label scores. Softmax converts these scores to probabilities. All conditions train this head; only adapter conditions also train coefficients and, where enabled, rotations. The original encoder weights remain frozen. A fixed 20% split of RTE's training set, drawn with seed 271828, supplies the inner-selection examples; training uses the remaining 80%. Conditions share the head initialization and data order within each seed. All train for 5,670 AdamW steps with effective batch size 32, adapter learning rate 10^{-2}, head learning rate 5×10^{-4}, linear decay, no warmup, no weight decay, and no interaction penalty. Other input and optimizer settings follow Appendix B.2; the zero-update initialization below replaces the scaled leading-plus-tail adapter's initialization. We report accuracy and cross-entropy on the inner-selection split at the fixed final step, not the best-scoring checkpoint. After the original report,

Table 16: All frozen confirmation endpoints on the held-aside official splits. Accuracy and F1 are percentages; loss is task cross-entropy. These scores were not used for selection.

Task	Arm	Seed	Accuracy	F1	Loss
RTE	UNREG	17	70.04	—	2.347
RTE	UNREG	42	72.20	—	2.461
RTE	UNREG	123	72.92	—	2.478
RTE	MIX	17	69.68	—	1.330
RTE	MIX	42	71.48	—	1.192
RTE	MIX	123	72.92	—	1.165
RTE	NORM	17	69.68	—	2.567
RTE	NORM	42	72.92	—	1.956
RTE	NORM	123	71.12	—	2.057
MRPC	UNREG	17	85.78	89.75	0.872
MRPC	UNREG	42	85.54	89.74	0.938
MRPC	UNREG	123	87.25	90.88	0.802
MRPC	MIX	17	84.31	88.97	0.425
MRPC	MIX	42	85.78	89.97	0.370
MRPC	MIX	123	86.27	90.11	0.388
MRPC	NORM	17	86.27	90.24	0.719
MRPC	NORM	42	85.05	89.35	0.829
MRPC	NORM	123	86.03	90.02	0.718

Table 17: Paired held-aside differences (percentage points), left minus right; all contrasts retained. Nominal 95% intervals are descriptive, not multiplicity-adjusted findings.

Task	Contrast	Metric	Seed 17	Seed 42	Seed 123	Mean±SD	Nominal 95% CI
RTE	MIX–NORM	accuracy	0.00	−1.44	1.81	0.12 ± 1.63	$[-3.92, 4.16]$
RTE	MIX–UNREG	accuracy	−0.36	−0.72	0.00	-0.36 ± 0.36	$[-1.26, 0.54]$
RTE	NORM–UNREG	accuracy	−0.36	0.72	−1.81	-0.48 ± 1.27	$[-3.63, 2.67]$
MRPC	MIX–NORM	accuracy	−1.96	0.74	0.25	-0.33 ± 1.44	$[-3.89, 3.24]$
MRPC	MIX–NORM	f1	−1.28	0.61	0.09	-0.19 ± 0.98	$[-2.62, 2.23]$
MRPC	MIX–UNREG	accuracy	−1.47	0.25	−0.98	-0.74 ± 0.88	$[-2.93, 1.46]$
MRPC	MIX–UNREG	f1	−0.79	0.23	−0.77	-0.44 ± 0.58	$[-1.89, 1.00]$
MRPC	NORM–UNREG	accuracy	0.49	−0.49	−1.23	-0.41 ± 0.86	$[-2.55, 1.73]$
MRPC	NORM–UNREG	f1	0.49	−0.38	−0.86	-0.25 ± 0.69	$[-1.95, 1.45]$

Table 18: Complete RTE band comparison, mean±sample SD, three seeds per row. Accuracy and loss use the inner-selection split. Off-diagonal energy is a percentage within the selected band, computed directly from the compact trained cores; diagonal residuals are float32 roundoff. The head-only backbone update is zero, so its energy fraction is undefined. All nonzero updates have structural off-band zero.

Condition	Accuracy (%)	Task loss (nat)	Within-band off-diag. (%)	ρ_F
Head only	59.77 ± 0.65	0.654 ± 0.002	—	0
Leading / diagonal	71.75 ± 1.71	1.785 ± 0.083	0.00 ± 0.00	0.3283 ± 0.0009
Leading / rotation	72.22 ± 1.01	2.075 ± 0.118	11.36 ± 0.41	0.3090 ± 0.0006
Middle / diagonal	71.02 ± 1.21	1.759 ± 0.027	0.00 ± 0.00	0.3287 ± 0.0025
Middle / rotation	72.22 ± 1.21	1.958 ± 0.044	10.16 ± 0.29	0.3201 ± 0.0030
Tail / diagonal	70.01 ± 0.23	2.001 ± 0.141	0.00 ± 0.00	0.2560 ± 0.0017
Tail / rotation	71.02 ± 3.30	2.111 ± 0.268	7.88 ± 0.31	0.2720 ± 0.0085

the same frozen endpoints were evaluated on all 277 official validation examples; Table 1 uses this held-out follow-up. The inner-selection outcomes below remain unchanged.

The separate RTE study crosses three 256-direction bands with diagonal coefficients or rotations in the last 64 directions, plus an original-backbone head-only reference. Leading, middle and tail bands are $[0, 256)$, $[256, 512)$ and $[512, 768)$. All 48 attention updates use $\Delta = U_B H V_B^\top$, start at zero, and omit the ambient scalers and leading-identity contribution. Three seeds (17, 42, 123) give 21 registered confirmations. Two seed-31415 timing pilots are recorded separately. Off-band energy is structurally zero, with a 10^{-8} float32 layer-test bound.

Each rotation uses PyTorch 2.5.1's `nn.utils.parametrizations.orthogonal` with `orthogonal_map="matrix_exp"` and the default `use_trivialization=True`. Its square weight is initialized to the identity before registration. The matrix exponential maps a skew-symmetric parameter matrix to an orthogonal output throughout optimization. Each pair stores $2q^2$ trainable entries, while its two orthogonal outputs have $q(q-1)$ combined degrees of freedom (Appendix D).

The final export contains all 21 confirmation outcomes and the two timing pilots. Figure 4 plots all seeds; Table 18 summarizes every arm; Table 19 reports all 18 registered paired contrasts; Table 20 gives every individual endpoint. All scores are inner-selection results. The freeze ledger joins all 21 fixed checkpoints to their validation hashes and the same registered protocol. Although repository revisions differ across launches, their experiment-source trees are identical; the supplement verifies all 46 recorded source-file fingerprints against a saved source snapshot. Within-band off-diagonal energy is computed directly from compact cores, with signed floating-point residuals retained in the raw export for diagonal arms.

The tail-capacity contrast compares tail-only and head-only training at the same fixed endpoint, with common initial heads and data order. The flexibility contrast compares rotated and coefficient adapters at the same band width and training budget. Leading and middle bands provide location controls. Row order and these contrasts were fixed before confirmation inspection, after disclosure of the tail pilot. The later emphasis on the tail question leaves the full registered population intact.

Rotated updates place 11.36%, 10.16% and 7.88% of their energy off the diagonal within the leading, middle and tail bands, respectively. The rotations therefore change the trained core in every band. Mean accuracy changes relative to coefficient adapters are $+0.47$, $+1.20$ and $+1.00$ points; all paired intervals include zero. Mean prediction loss increases by 0.290, 0.199 and 0.110 nats. The diagonal tail's accuracy gain over head-only training occurs in all three seeds, together with a mean loss increase of 1.347 nats. These results separate the ability to learn useful classifications from improvements in the probabilities assigned to them.

The comparisons here use the fixed endpoints; we make no convergence-rate or optimizer-stability claim. A zero-initialized core gives zero rotation gradients at insertion; the final nonzero within-band off-diagonal energy shows that training leaves the diagonal family. Rotations add trainable parameters, so the comparison holds spectral support and training budget fixed while varying flexibility.

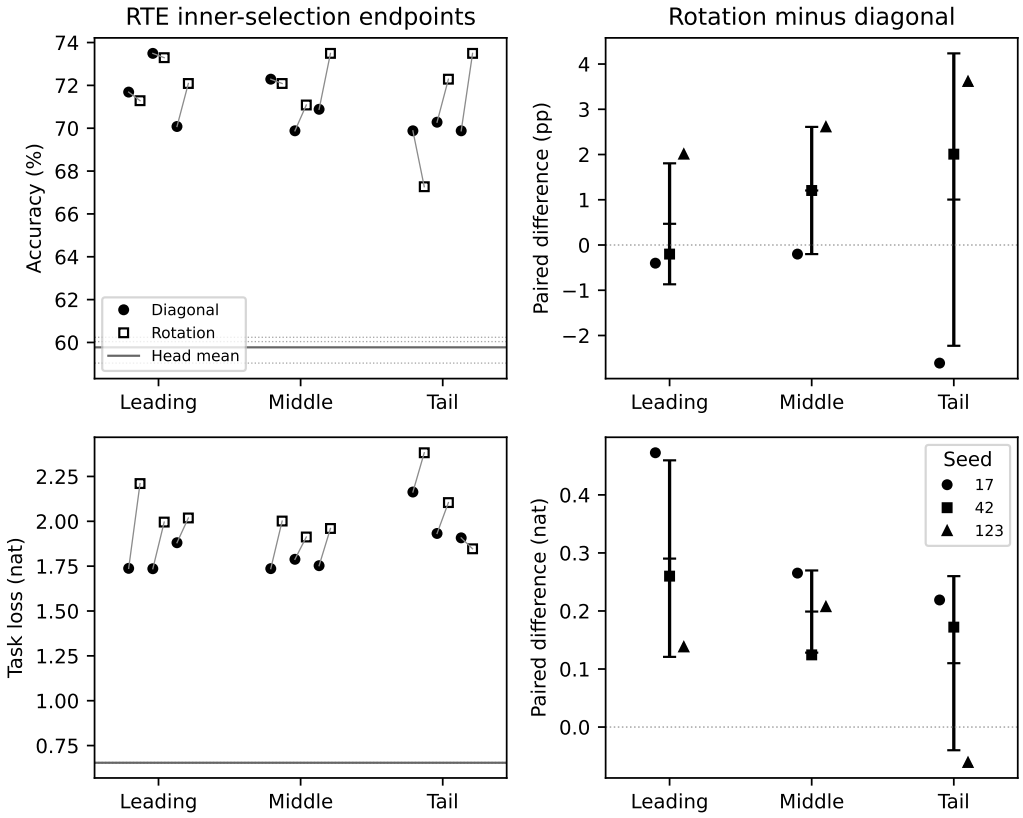


Figure 4: All three seeds for each band and flexibility level, at the fixed inner-selection endpoint. Left: diagonal and rotated outcomes are joined within seed; horizontal lines show head-only references. Right: every paired rotation-minus-diagonal difference, with mean and sample SD. Rotations create additional within-band interaction without a consistent accuracy improvement; the loss panels show the accompanying confidence-sensitive behavior.

Table 19: All registered paired RTE contrasts at the fixed endpoint on inner-selection examples. Entries are mean±sample SD of the three seed differences and nominal 95% paired t intervals (two degrees of freedom). Rotation-gain differences compare rotation-minus-diagonal changes between bands. Intervals are descriptive and unadjusted for multiple comparisons.

Contrast	Accuracy (pp)	Interval	Task loss (nat)	Interval
Leading / diagonal – head	11.98 ± 1.71	$[7.74, 16.22]$	1.130 ± 0.085	$[0.920, 1.341]$
Leading / rotation – head	12.45 ± 0.53	$[11.13, 13.77]$	1.421 ± 0.117	$[1.131, 1.711]$
Middle / diagonal – head	11.24 ± 1.84	$[6.67, 15.82]$	1.105 ± 0.027	$[1.038, 1.171]$
Middle / rotation – head	12.45 ± 1.41	$[8.96, 15.94]$	1.304 ± 0.044	$[1.194, 1.413]$
Tail / diagonal – head	10.24 ± 0.53	$[8.92, 11.56]$	1.347 ± 0.140	$[1.000, 1.693]$
Tail / rotation – head	11.24 ± 2.70	$[4.53, 17.96]$	1.457 ± 0.266	$[0.796, 2.118]$
LEAD rotation – diagonal	0.47 ± 1.34	$[-2.85, 3.79]$	0.290 ± 0.169	$[-0.130, 0.711]$
MID rotation – diagonal	1.20 ± 1.41	$[-2.29, 4.70]$	0.199 ± 0.071	$[0.023, 0.375]$
TAIL rotation – diagonal	1.00 ± 3.23	$[-7.02, 9.03]$	0.110 ± 0.150	$[-0.263, 0.483]$
MID – LEAD / diagonal	-0.74 ± 2.49	$[-6.93, 5.46]$	-0.026 ± 0.093	$[-0.256, 0.204]$
MID – LEAD / rotation	-0.00 ± 1.94	$[-4.81, 4.81]$	-0.117 ± 0.081	$[-0.317, 0.083]$
Rotation gain: MID – LEAD	0.74 ± 0.61	$[-0.79, 2.26]$	-0.091 ± 0.144	$[-0.448, 0.265]$
TAIL – LEAD / diagonal	-1.74 ± 1.51	$[-5.48, 2.00]$	0.216 ± 0.200	$[-0.280, 0.713]$
TAIL – LEAD / rotation	-1.20 ± 2.72	$[-7.95, 5.54]$	0.036 ± 0.183	$[-0.418, 0.491]$
Rotation gain: TAIL – LEAD	0.54 ± 2.40	$[-5.42, 6.49]$	-0.180 ± 0.084	$[-0.390, 0.030]$
TAIL – MID / diagonal	-1.00 ± 1.41	$[-4.50, 2.49]$	0.242 ± 0.160	$[-0.156, 0.640]$
TAIL – MID / rotation	-1.20 ± 3.19	$[-9.12, 6.71]$	0.153 ± 0.249	$[-0.466, 0.772]$
Rotation gain: TAIL – MID	-0.20 ± 1.92	$[-4.96, 4.56]$	-0.089 ± 0.162	$[-0.492, 0.314]$

B.8 COMMONSENSEQA DECODER COHORT

The Qwen2.5-1.5B-Instruct cohort adapts all 56 query/output attention projections. Each projection has width 1,536; the leading, middle and tail bands each contain 512 directions, and the rotated form acts within a 128-direction block. All six conditions use 842 AdamW updates, learning rate 10^{-3}, effective batch size 16 and the same three seeds. Evaluation uses all 1,221 public-validation questions, with unmerged bf16 choice accuracy and native choice NLL. The current GSM8K recipe and results are reported in Appendix B.10.

Table 20: All 21 strict-band/head confirmations and two timing pilots. Scores use fixed endpoints on the inner-selection split; each row is one run. Pilots are excluded from every confirmation summary. Adapter counts exclude the common 592,130-parameter trained head. Band updates remain inside their selected pretrained span by construction.

Task	Band	Core	Seed	Acc. (%)	Task loss	Adapter params
All 21 confirmation endpoints						
RTE	Lead	Diagonal	17	71.69	1.738	12,288
RTE	Lead	Diagonal	42	73.49	1.736	12,288
RTE	Lead	Diagonal	123	70.08	1.881	12,288
RTE	Lead	Rotated 64	17	71.29	2.210	405,504
RTE	Lead	Rotated 64	42	73.29	1.996	405,504
RTE	Lead	Rotated 64	123	72.09	2.019	405,504
RTE	Mid	Diagonal	17	72.29	1.736	12,288
RTE	Mid	Diagonal	42	69.88	1.788	12,288
RTE	Mid	Diagonal	123	70.88	1.753	12,288
RTE	Mid	Rotated 64	17	72.09	2.001	405,504
RTE	Mid	Rotated 64	42	71.08	1.913	405,504
RTE	Mid	Rotated 64	123	73.49	1.960	405,504
RTE	Tail	Diagonal	17	69.88	2.163	12,288
RTE	Tail	Diagonal	42	70.28	1.932	12,288
RTE	Tail	Diagonal	123	69.88	1.908	12,288
RTE	Tail	Rotated 64	17	67.27	2.382	405,504
RTE	Tail	Rotated 64	42	72.29	2.104	405,504
RTE	Tail	Rotated 64	123	73.49	1.847	405,504
RTE	Head only	—	17	59.04	0.656	0
RTE	Head only	—	42	60.24	0.655	0
RTE	Head only	—	123	60.04	0.652	0
Timing pilots (excluded from confirmations)						
RTE	Tail	Diagonal	31415	70.28	1.884	12,288
RTE	Tail	Rotated 64	31415	70.48	2.104	405,504

Table 21: Qwen2.5-1.5B-Instruct CommonsenseQA, seeds 17, 42 and 123: mean±sample SD at fixed endpoints. Frozen is one reference. Accuracy is percent; native NLL is per answer choice.

Condition	Accuracy	Native NLL
Frozen	70.60	1.9129
Leading coefficients	75.51 ± 0.38	0.6651 ± 0.0007
Leading + rotations	76.71 ± 0.29	0.6258 ± 0.0014
Middle coefficients	75.70 ± 0.17	0.6623 ± 0.0010
Middle + rotations	76.69 ± 0.37	0.6232 ± 0.0016
Tail coefficients	76.03 ± 0.19	0.6493 ± 0.0004
Tail + rotations	77.15 ± 0.08	0.6128 ± 0.0009

B.9 Random-frame controls: RTE and CommonsenseQA

The random-frame control repeats the CommonsenseQA band study with an equal-width random frame whose input and output bases are independent random orthonormal matrices rather than pretrained singular vectors. Frames are drawn once per training seed and adapted matrix and shared by the coefficient and rotated forms; the coefficient form learns 512 diagonal entries and the rotated form additionally rotates the last 128 directions, exactly as in the spectral bands. The recipe, adapted modules, seeds (17, 42, 123), inner-selection and held-out splits, unmerged bf16 evaluation, choice-probability renormalization and selection-fitted temperature are those of Appendix B.8; the six random runs were read through the same sealed collector as the 18 original band runs and the frozen reference. Paired contrasts use the same seed in both arms; the frozen reference is scored once and paired with every seed. These controls were added after the original band study. Under the

registered comparison rule, a random-minus-band accuracy difference is *equivalent* when its nominal 90% interval lies inside ± 2 points, *noninferior* when its one-sided 95% lower bound exceeds -2 points, and *resolved* when its nominal 95% interval excludes zero; every satisfied reading is reported. Geometry is measured on the saved endpoints in the pretrained singular frame, pooling squared energies over the 56 adapted matrices. For this two-thirds/one-third partition, a dimensionally uniform update has cross share $4/9$; that value is a descriptive reference, not a prediction for a learned update. Every random update lies inside its random frame to within 10^{-9} of its energy. Acceptance of the export, its per-example scoring and reload bindings is recorded separately from this analysis.

On RTE, the same construction uses 256-direction random frames with rotations over the last 64 directions under the band study's recipe (Appendix B.7), with reported values from a CPU re-evaluation of the frozen endpoints on the 277 held-out validation examples of the locked split, which reproduces the training-time GPU accuracies exactly and the losses within the recorded tolerance, and a temperature fitted on inner-selection logits; the head-only reference is trained per seed and paired by seed. Three RTE attempts were interrupted by a scheduling error and replaced; the replacements reproduce identical compared training state over the interrupted prefixes, and the six controller reports were audited before acceptance. The registered readings apply to both tasks; on RTE no equivalence is established, and the tables report every coexisting finding. Tables 22–24 give the RTE per-seed outcomes and contrasts; Tables 25–27 give the CommonsenseQA ones.

Table 22: RTE random-frame control, per seed. Held-out accuracy, task NLL before and after selection-fitted temperature, the fitted temperature, pooled cross and leading–leading shares in the pretrained frame, and pooled relative update norm.

Form	Seed	Run	Acc. (%)	NLL	Corr. NLL	T	Cross (%)	LL (%)	ρ_F
Coefficients	17	`81ffa8a13d50`	71.84	1.779	0.551	7.53	44.45	44.45	0.2720
	42	`111043ad8542`	73.65	1.545	0.532	6.67	44.47	44.42	0.2689
	123	`d64212f9f35f`	70.40	1.729	0.563	7.83	44.43	44.49	0.2692
+ rotations	17	`44c0fb5bdb87`	73.65	1.765	0.524	7.70	44.44	44.44	0.2646
	42	`6dfc059c48fc`	76.53	1.644	0.501	7.58	44.45	44.43	0.2625
	123	`1aa250ed2b4d`	72.20	1.790	0.540	8.12	44.48	44.39	0.2568

Table 23: RTE random-frame accuracy contrasts over three paired seeds, random minus comparator, in percentage points: nominal 95% and 90% intervals and the one-sided 95% lower bound. Readings: Improves, 95% lower bound above zero against the reference; E, equivalent within two points (90% interval); NI, noninferior (one-sided bound above -2). Unresolved differences are not equivalence.

Form	Comparator	Acc. Δ	95%	90%	Lower	Reading
Coefficients	Leading	-0.72	$[-7.84, +6.40]$	$[-5.553, +4.109]$	-5.553	Unresolved
	Middle	$+2.05$	$[-1.10, +5.20]$	$[-0.092, +4.183]$	-0.092	NI
	Tail	$+2.77$	$[-4.75, +10.29]$	$[-2.336, +7.872]$	-2.336	Unresolved
	Head-only reference	$+13.84$	$[+10.21, +17.46]$	$[+11.379, +16.298]$	$+11.379$	Improves
+ rotations	Leading	-0.48	$[-6.97, +6.01]$	$[-4.884, +3.921]$	-4.884	Unresolved
	Middle	$+2.41$	$[-4.08, +8.89]$	$[-1.996, +6.810]$	-1.996	NI
	Tail	$+5.78$	$[-2.87, +14.42]$	$[-0.093, +11.645]$	-0.093	NI
	Head-only reference	$+16.00$	$[+10.83, +21.18]$	$[+12.491, +19.519]$	$+12.491$	Improves

B.10 GSM8K RESULTS

This post hoc cohort compares all eight strict conditions under a 421-step, 10^{-4} recipe with 21 warmup steps, linear decay and effective batch size 16. The adapted modules, band widths and rotation block are fixed across arms; three seeds (17, 42, 123) pair their training runs. Each random frame is drawn per seed and matrix and shared by its coefficient and rotated forms. Greedy unmerged-bf16 generation uses a 640-token cap and the same fixed 768-example subset for every arm and the frozen reference; completion NLL uses all 1,319 test examples. The orthogonal map is cached within each no-grad evaluation call after a token-and-NLL parity check, and no temperature is fitted. The recipe was selected after earlier GSM8K tests and this test split had already been evaluated; these results are not untouched-test confirmation. No equivalence margin was registered for this cohort. Random-frame updates place about 44% of their energy in the leading–tail cross blocks of

Table 24: RTE random-frame loss contrasts over three paired seeds, random minus comparator, in nats per answer: native and temperature-corrected NLL with nominal 95% intervals.

Form	Comparator	Native Δ	95%	Corrected Δ	95%
Coefficients	Leading	-0.0985	$[-0.50399, +0.30702]$	-0.006	$[-0.064, +0.051]$
	Middle	-0.0418	$[-0.40172, +0.31808]$	-0.015	$[-0.055, +0.025]$
	Tail	-0.3232	$[-0.88509, +0.23866]$	-0.021	$[-0.074, +0.032]$
	Head-only reference	$+1.0046$	$[+0.70020, +1.30895]$	-0.133	$[-0.167, -0.100]$
+ rotations	Leading	-0.1991	$[-0.91345, +0.51520]$	-0.019	$[-0.103, +0.066]$
	Middle	-0.3138	$[-0.62744, -0.00009]$	-0.028	$[-0.071, +0.014]$
	Tail	-0.5174	$[-1.23816, +0.20340]$	-0.062	$[-0.171, +0.047]$
	Head-only reference	$+1.0533$	$[+0.86215, +1.24440]$	-0.160	$[-0.204, -0.117]$

Table 25: CommonsenseQA random-frame control, per seed. Held-out choice accuracy, choice NLL before and after selection-fitted temperature, the fitted temperature, pooled cross and leading–leading shares in the pretrained frame, and pooled relative update norm.

Form	Seed	Run	Acc. (%)	NLL	Corr. NLL	T	Cross (%)	LL (%)	ρ_F
Coefficients	17	`c96f9811c85d`	75.68	0.6554	0.6553	1.006	44.45	44.43	0.0204
	42	`4dc6ec24f4be`	75.51	0.6558	0.6558	0.999	44.44	44.45	0.0203
	123	`ff5282b93ac9`	75.84	0.6574	0.6574	1.000	44.43	44.48	0.0203
+ rotations	17	`1b1a7c4e9e7e`	77.72	0.6147	0.6141	1.022	44.45	44.43	0.0174
	42	`2372adb7282f`	77.15	0.6199	0.6198	1.004	44.45	44.44	0.0174
	123	`d80004572a1c`	77.31	0.6181	0.6179	1.009	44.44	44.46	0.0177

Table 26: CommonsenseQA random-frame contrasts over three paired seeds. Accuracy differences are percentage points, random minus comparator, except the final row (rotations minus coefficients). Intervals are nominal 95% and 90%; Lower is the one-sided 95% lower bound. E: equivalent within two points; NI: noninferior within two points; Above: the nominal 95% interval excludes zero positively. These findings can coexist. Intervals are unadjusted.

Form	Comparator	Acc. Δ	95%	90%	Lower	Reading
Coefficients	Leading	$+0.16$	$[-0.97, +1.30]$	$[-0.60, +0.93]$	-0.60	E; NI
	Middle	-0.03	$[-0.80, +0.74]$	$[-0.55, +0.50]$	-0.55	E; NI
	Tail	-0.35	$[-0.98, +0.27]$	$[-0.78, +0.07]$	-0.78	E; NI
	Frozen reference	$+5.08$	$[+4.67, +5.48]$	$[+4.80, +5.35]$	$+4.80$	Improves
+ rotations	Leading	$+0.68$	$[+0.06, +1.30]$	$[+0.26, +1.10]$	$+0.26$	Above; E; NI
	Middle	$+0.71$	$[-0.33, +1.75]$	$[+0.00, +1.42]$	$+0.00$	E; NI
	Tail	$+0.25$	$[-0.36, +0.86]$	$[-0.17, +0.66]$	-0.17	E; NI
	Frozen reference	$+6.80$	$[+6.06, +7.53]$	$[+6.30, +7.30]$	$+6.30$	Improves
Rot. $-$ coeff.	Random frame	$+1.72$	$[+0.99, +2.45]$	$[+1.22, +2.22]$	$+1.22$	Above

Table 27: CommonsenseQA random-frame choice-loss contrasts over three paired seeds. Differences are nats per answer, random minus comparator, except the final row (rotations minus coefficients). Intervals are nominal 95% and unadjusted. Negative differences favor the first arm.

Form	Comparator	NLL Δ	95% interval
Coefficients	Leading	-0.0090	$[-0.0116, -0.0063]$
	Middle	-0.0061	$[-0.0102, -0.0020]$
	Tail	$+0.0069$	$[+0.0054, +0.0084]$
	Frozen reference	-1.2568	$[-1.2593, -1.2542]$
+ rotations	Leading	-0.0082	$[-0.0121, -0.0043]$
	Middle	-0.0056	$[-0.0087, -0.0025]$
	Tail	$+0.0048$	$[+0.0002, +0.0093]$
	Frozen reference	-1.2953	$[-1.3018, -1.2888]$
Rot. $-$ coeff.	Random frame	-0.0386	$[-0.0447, -0.0325]$

the pretrained frame; strictly confined bands have none by construction. Tables 28 and 29 report every arm and paired contrast. All eight arms lower completion NLL relative to the frozen reference, but none has a resolved accuracy gain; tail rotations lower accuracy by 2.26 percentage points, with nominal interval $[-2.75, -1.76]$.

Table 28: GSM8K results. Exact match on the fixed 768-example subset and native completion NLL on all 1,319 test examples, mean±SD over three seeds; the frozen reference is scored once on the same subset. Cross is the pooled leading–tail cross share in the pretrained frame (measured for random frames); ρ_F is the pooled relative update norm.

Support	Update	Exact match (%)	Native NLL	Cross (%)	ρ_F
Leading	Coefficients	59.11 ± 0.34	0.5439 ± 0.0002	–	0.0061
	+ rotations	58.51 ± 0.89	0.5285 ± 0.0001	–	0.0063
Middle	Coefficients	59.20 ± 0.62	0.5405 ± 0.0000	–	0.0062
	+ rotations	57.73 ± 1.46	0.5233 ± 0.0001	–	0.0063
Tail	Coefficients	58.94 ± 0.87	0.5297 ± 0.0001	–	0.0058
	+ rotations	56.86 ± 0.20	0.5093 ± 0.0000	–	0.0059
Random frame	Coefficients	59.98 ± 0.72	0.5365 ± 0.0001	44.4	0.0060
	+ rotations	56.68 ± 1.04	0.5198 ± 0.0001	44.4	0.0062
Frozen reference		59.11	0.5934	–	–

Table 29: GSM8K paired contrasts over three seeds. Exact-match differences are percentage points; NLL differences are nats per completion token; intervals are nominal paired 95% intervals, unadjusted. The frozen reference is paired with each seed. Unresolved differences are not equivalence.

Contrast	Comparator or family	EM Δ	95%	NLL Δ	95%
Leading, Coefficients	Frozen reference	-0.00	$[-0.86, +0.86]$	-0.0495	$[-0.0499, -0.0490]$
Leading, rotations	Frozen reference	-0.61	$[-2.81, +1.59]$	-0.0649	$[-0.0651, -0.0648]$
Middle, Coefficients	Frozen reference	$+0.09$	$[-1.44, +1.62]$	-0.0529	$[-0.0530, -0.0528]$
Middle, rotations	Frozen reference	-1.39	$[-5.02, +2.25]$	-0.0701	$[-0.0702, -0.0699]$
Tail, Coefficients	Frozen reference	-0.17	$[-2.33, +1.98]$	-0.0637	$[-0.0640, -0.0634]$
Tail, rotations	Frozen reference	-2.26	$[-2.75, -1.76]$	-0.0840	$[-0.0841, -0.0840]$
Random frame, Coefficients	Frozen reference	$+0.87$	$[-0.91, +2.65]$	-0.0568	$[-0.0572, -0.0565]$
Random frame, rotations	Frozen reference	-2.43	$[-5.02, +0.16]$	-0.0736	$[-0.0738, -0.0734]$
Leading − middle	Coefficients	-0.09	$[-1.75, +1.57]$	$+0.0035$	$[+0.0029, +0.0040]$
Leading − tail	Coefficients	$+0.17$	$[-2.83, +3.18]$	$+0.0142$	$[+0.0139, +0.0146]$
Middle − tail	Coefficients	$+0.26$	$[-2.39, +2.91]$	$+0.0108$	$[+0.0105, +0.0110]$
Leading − middle	rotations	$+0.78$	$[-2.60, +4.16]$	$+0.0051$	$[+0.0050, +0.0052]$
Leading − tail	rotations	$+1.65$	$[-0.99, +4.28]$	$+0.0191$	$[+0.0189, +0.0193]$
Middle − tail	rotations	$+0.87$	$[-2.74, +4.47]$	$+0.0140$	$[+0.0137, +0.0142]$
Random − leading	Coefficients	$+0.87$	$[-0.36, +2.09]$	-0.0074	$[-0.0077, -0.0071]$
Random − middle	Coefficients	$+0.78$	$[-0.51, +2.08]$	-0.0039	$[-0.0043, -0.0035]$
Random − tail	Coefficients	$+1.04$	$[-2.60, +4.69]$	$+0.0068$	$[+0.0064, +0.0073]$
Random − leading	rotations	-1.82	$[-5.47, +1.82]$	-0.0086	$[-0.0089, -0.0083]$
Random − middle	rotations	-1.04	$[-7.21, +5.13]$	-0.0035	$[-0.0039, -0.0031]$
Random − tail	rotations	-0.17	$[-3.00, +2.65]$	$+0.0105$	$[+0.0103, +0.0106]$
Rotations − coefficients	Leading	-0.61	$[-2.20, +0.99]$	-0.0155	$[-0.0159, -0.0151]$
Rotations − coefficients	Middle	-1.48	$[-4.48, +1.53]$	-0.0171	$[-0.0172, -0.0170]$
Rotations − coefficients	Tail	-2.08	$[-4.05, -0.12]$	-0.0204	$[-0.0207, -0.0200]$
Rotations − coefficients	Random frame	-3.30	$[-7.67, +1.07]$	-0.0167	$[-0.0169, -0.0166]$

B.11 Second decoder family: Llama-3.1-8B-Instruct on CommonsenseQA

To test whether the location findings depend on the Qwen family, we repeat the CommonsenseQA study on the official Llama-3.1-8B-Instruct checkpoint at a pinned revision, adapting all 64 query and output projections (width 4,096) with 512-direction bands. Registration fixed the leading band $[0, 512)$ and the tail band $[3584, 4096)$ and omitted the middle band before this cohort's held-out evaluation; an equal-width random frame is drawn per seed and adapted matrix by the rule of Appendix B.9. The recipe is the Qwen CommonsenseQA recipe with only model-specific fields changed: 842 updates at learning rate 10^{-3}, effective batch size 16, 160-token inputs, no warmup, an

inner split of ten percent of the training set (seed 271828), and one evaluation on the 1,221 public validation questions, which serve as the held-out split; the frozen checkpoint is scored once on the same questions. Choice accuracy and choice NLL are the two co-primary outcomes; no temperature is fitted and nothing is generated. The accepted original core comprises all 18 confirmations (three supports, two adapter families, seeds 17, 42 and 123) and the frozen reference. Coefficient and rotated forms have 32,768 and 2,129,920 trainable parameters, respectively; their comparison does not isolate rotation at fixed capacity. Native scoring uses the registered bf16 path, distinct from the float32 checkpoint interventions below. Nothing is pooled with Qwen, and no equivalence or noninferiority rule transfers to this cohort. All three leading coefficient runs answer 944 of 1,221 questions correctly, with different predictions, losses and checkpoints. Tables 30–32 retain all outcomes and nominal paired contrasts.

Tail outperforms leading support in accuracy and choice NLL in both families. Random support outperforms leading in accuracy but its contrast against tail is unresolved in the original three seeds; its choice NLL is higher than tail's. The post hoc five-seed view resolves the rotated-family accuracy contrast in favor of tail (Table 36); the coefficient-family contrast remains unresolved. Rotated forms improve both outcomes across all three supports. These are comparisons under the common recipe on this model.

Table 30: Llama-3.1-8B-Instruct on CommonsenseQA, complete original six-arm core. C denotes coefficients and R coefficients plus rotations. Held-out choice accuracy and choice NLL, mean±SD over three seeds; ρ_F is the pooled relative update norm; the frozen reference is scored once.

Support	Update	Accuracy (%)	Choice NLL	ρ_F
Leading	C	77.31 ± 0.00	0.6028 ± 0.0016	0.0146
Leading	R	79.01 ± 0.31	0.5567 ± 0.0008	0.0139
Tail	C	78.13 ± 0.22	0.5720 ± 0.0017	0.0112
Tail	R	80.48 ± 0.50	0.5303 ± 0.0020	0.0105
Random	C	78.05 ± 0.22	0.5938 ± 0.0023	0.0135
Random	R	79.80 ± 0.13	0.5456 ± 0.0015	0.0126
Frozen reference		76.66	0.9739	–

Table 31: Llama-3.1-8B-Instruct paired contrasts over three seeds. Accuracy differences are percentage points and NLL differences nats per answer, first minus second, with nominal 95% intervals; unadjusted, and unresolved differences are not equivalence. A zero-width accuracy interval reflects identical observed seed differences, not zero population uncertainty.

Contrast	Accuracy difference [95%]	NLL difference [95%]
Leading C − Tail C	-0.82 $[-1.36, -0.28]$	$+0.0309$ $[+0.0296, +0.0321]$
Random C − Leading C	$+0.74$ $[+0.20, +1.28]$	-0.0091 $[-0.0110, -0.0071]$
Random C − Tail C	-0.08 $[-1.01, +0.85]$	$+0.0218$ $[+0.0201, +0.0235]$
Leading R − Leading C	$+1.69$ $[+0.92, +2.46]$	-0.0462 $[-0.0506, -0.0417]$
Leading C − Frozen	$+0.66$ $[+0.66, +0.66]$	-0.3710 $[-0.3750, -0.3671]$
Leading R − Frozen	$+2.35$ $[+1.58, +3.12]$	-0.4172 $[-0.4193, -0.4151]$
Random R − Random C	$+1.75$ $[+1.24, +2.26]$	-0.0482 $[-0.0558, -0.0406]$
Random C − Frozen	$+1.39$ $[+0.85, +1.93]$	-0.3801 $[-0.3858, -0.3744]$
Random R − Frozen	$+3.14$ $[+2.83, +3.45]$	-0.4283 $[-0.4321, -0.4245]$
Leading R − Tail R	-1.47 $[-2.90, -0.05]$	$+0.0264$ $[+0.0229, +0.0298]$
Random R − Leading R	$+0.79$ $[+0.28, +1.30]$	-0.0111 $[-0.0132, -0.0090]$
Random R − Tail R	-0.68 $[-1.80, +0.44]$	$+0.0153$ $[+0.0138, +0.0167]$
Tail R − Tail C	$+2.35$ $[+0.82, +3.87]$	-0.0417 $[-0.0490, -0.0343]$
Tail C − Frozen	$+1.47$ $[+0.94, +2.01]$	-0.4019 $[-0.4061, -0.3977]$
Tail R − Frozen	$+3.82$ $[+2.58, +5.07]$	-0.4436 $[-0.4485, -0.4387]$

Table 32: Every original Llama8B endpoint on CommonsenseQA. Correct counts are out of 1,221; accuracy is percent and NLL is nats per answer. C denotes coefficients and R coefficients plus rotations.

Support	Form	Seed	Correct	Accuracy (%)	Choice NLL
Leading	C	17	944	77.31	0.60283
Leading	C	42	944	77.31	0.60444
Leading	C	123	944	77.31	0.60123
Leading	R	17	969	79.36	0.55572
Leading	R	42	963	78.87	0.55717
Leading	R	123	962	78.79	0.55715
Tail	C	17	951	77.89	0.57138
Tail	C	42	956	78.30	0.57388
Tail	C	123	955	78.21	0.57065
Tail	R	17	984	80.59	0.52846
Tail	R	42	976	79.93	0.53010
Tail	R	123	988	80.92	0.53239
Random	C	17	955	78.21	0.59330
Random	C	42	954	78.13	0.59628
Random	C	123	950	77.81	0.59173
Random	R	17	976	79.93	0.54395
Random	R	42	973	79.69	0.54577
Random	R	123	974	79.77	0.54700
Frozen reference			936	76.66	0.97388

B.12 Post hoc Llama middle-band supplement

The middle band supports adaptation but does not match the tail under this recipe. We added middle-band C and R (run labels MID_DIAG and MID_ROT128) at seeds 17, 42 and 123, with the same 842-step recipe, modules, band width and native bf16 scoring as the original Llama core. The middle support is $[1792, 2304)$. Its registration followed inspection of original outcomes, so these six runs are a separate post hoc supplement. Table 33 labels the combined 24 trained endpoints; the original 18-run view and its frozen reference remain unchanged. The frozen bf16 reference is reused, not re-evaluated or replaced by the intervention's float32 reference.

Both middle forms improve on the frozen reference. Tail and random support exceed middle accuracy in both families, while leading–middle differences are unresolved. Rotation improves both middle-band outcomes. Table 34 gives the nine added pairs (18 metric contrasts); together with Table 31, it displays all 48 combined metric contrasts. These are nominal, unadjusted three-seed intervals, not equivalence tests.

Table 33: Llama original core plus post hoc MID supplement: a labelled combined 24-run view. Mean±sample SD over three seeds; C: coefficients, R: coefficients plus rotations. Original rows retain their separate original-cohort interpretation. ρ_F is pooled relative update norm.

Support/form	Cohort	Accuracy (%)	Choice NLL	ρ_F
Leading C	Original core	77.31 ± 0.00	0.6028 ± 0.0016	0.0146
Leading R	Original core	79.01 ± 0.31	0.5567 ± 0.0008	0.0139
Middle C	MID supplement	77.42 ± 0.05	0.6026 ± 0.0020	0.0146
Middle R	MID supplement	79.14 ± 0.05	0.5547 ± 0.0019	0.0134
Tail C	Original core	78.13 ± 0.22	0.5720 ± 0.0017	0.0112
Tail R	Original core	80.48 ± 0.50	0.5303 ± 0.0020	0.0105
Random C	Original core	78.05 ± 0.22	0.5938 ± 0.0023	0.0135
Random R	Original core	79.80 ± 0.13	0.5456 ± 0.0015	0.0126
Frozen	Reused original	76.66	0.9739	–

Table 34: All new middle-band pairs in the labelled combined Llama view. First minus second, nominal paired 95% intervals; accuracy differences in percentage points and choice NLL in nats per answer. The zero-width Middle R–C accuracy interval reflects identical observed seed differences, not zero population uncertainty.

Contrast	Accuracy difference [95%]	Choice NLL difference [95%]
Leading C − Middle C	-0.11 $[-0.23, +0.01]$	$+0.0002$ $[-0.0024, +0.0029]$
Middle C − Tail C	-0.71 $[-1.36, -0.06]$	$+0.0306$ $[+0.0293, +0.0320]$
Random C − Middle C	$+0.63$ $[+0.16, +1.10]$	-0.0088 $[-0.0111, -0.0066]$
Middle R − Middle C	$+1.72$ $[+1.72, +1.72]$	-0.0479 $[-0.0536, -0.0422]$
Middle C − Frozen	$+0.76$ $[+0.65, +0.88]$	-0.3713 $[-0.3762, -0.3664]$
Middle R − Frozen	$+2.48$ $[+2.37, +2.60]$	-0.4192 $[-0.4240, -0.4144]$
Leading R − Middle R	-0.14 $[-0.79, +0.52]$	$+0.0020$ $[-0.0008, +0.0049]$
Middle R − Tail R	-1.34 $[-2.56, -0.11]$	$+0.0243$ $[+0.0226, +0.0261]$
Random R − Middle R	$+0.66$ $[+0.45, +0.86]$	-0.0091 $[-0.0103, -0.0079]$

Table 35: All six post hoc Llama MID endpoints. Correct counts are out of 1,221; the original frozen and other 18 endpoints are retained in Table 32.

Support/form	Seed	Correct	Accuracy (%)	Choice NLL
Middle C	17	946	77.48	0.60136
Middle C	42	945	77.40	0.60484
Middle C	123	945	77.40	0.60154
Middle R	17	967	79.20	0.55250
Middle R	42	966	79.12	0.55524
Middle R	123	966	79.12	0.55623

B.13 LLAMA8B: TWO ADDITIONAL SEEDS AND A SEPARATE FIVE-SEED VIEW

The five-seed views in Table 36 retain similar accuracy-effect estimates while narrowing both nominal intervals below zero. The original three-seed comparisons remain primary; these cumulative follow-ups are post hoc.

The five-seed location view retains the lowest observed choice NLL for tail support in both families. This post hoc supplement adds seeds 2026 and 2027 across all eight arms, using the unchanged 842-update recipe. The 16 new endpoints join the original 18-run core and separately registered six-run middle-band supplement. Their original three-seed records remain unchanged; the combined view contains 40 trained endpoints and one reused frozen reference. Tables 37–39 separate the new-two and combined-five summaries and retain all 48 paired metric contrasts.

All native outcomes use the same bf16-autocast path and 1,221 public validation questions. The new seeds were registered after original outcomes were available. Combined intervals use the five paired training seeds, with four degrees of freedom; they are exploratory, nominal and unadjusted for multiplicity. Post hoc status follows the full view and its seed population, including comparisons against the reused frozen model. No equivalence classification transfers to this supplement.

Table 36: Llama CommonsenseQA accuracy differences (percentage points) before and after adding two seeds. Brackets give nominal 95% intervals across paired training seeds. The cumulative five-seed views are post hoc; original three-seed comparisons remain primary. Location uses native scores; checkpoint intervention uses its separate float32 path (Appendices B.13 and B.18).

Accuracy contrast	Original three	Post hoc five
Location: Random R − Tail R	-0.68 $[-1.80, +0.44]$	-0.64 $[-1.05, -0.23]$
Random R intervention: zero cross − shrink	-0.60 $[-1.54, +0.34]$	-0.54 $[-0.93, -0.15]$

Table 37: Llama8B native post hoc supplement. Mean±sample SD for the two new seeds and combined five seeds. C: coefficients; R: coefficients plus rotations. The original core and middle-band three-seed records remain separate.

Arm	New-two acc. (%)	New-two NLL	Five-seed acc. (%)	Five-seed NLL
Leading C	77.76 ± 0.06	0.60411 ± 0.00092	77.49 ± 0.25	0.60335 ± 0.00141
Leading R	78.99 ± 0.06	0.55678 ± 0.00254	79.00 ± 0.22	0.55672 ± 0.00140
Middle C	77.48 ± 0.35	0.60520 ± 0.00286	77.44 ± 0.18	0.60363 ± 0.00245
Middle R	79.12 ± 0.12	0.55460 ± 0.00357	79.13 ± 0.07	0.55463 ± 0.00225
Tail C	78.46 ± 0.12	0.57353 ± 0.00133	78.26 ± 0.24	0.57260 ± 0.00161
Tail R	80.26 ± 0.35	0.53035 ± 0.00002	80.39 ± 0.41	0.53033 ± 0.00140
Random C	77.68 ± 0.06	0.59797 ± 0.00389	77.90 ± 0.26	0.59545 ± 0.00342
Random R	79.69 ± 0.23	0.54626 ± 0.00357	79.75 ± 0.16	0.54585 ± 0.00212

Table 38: All 16 additional native Llama8B endpoints. Correct counts have denominator 1,221.

Arm	Seed	Correct	Accuracy (%)	Choice NLL
Leading C	2026	950	77.81	0.60346
Leading C	2027	949	77.72	0.60477
Leading R	2026	965	79.03	0.55499
Leading R	2027	964	78.95	0.55858
Middle C	2026	949	77.72	0.60317
Middle C	2027	943	77.23	0.60722
Middle R	2026	967	79.20	0.55207
Middle R	2027	965	79.03	0.55712
Tail C	2026	957	78.38	0.57259
Tail C	2027	959	78.54	0.57447
Tail R	2026	983	80.51	0.53033
Tail R	2027	977	80.02	0.53036
Random C	2026	949	77.72	0.59522
Random C	2027	948	77.64	0.60071
Random R	2026	975	79.85	0.54373
Random R	2027	971	79.52	0.54878

Table 39: All 48 paired metric contrasts in the separate five-seed Llama8B location view. Accuracy differences are percentage points; brackets are nominal 95% paired intervals ($df = 4$). Every row belongs to the post hoc combined view, including frozen-reference contrasts.

Contrast	Δ accuracy (pp)	Δ choice NLL
Leading C − Middle C	$+0.05$ $[-0.28, +0.38]$	-0.00028 $[-0.00205, +0.00149]$
Leading C − Tail C	-0.77 $[-1.00, -0.54]$	$+0.03075$ $[+0.03020, +0.03130]$
Middle C − Tail C	-0.82 $[-1.23, -0.41]$	$+0.03103$ $[+0.02975, +0.03231]$
Random C − Leading C	$+0.41$ $[-0.18, +1.00]$	-0.00790 $[-0.01069, -0.00511]$
Random C − Middle C	$+0.46$ $[+0.08, +0.84]$	-0.00818 $[-0.00966, -0.00670]$
Random C − Tail C	-0.36 $[-0.95, +0.23]$	$+0.02285$ $[+0.02039, +0.02532]$
Leading R − Leading C	$+1.51$ $[+1.09, +1.92]$	-0.04662 $[-0.04866, -0.04459]$
Leading C − Frozen	$+0.84$ $[+0.53, +1.14]$	-0.37053 $[-0.37229, -0.36878]$
Leading R − Frozen	$+2.34$ $[+2.07, +2.62]$	-0.41716 $[-0.41890, -0.41542]$
Middle R − Middle C	$+1.69$ $[+1.53, +1.84]$	-0.04900 $[-0.05174, -0.04625]$
Middle C − Frozen	$+0.79$ $[+0.56, +1.01]$	-0.37025 $[-0.37330, -0.36721]$
Middle R − Frozen	$+2.47$ $[+2.39, +2.56]$	-0.41925 $[-0.42204, -0.41645]$
Random R − Random C	$+1.85$ $[+1.58, +2.13]$	-0.04960 $[-0.05320, -0.04601]$
Random C − Frozen	$+1.24$ $[+0.93, +1.56]$	-0.37843 $[-0.38268, -0.37418]$
Random R − Frozen	$+3.10$ $[+2.90, +3.29]$	-0.42803 $[-0.43067, -0.42540]$
Leading R − Middle R	-0.13 $[-0.37, +0.10]$	$+0.00209$ $[+0.00089, +0.00329]$
Leading R − Tail R	-1.39 $[-1.94, -0.84]$	$+0.02639$ $[+0.02441, +0.02838]$
Middle R − Tail R	-1.26 $[-1.74, -0.79]$	$+0.02430$ $[+0.02201, +0.02659]$
Random R − Leading R	$+0.75$ $[+0.53, +0.97]$	-0.01087 $[-0.01193, -0.00981]$
Random R − Middle R	$+0.62$ $[+0.51, +0.74]$	-0.00879 $[-0.00944, -0.00813]$
Random R − Tail R	-0.64 $[-1.05, -0.23]$	$+0.01552$ $[+0.01322, +0.01782]$
Tail R − Tail C	$+2.13$ $[+1.41, +2.84]$	-0.04227 $[-0.04519, -0.03935]$
Tail C − Frozen	$+1.61$ $[+1.30, +1.91]$	-0.40128 $[-0.40329, -0.39928]$
Tail R − Frozen	$+3.73$ $[+3.22, +4.25]$	-0.44355 $[-0.44528, -0.44182]$

B.14 QWEN2.5-3B-INSTRUCT ON COMMONSENSEQA

The complete original cohort compares leading, middle, tail and random support in both adapter families, with seeds 17, 42 and 123 plus a frozen reference: 24 trained runs and one reference. We adapt all 72 query/output matrices of width 2,048, with bands of width 512 starting at 0, 768 and 1,536. The common recipe uses 842 updates, learning rate 10^{-3}, effective batch 16, 160-token inputs and no warmup. Choice accuracy and choice NLL are co-primary; all runs score the same 1,221 public validation questions, separately from the 974 selection questions. Coefficients and rotated forms have 36,864 and 2,396,160 trainable parameters. This is a separate model check, not a model-only intervention or a pooled scale comparison.

The pilot failed its prespecified accuracy-movement criterion. The cohort was admitted under a post hoc validity rule fixed after selection-pilot evaluation and before confirmation training. Only selection examples were scored in the pilot, although preprocessing included all splits. The original three-seed cohort is reported in full, without changes to its population or recipe based on held-out outcomes. Separately accepted post hoc seeds appear in Appendix B.15, and checkpoint interventions in Appendix B.20.

Tail support has higher accuracy and lower choice NLL than leading, middle and random support in both families. Rotated forms improve both outcomes across all four supports, with more trainable parameters. Tables 40–42 report every seed and all 48 paired metric contrasts. Intervals are nominal and unadjusted; no registered equivalence classification transfers to this cohort.

Table 40: Qwen2.5-3B-Instruct, complete original CommonsenseQA cohort. Mean±sample SD over three seeds. C: coefficients; R: coefficients plus rotations. ρ_F is pooled relative update norm.

Support	Form	Accuracy (%)	Choice NLL	ρ_F
Leading	C	78.27 ± 0.21	0.6170 ± 0.0028	0.0264
Leading	R	80.75 ± 0.22	0.5402 ± 0.0021	0.0193
Middle	C	78.38 ± 0.30	0.6195 ± 0.0031	0.0260
Middle	R	80.89 ± 0.24	0.5419 ± 0.0015	0.0181
Tail	C	79.77 ± 0.08	0.5719 ± 0.0001	0.0188
Tail	R	82.23 ± 0.16	0.5135 ± 0.0022	0.0146
Random	C	78.65 ± 0.19	0.5969 ± 0.0035	0.0229
Random	R	81.22 ± 0.13	0.5325 ± 0.0028	0.0172
Frozen reference		76.99	2.9985	–

Table 41: Every original Qwen3B paired contrast: first minus second, nominal 95% Student-t interval over three seeds. Accuracy differences are percentage points and NLL differences nats per answer. C: coefficients; R: coefficients plus rotations.

Contrast	Accuracy difference [95%]	NLL difference [95%]
Leading C − Middle C	-0.11 $[-0.58, +0.36]$	-0.0025 $[-0.0052, +0.0002]$
Leading C − Tail C	-1.50 $[-2.22, -0.79]$	$+0.0452$ $[+0.0385, +0.0518]$
Middle C − Tail C	-1.39 $[-2.28, -0.51]$	$+0.0476$ $[+0.0402, +0.0551]$
Random C − Leading C	$+0.38$ $[-0.27, +1.04]$	-0.0201 $[-0.0227, -0.0175]$
Random C − Middle C	$+0.27$ $[-0.24, +0.79]$	-0.0226 $[-0.0238, -0.0214]$
Random C − Tail C	-1.12 $[-1.63, -0.61]$	$+0.0250$ $[+0.0167, +0.0334]$
Leading R − Leading C	$+2.48$ $[+1.44, +3.53]$	-0.0769 $[-0.0842, -0.0695]$
Leading C − Frozen	$+1.28$ $[+0.77, +1.80]$	-2.3814 $[-2.3883, -2.3745]$
Leading R − Frozen	$+3.77$ $[+3.23, +4.31]$	-2.4583 $[-2.4636, -2.4530]$
Middle R − Middle C	$+2.51$ $[+1.20, +3.82]$	-0.0776 $[-0.0816, -0.0736]$
Middle C − Frozen	$+1.39$ $[+0.66, +2.13]$	-2.3790 $[-2.3867, -2.3712]$
Middle R − Frozen	$+3.90$ $[+3.32, +4.49]$	-2.4566 $[-2.4604, -2.4527]$
Random R − Random C	$+2.57$ $[+2.26, +2.88]$	-0.0644 $[-0.0683, -0.0604]$
Random C − Frozen	$+1.67$ $[+1.20, +2.14]$	-2.4016 $[-2.4102, -2.3930]$
Random R − Frozen	$+4.23$ $[+3.92, +4.54]$	-2.4659 $[-2.4730, -2.4589]$
Leading R − Middle R	-0.14 $[-0.25, -0.02]$	-0.0017 $[-0.0055, +0.0020]$
Leading R − Tail R	-1.47 $[-1.83, -1.12]$	$+0.0266$ $[+0.0263, +0.0269]$
Middle R − Tail R	-1.34 $[-1.65, -1.03]$	$+0.0284$ $[+0.0244, +0.0324]$
Random R − Leading R	$+0.46$ $[-0.05, +0.98]$	-0.0076 $[-0.0114, -0.0039]$
Random R − Middle R	$+0.33$ $[-0.28, +0.94]$	-0.0094 $[-0.0128, -0.0059]$
Random R − Tail R	-1.01 $[-1.60, -0.42]$	$+0.0190$ $[+0.0150, +0.0230]$
Tail R − Tail C	$+2.46$ $[+2.10, +2.81]$	-0.0583 $[-0.0636, -0.0530]$
Tail C − Frozen	$+2.78$ $[+2.58, +2.99]$	-2.4266 $[-2.4269, -2.4263]$
Tail R − Frozen	$+5.24$ $[+4.83, +5.65]$	-2.4849 $[-2.4904, -2.4795]$

Table 42: Every original Qwen3B endpoint on CommonsenseQA. Correct counts are out of 1,221; accuracy is percent and NLL is nats per answer. C denotes coefficients and R coefficients plus rotations.

Support	Form	Seed	Correct	Accuracy (%)	Choice NLL
Leading	C	17	953	78.05	0.61456
Leading	C	42	958	78.46	0.61649
Leading	C	123	956	78.30	0.62004
Leading	R	17	989	81.00	0.54112
Leading	R	42	984	80.59	0.53774
Leading	R	123	985	80.67	0.54168
Middle	C	17	953	78.05	0.61770
Middle	C	42	958	78.46	0.61772
Middle	C	123	960	78.62	0.62312
Middle	R	17	991	81.16	0.54125
Middle	R	42	986	80.75	0.54083
Middle	R	123	986	80.75	0.54368
Tail	C	17	975	79.85	0.57181
Tail	C	42	973	79.69	0.57180
Tail	C	123	974	79.77	0.57200
Tail	R	17	1006	82.39	0.51462
Tail	R	42	1004	82.23	0.51103
Tail	R	123	1002	82.06	0.51499
Random	C	17	959	78.54	0.59458
Random	C	42	959	78.54	0.59528
Random	C	123	963	78.87	0.60090
Random	R	17	992	81.24	0.53203
Random	R	42	990	81.08	0.52999
Random	R	123	993	81.33	0.53562
Frozen reference			940	76.99	2.99848

B.15 QWEN3B: TWO ADDITIONAL SEEDS AND A SEPARATE FIVE-SEED VIEW

Two additional seeds preserve the direction of the original location comparisons. This post hoc supplement adds seeds 2026 and 2027 to every leading, middle, tail and random arm in both families: 16 new runs at the unchanged 842-update endpoint. It was registered after original outcomes had been inspected. The original three-seed cohort remains the primary report; Tables 43–45 report the new seeds and the separate five-seed view. Tail support has higher accuracy and lower choice NLL than the other supports in both families; rotated forms improve both outcomes at each support. These are comparisons between adapter families with different parameter counts.

All runs use the original native bf16-autocast scoring path and the same 1,221 public validation questions. The frozen reference is reused, not rerun or counted as an additional seed. Five-seed intervals use paired differences across seeds 17, 42, 123, 2026 and 2027, with a Student-t critical value at four degrees of freedom. They are nominal, exploratory and unadjusted for multiplicity; they do not establish equivalence or transfer the earlier decision categories. The analysis supplement retains the complete per-example arrays, original population and all 48 paired metric contrasts.

Table 43: Qwen3B post hoc native supplement. Mean±sample SD for the two new seeds and the combined five seeds; original-three results remain in Table 40. C: coefficients; R: coefficients plus rotations.

Arm	New-two acc. (%)	New-two NLL	Five-seed acc. (%)	Five-seed NLL
Leading C	78.17 ± 0.29	0.61442 ± 0.00017	78.23 ± 0.21	0.61599 ± 0.00243
Leading R	80.84 ± 0.35	0.53951 ± 0.00117	80.79 ± 0.24	0.53991 ± 0.00166
Middle C	78.46 ± 0.12	0.61533 ± 0.00017	78.41 ± 0.22	0.61784 ± 0.00318
Middle R	80.71 ± 0.17	0.54095 ± 0.00240	80.82 ± 0.21	0.54153 ± 0.00170
Tail C	79.93 ± 0.00	0.57002 ± 0.00130	79.84 ± 0.11	0.57113 ± 0.00121
Tail R	81.74 ± 0.46	0.51478 ± 0.00463	82.03 ± 0.37	0.51404 ± 0.00286
Random C	78.87 ± 0.35	0.59891 ± 0.00154	78.74 ± 0.25	0.59772 ± 0.00279
Random R	80.75 ± 0.12	0.53146 ± 0.00247	81.03 ± 0.28	0.53211 ± 0.00244

Table 44: All 16 additional native Qwen3B endpoints. Correct counts have denominator 1,221; the recipe and evaluation path match the original cohort.

Arm	Seed	Correct	Accuracy (%)	Choice NLL
Leading C	2026	957	78.38	0.61430
Leading C	2027	952	77.97	0.61455
Leading R	2026	990	81.08	0.53868
Leading R	2027	984	80.59	0.54033
Middle C	2026	957	78.38	0.61546
Middle C	2027	959	78.54	0.61521
Middle R	2026	987	80.84	0.53926
Middle R	2027	984	80.59	0.54265
Tail C	2026	976	79.93	0.56910
Tail C	2027	976	79.93	0.57093
Tail R	2026	1002	82.06	0.51151
Tail R	2027	994	81.41	0.51805
Random C	2026	960	78.62	0.60000
Random C	2027	966	79.12	0.59782
Random R	2026	987	80.84	0.52972
Random R	2027	985	80.67	0.53321

Table 45: All 48 paired metric contrasts in the separate post hoc five-seed native Qwen3B view. Each row gives accuracy difference in percentage points and choice-NLL difference, with nominal 95% paired intervals ($df = 4$).

Contrast	Δ accuracy (pp)	Δ choice NLL
Leading C − Middle C	-0.18 $[-0.50, +0.14]$	-0.00185 $[-0.00330, -0.00041]$
Leading C − Tail C	-1.61 $[-1.96, -1.25]$	$+0.04486$ $[+0.04235, +0.04737]$
Middle C − Tail C	-1.43 $[-1.75, -1.10]$	$+0.04671$ $[+0.04350, +0.04993]$
Random C − Leading C	$+0.51$ $[+0.00, +1.01]$	-0.01827 $[-0.02170, -0.01484]$
Random C − Middle C	$+0.33$ $[+0.08, +0.58]$	-0.02012 $[-0.02443, -0.01582]$
Random C − Tail C	-1.10 $[-1.38, -0.81]$	$+0.02659$ $[+0.02227, +0.03090]$
Leading R − Leading C	$+2.56$ $[+2.17, +2.95]$	-0.07608 $[-0.07906, -0.07310]$
Leading C − Frozen	$+1.24$ $[+0.98, +1.51]$	-2.38249 $[-2.38551, -2.37947]$
Leading R − Frozen	$+3.80$ $[+3.51, +4.09]$	-2.45857 $[-2.46062, -2.45651]$
Middle R − Middle C	$+2.41$ $[+1.88, +2.93]$	-0.07631 $[-0.07936, -0.07325]$
Middle C − Frozen	$+1.43$ $[+1.15, +1.70]$	-2.38064 $[-2.38458, -2.37669]$
Middle R − Frozen	$+3.83$ $[+3.57, +4.10]$	-2.45694 $[-2.45906, -2.45483]$
Random R − Random C	$+2.29$ $[+1.74, +2.85]$	-0.06560 $[-0.06914, -0.06207]$
Random C − Frozen	$+1.75$ $[+1.44, +2.06]$	-2.40076 $[-2.40422, -2.39730]$
Random R − Frozen	$+4.05$ $[+3.70, +4.39]$	-2.46636 $[-2.46939, -2.46334]$
Leading R − Middle R	-0.03 $[-0.24, +0.18]$	-0.00162 $[-0.00315, -0.00009]$
Leading R − Tail R	-1.24 $[-1.66, -0.83]$	$+0.02587$ $[+0.02336, +0.02838]$
Middle R − Tail R	-1.21 $[-1.51, -0.91]$	$+0.02749$ $[+0.02502, +0.02997]$
Random R − Leading R	$+0.25$ $[-0.19, +0.68]$	-0.00780 $[-0.00938, -0.00621]$
Random R − Middle R	$+0.21$ $[-0.08, +0.51]$	-0.00942 $[-0.01065, -0.00819]$
Random R − Tail R	-1.00 $[-1.30, -0.70]$	$+0.01807$ $[+0.01557, +0.02058]$
Tail R − Tail C	$+2.19$ $[+1.65, +2.74]$	-0.05709 $[-0.06058, -0.05360]$
Tail C − Frozen	$+2.85$ $[+2.72, +2.98]$	-2.42735 $[-2.42885, -2.42585]$
Tail R − Frozen	$+5.05$ $[+4.58, +5.51]$	-2.48444 $[-2.48799, -2.48088]$

B.16 A FIXED-RECIPE QWEN3B LORA COMPARISON AND SPECTRAL PROFILE

LoRA's accuracy contrast with tail rotations is unresolved: $+0.49$ percentage points $[-1.12, +2.11]$; its choice NLL is higher. We train rank-8 LoRA with scaling $\alpha = 8$ and zero dropout on the same 72 query/output matrices, using the Qwen3B CommonsenseQA recipe: 842 updates, learning rate 10^{-3}, effective batch 16 and 160-token inputs. Its three seeds are 17, 42 and 123. The trainable counts are approximately matched (2,359,296 LoRA parameters and 2,396,160 for the rotated family), while their structural freedoms differ. This is a post hoc comparison under a common recipe, not a separately tuned benchmark. Original Qwen3B outcomes and a selection-only LoRA diagnostic preceded confirmation registration; the three confirmation held-out evaluations followed it.

Native bf16 scoring uses all 1,221 public validation questions and reuses the accepted Qwen3B frozen reference. LoRA's accuracy contrast against tail rotations is unresolved, while its choice NLL is higher. The other seven adapter accuracy contrasts favor LoRA under nominal paired intervals. Table 47 retains all nine paired comparisons and both co-primary outcomes; full-vocabulary label NLL is a secondary diagnostic. These three LoRA seeds are not extended or pooled with later spectral seeds.

Table 46: Post hoc Qwen3B LoRA comparison on CommonsenseQA. Native bf16 accuracy and choice NLL, mean±sample SD over three seeds; the original frozen reference is reused.

Adapter	Accuracy (%)	Choice NLL
LoRA r8	82.72 ± 0.50	0.53760 ± 0.00200
Tail R	82.23 ± 0.16	0.51355 ± 0.00219
Frozen (reused)	76.99	2.99848

The profile describes where the learned update places energy in the starting checkpoint's frame. We sum block energies across 72 matrices before normalizing, with $L = [0, 1536)$ and $T = [1536, 2048)$. LoRA places about 37% in the cross blocks (Table 48); the dimension-only reference is 37.5%. This

Table 47: Every LoRA comparison with the original Qwen3B reference and eight arms. LoRA minus comparator, nominal unadjusted paired 95% intervals. Accuracy differences are percentage points; losses are nats per answer. Full-vocabulary NLL is secondary.

Contrast	Accuracy [95%]	Choice NLL [95%]	Full-vocabulary NLL [95%]
LoRA r8 − Frozen	$+5.73$ $[+4.50, +6.97]$	-2.4609 $[-2.4658, -2.4559]$	-2.4609 $[-2.4659, -2.4560]$
LoRA r8 − Leading C	$+4.45$ $[+3.21, +5.69]$	-0.0794 $[-0.0823, -0.0766]$	-0.0820 $[-0.0850, -0.0791]$
LoRA r8 − Middle C	$+4.34$ $[+3.53, +5.15]$	-0.0819 $[-0.0870, -0.0768]$	-0.0840 $[-0.0891, -0.0789]$
LoRA r8 − Tail C	$+2.95$ $[+1.68, +4.22]$	-0.0343 $[-0.0390, -0.0295]$	-0.0350 $[-0.0396, -0.0303]$
LoRA r8 − Random C	$+4.07$ $[+3.30, +4.84]$	-0.0593 $[-0.0647, -0.0540]$	-0.0609 $[-0.0664, -0.0555]$
LoRA r8 − Leading R	$+1.97$ $[+0.43, +3.50]$	-0.0026 $[-0.0100, +0.0049]$	-0.0028 $[-0.0103, +0.0048]$
LoRA r8 − Middle R	$+1.83$ $[+0.18, +3.47]$	-0.0043 $[-0.0081, -0.0006]$	-0.0044 $[-0.0083, -0.0006]$
LoRA r8 − Tail R	$+0.49$ $[-1.12, +2.11]$	$+0.0240$ $[+0.0163, +0.0318]$	$+0.0240$ $[+0.0163, +0.0318]$
LoRA r8 − Random R	$+1.50$ $[+0.46, +2.55]$	$+0.0050$ $[-0.0014, +0.0115]$	$+0.0049$ $[-0.0015, +0.0114]$

allocation does not show that learned directions are random or that the allocation causes performance. The profile reconstructs saved updates on CPU; the float32 intervention below is a separate evaluation.

Table 48: Every native LoRA endpoint and its spectral profile. Correct counts are out of 1,221; block columns and Cross are percentages of pooled squared update norm. ρ_F is the training record's pooled relative update norm.

Seed	Correct	Accuracy (%)	Choice NLL	ρ_F	LL	LT	TL	TT	Cross
17	1006	82.39	0.53542	0.0468	57.53	19.09	17.49	5.89	36.58
42	1007	82.47	0.53803	0.0445	57.16	19.31	17.54	5.99	36.85
123	1017	83.29	0.53935	0.0426	57.18	19.42	17.41	5.98	36.84

B.17 Llama checkpoint intervention: cross-region components versus shrinkage

Removing cross components costs more choice loss than shrinking the update. We intervene on the six fixed random-frame Llama checkpoints: coefficients and rotated adapters, each at seeds 17, 42 and 123 after 842 updates. The rotated family is primary and coefficients are secondary. The study is post hoc: frozen and some trained held-out outcomes were already inspected. The intervention protocol fixed the variants and contrasts before their held-out evaluations; selection examples were used for numerical parity and timing. All five variants and one matching frozen reference are evaluated on all 1,221 CommonsenseQA validation questions. No checkpoint is retrained or selected by intervention outcome.

The control matches each matrix's remaining update norm. In the starting weight's singular frame, let Δ_{within} contain the two within-region blocks and Δ_{cross} the two cross-region blocks of the attained update, so $\Delta = \Delta_{\text{within}} + \Delta_{\text{cross}}$. Here the tail is $[3584, 4096)$ and the leading region is its entire complement $[0, 3584)$, across all 64 query and output projections. For $g \in \{1/2, 0\}$, the intervention is $\Delta_{\text{within}} + g\Delta_{\text{cross}}$; its comparator is $s_g\Delta$, where $s_g = \|\Delta_{\text{within}} + g\Delta_{\text{cross}}\|_F / \|\Delta\|_F$ is computed separately for each matrix. Thus each pair has the same remaining Frobenius norm, while the comparator retains the original block proportions. The shrink rows identify their matching gate g. The original update is the fifth variant. Removal changes direction and can change rank; matched remaining norm does not match distance from the trained update or isolate an effect of rank.

Every intervention uses one common float32 path. The whole model is upcast, including frozen projections; autocast and TF32 are disabled, evaluation batch size is eight, and the base weights are restored and verified between variants. Selection checks compare reconstructed and native original updates in float32: predictions agree exactly, maximum absolute choice-logit difference is 8.6×10^{-5} and maximum mean-NLL difference is 4.4×10^{-7}. Native bf16 reload checks are separate. On held-out questions, switching the original checkpoint from native bf16 to float32 changes 17/9/11 predictions for the rotated family and 10/23/14 for coefficients. These path differences do not enter

the intervention contrast, and the float32 frozen reference below is separate from the native reference in Table 30. Exported matrix reconstruction and remaining-norm residuals satisfy the declared tolerances; the observed float32 reconstruction residual is a supporting diagnostic, not a separately enforced runtime gate.

Table 49: Post hoc Llama checkpoint intervention on CommonsenseQA. Mean±sample SD over the same three seeds; all evaluations use 1,221 questions and the common float32 path. "Shrink" rescales the original update separately in each matrix to match the corresponding intervention's remaining norm.

Family	Variant	Accuracy (%)	Choice NLL
Random R	Original	79.69 ± 0.22	0.5448 ± 0.0024
Random R	Half cross	79.14 ± 0.21	0.5559 ± 0.0015
Random R	Shrink ($g = 1/2$)	79.31 ± 0.13	0.5495 ± 0.0019
Random R	Zero cross	78.62 ± 0.25	0.5789 ± 0.0007
Random R	Shrink ($g = 0$)	79.22 ± 0.21	0.5516 ± 0.0018
Random C	Original	77.91 ± 0.13	0.5924 ± 0.0017
Random C	Half cross	77.72 ± 0.00	0.6020 ± 0.0010
Random C	Shrink ($g = 1/2$)	77.78 ± 0.09	0.5958 ± 0.0012
Random C	Zero cross	77.53 ± 0.13	0.6272 ± 0.0015
Random C	Shrink ($g = 0$)	77.83 ± 0.05	0.5976 ± 0.0010
Frozen (same float32 path)		76.58	0.9749

Table 50: All registered Llama intervention contrasts, first minus second: mean and nominal paired 95% interval over three seeds (two degrees of freedom), unadjusted. Zero-cross minus its matched shrink is primary within each family; the rotated family is primary overall. Accuracy is percentage points and NLL is nats per answer.

Family	Contrast	Accuracy difference [95% interval]	NLL difference [95% interval]
Random R	Zero cross − shrink ($g = 0$)	-0.60 $[-1.54, +0.34]$	$+0.0273$ $[+0.0245, +0.0301]$
Random R	Original − zero cross	$+1.06$ $[+0.18, +1.95]$	-0.0341 $[-0.0384, -0.0298]$
Random R	Half cross − shrink ($g = 1/2$)	-0.16 $[-0.37, +0.04]$	$+0.0064$ $[+0.0053, +0.0075]$
Random R	Original − half cross	$+0.55$ $[+0.24, +0.86]$	-0.0111 $[-0.0135, -0.0087]$
Random C	Zero cross − shrink ($g = 0$)	-0.30 $[-0.72, +0.12]$	$+0.0295$ $[+0.0264, +0.0327]$
Random C	Original − zero cross	$+0.38$ $[+0.07, +0.69]$	-0.0347 $[-0.0402, -0.0293]$
Random C	Half cross − shrink ($g = 1/2$)	-0.05 $[-0.29, +0.18]$	$+0.0063$ $[+0.0042, +0.0083]$
Random C	Original − half cross	$+0.19$ $[-0.12, +0.50]$	-0.0096 $[-0.0135, -0.0057]$

Both families have higher choice loss after complete cross removal than after matched shrinking; the same loss ordering holds for the half-cross pair. The original three-seed primary accuracy intervals include zero, which establishes neither preserved accuracy nor equivalence. With five seeds, the post hoc complete-removal contrast has a negative accuracy interval for the rotated family, while the coefficient-family contrast remains unresolved (Table 36). Against its matched shrink, complete removal changes 19/18/22 predictions for the rotated family and 13/15/14 for coefficients. Original cross energy is close to the dimension-only reference $2(3584)(512)/4096^2 = 21.875\%$; that agreement describes allocation and does not identify its cause. The experiment measures functional sensitivity of these fixed checkpoints, not training-time necessity or unique importance of the cross blocks.

B.18 Llama checkpoint intervention: additional seeds

The five-seed intervention view preserves the choice-loss cost of removing cross-region components relative to matched shrinking in both families. Its rotated-family accuracy contrast is negative with a nominal interval below zero; the coefficient-family accuracy contrast remains unresolved. This is a separate post hoc view. The original three-seed result remains primary and unchanged, and the additional two-seed estimates remain visible in Tables 52 and 53.

The extension evaluates four checkpoints under five variants, plus the identical float32 frozen reference: 21 evaluations. The combined view contains ten checkpoints under five variants and

Table 51: Every Llama checkpoint intervention outcome. Correct counts are out of 1,221; all three seeds and all five variants are retained. The same-path frozen model answers 935 questions correctly, with accuracy 76.58% and choice NLL 0.97489. Machine-readable records retain per-example predictions, losses, source bindings, geometry and numerical-path differences.

Family	Seed	Variant	Correct	Accuracy (%)	Choice NLL
Random R	17	Original	976	79.93	0.54228
Random R	17	Half cross	969	79.36	0.55449
Random R	17	Shrink ($g = 1/2$)	970	79.44	0.54758
Random R	17	Zero cross	960	78.62	0.57823
Random R	17	Shrink ($g = 0$)	970	79.44	0.54994
Random R	42	Original	971	79.52	0.54512
Random R	42	Half cross	966	79.12	0.55572
Random R	42	Shrink ($g = 1/2$)	968	79.28	0.54949
Random R	42	Zero cross	963	78.87	0.57904
Random R	42	Shrink ($g = 0$)	965	79.03	0.55154
Random R	123	Original	972	79.61	0.54703
Random R	123	Half cross	964	78.95	0.55752
Random R	123	Shrink ($g = 1/2$)	967	79.20	0.55142
Random R	123	Zero cross	957	78.38	0.57953
Random R	123	Shrink ($g = 0$)	967	79.20	0.55345
Random C	17	Original	953	78.05	0.59271
Random C	17	Half cross	949	77.72	0.60308
Random C	17	Shrink ($g = 1/2$)	951	77.89	0.59638
Random C	17	Zero cross	947	77.56	0.62888
Random C	17	Shrink ($g = 0$)	950	77.81	0.59837
Random C	42	Original	951	77.89	0.59398
Random C	42	Half cross	949	77.72	0.60179
Random C	42	Shrink ($g = 1/2$)	949	77.72	0.59649
Random C	42	Zero cross	948	77.64	0.62620
Random C	42	Shrink ($g = 0$)	950	77.81	0.59808
Random C	123	Original	950	77.81	0.59058
Random C	123	Half cross	949	77.72	0.60118
Random C	123	Shrink ($g = 1/2$)	949	77.72	0.59442
Random C	123	Zero cross	945	77.40	0.62643
Random C	123	Shrink ($g = 0$)	951	77.89	0.59644

one shared frozen reference. Random R remains the primary family and random C the secondary family. The registered two-seed intervals use one degree of freedom and are descriptive; combined intervals use four degrees of freedom. Both are nominal, unadjusted and based on paired training seeds. Tables 54 and 55 report the full combined results.

All treatment contrasts use the common dense float32 path. Changes between native bf16 and float32 evaluation remain separate. Matched shrinking controls each matrix's remaining update norm, while effective rank and distance from the trained update differ. This checkpoint sensitivity does not identify training-time necessity, an isolated rank effect or a unique mechanism; an unresolved accuracy contrast is not equivalence.

Table 52: All 20 additional Llama checkpoint-treatment endpoints, seeds 2026 and 2027. Correct counts have denominator 1,221; all treatments use the common dense float32 path. The shared frozen reference appears in Table 54.

Family	Seed	Variant	Correct	Accuracy (%)	Choice NLL
Random R	2026	Original	968	79.28	0.54220
Random R	2026	Half cross	965	79.03	0.55557
Random R	2026	Shrink ($g = 1/2$)	964	78.95	0.54802
Random R	2026	Zero cross	961	78.71	0.58100
Random R	2026	Shrink ($g = 0$)	964	78.95	0.55057
Random R	2027	Original	969	79.36	0.54841
Random R	2027	Half cross	968	79.28	0.55816
Random R	2027	Shrink ($g = 1/2$)	967	79.20	0.55250
Random R	2027	Zero cross	960	78.62	0.57891
Random R	2027	Shrink ($g = 0$)	968	79.28	0.55443
Random C	2026	Original	950	77.81	0.59432
Random C	2026	Half cross	952	77.97	0.60374
Random C	2026	Shrink ($g = 1/2$)	947	77.56	0.59749
Random C	2026	Zero cross	947	77.56	0.62924
Random C	2026	Shrink ($g = 0$)	947	77.56	0.59931
Random C	2027	Original	949	77.72	0.59923
Random C	2027	Half cross	946	77.48	0.60710
Random C	2027	Shrink ($g = 1/2$)	950	77.81	0.60182
Random C	2027	Zero cross	945	77.40	0.63016
Random C	2027	Shrink ($g = 0$)	947	77.56	0.60345

Table 53: All eight Llama intervention contrasts for the two additional seeds. Accuracy differences are percentage points. Brackets are the protocol-specified nominal 95% paired intervals ($df = 1$); these wide intervals are descriptive. Per-seed endpoints are in Table 52.

Family	Contrast	Δ accuracy (pp)	Δ choice NLL
Random R	Zero cross − Shrink ($g = 0$)	-0.45 $[-3.05, +2.15]$	$+0.02746$ $[-0.01033, +0.06524]$
Random R	Original − Zero cross	$+0.66$ $[-0.39, +1.70]$	-0.03465 $[-0.08742, +0.01812]$
Random R	Half cross − Shrink ($g = 1/2$)	$+0.08$ $[+0.08, +0.08]$	$+0.00661$ $[-0.00538, +0.01860]$
Random R	Original − Half cross	$+0.16$ $[-0.88, +1.20]$	-0.01156 $[-0.03455, +0.01143]$
Random C	Zero cross − Shrink ($g = 0$)	-0.08 $[-1.12, +0.96]$	$+0.02832$ $[+0.00783, +0.04880]$
Random C	Original − Zero cross	$+0.29$ $[-0.23, +0.81]$	-0.03292 $[-0.05829, -0.00755]$
Random C	Half cross − Shrink ($g = 1/2$)	$+0.04$ $[-4.64, +4.72]$	$+0.00576$ $[-0.00045, +0.01198]$
Random C	Original − Half cross	$+0.04$ $[-2.56, +2.64]$	-0.00864 $[-0.01852, +0.00123]$

Table 54: Separate post hoc five-seed Llama checkpoint-intervention supplement. Mean±sample SD over the five training seeds, plus the single reused float32 frozen reference.

Family	Variant	Accuracy (%)	Choice NLL
Random R	Original	79.54 ± 0.26	0.54501 ± 0.00278
Random R	Half cross	79.15 ± 0.17	0.55629 ± 0.00151
Random R	Shrink ($g = 1/2$)	79.21 ± 0.18	0.54980 ± 0.00213
Random R	Zero cross	78.64 ± 0.18	0.57934 ± 0.00104
Random R	Shrink ($g = 0$)	79.18 ± 0.20	0.55198 ± 0.00190
Random C	Original	77.85 ± 0.12	0.59417 ± 0.00319
Random C	Half cross	77.72 ± 0.17	0.60338 ± 0.00231
Random C	Shrink ($g = 1/2$)	77.74 ± 0.12	0.59732 ± 0.00275
Random C	Zero cross	77.51 ± 0.11	0.62818 ± 0.00177
Random C	Shrink ($g = 0$)	77.72 ± 0.15	0.59913 ± 0.00263
Frozen	Reused float32 reference	76.58	0.97489

Table 55: All eight paired Llama intervention contrasts in the post hoc five-seed view. Accuracy differences are percentage points; brackets are nominal 95% paired intervals ($df = 4$). Zero cross minus matched shrinking is the primary contrast within each family; R remains the primary family and C secondary.

Family	Contrast	Δ accuracy (pp)	Δ choice NLL
Random R	Zero cross − Shrink ($g = 0$)	-0.54 $[-0.93, -0.15]$	$+0.02736$ $[+0.02457, +0.03015]$
Random R	Original − Zero cross	$+0.90$ $[+0.48, +1.33]$	-0.03433 $[-0.03830, -0.03036]$
Random R	Half cross − Shrink ($g = 1/2$)	-0.07 $[-0.25, +0.12]$	$+0.00649$ $[+0.00557, +0.00741]$
Random R	Original − Half cross	$+0.39$ $[+0.10, +0.68]$	-0.01128 $[-0.01311, -0.00945]$
Random C	Zero cross − Shrink ($g = 0$)	-0.21 $[-0.44, +0.01]$	$+0.02905$ $[+0.02707, +0.03103]$
Random C	Original − Zero cross	$+0.34$ $[+0.21, +0.48]$	-0.03402 $[-0.03690, -0.03113]$
Random C	Half cross − Shrink ($g = 1/2$)	-0.02 $[-0.36, +0.32]$	$+0.00606$ $[+0.00515, +0.00696]$
Random C	Original − Half cross	$+0.13$ $[-0.10, +0.37]$	-0.00921 $[-0.01087, -0.00756]$

B.19 QWEN-1.5B CHECKPOINT INTERVENTION

Cross removal raises loss, with a coefficient accuracy cost. We apply the same intervention design to the six original Qwen2.5-1.5B-Instruct random-frame checkpoints on CommonsenseQA: Random R (run label RAND_ROT128, primary family) and Random C (RAND_DIAG, secondary), each at seeds 17, 42 and 123 and the fixed 842-update endpoint. Every checkpoint has all five variants on all 1,221 validation questions: 30 evaluations in total. The sealed population contains no new float32 frozen evaluation, so we make no matched-path frozen comparison. The study is post hoc on previously inspected validation data. The final protocol, all variants and the cost-based admission decision were fixed before held-out treatment evaluation; 974 selection questions were used for parity and timing.

The partition differs from Llama's. All 56 query/output projections have width 1,536. The tail is $[1024, 1536)$, and the leading region is its full complement $[0, 1024)$. With Δ_{within} and Δ_{cross} defined as in Appendix B.17, we compare $\Delta_{\text{within}} + g\Delta_{\text{cross}}$ with $s_g\Delta$ for $g \in \{1/2, 0\}$, matching the remaining update norm separately in every matrix. Original cross energy spans 44.428–44.454%, close to the $4/9$ dimension-only reference for this partition. The Qwen and Llama effect magnitudes are separate model/protocol observations: neither a pure model-family effect nor a scale effect is isolated.

A common float32 path separates the intervention from rounding. Original and treated updates use float32 inference, with autocast and TF32 disabled, batch size eight, and base weights restored and verified between variants. All six native-versus-dense float32 selection checks have identical predictions; maximum absolute choice-logit difference is 1.05×10^{-4} and maximum mean-NLL difference 4.49×10^{-7}. Finite per-matrix reconstruction and norm-match gates pass, with respective maximum relative residuals 3.35×10^{-6} and 8.22×10^{-9}. Native bf16 reload checks are separate. Switching the original held-out evaluation from bf16 to float32 changes 8/11/19 predictions for rota-

tions and 20/24/16 for coefficients; these changes are not intervention effects and are not substituted as the treatment baseline.

Table 56: Post hoc Qwen2.5-1.5B checkpoint intervention on CommonsenseQA. Mean±sample SD over three seeds; all 30 evaluations use the same 1,221 questions and float32 path. Shrink rows match the intervention's remaining norm per matrix, using the indicated gate g. No new matching float32 frozen evaluation is included.

Family	Variant	Accuracy (%)	Choice NLL
Random R	Original	77.48 ± 0.08	0.6178 ± 0.0025
Random R	Half cross	76.39 ± 0.29	0.6660 ± 0.0014
Random R	Shrink ($g = 1/2$)	76.85 ± 0.29	0.6471 ± 0.0019
Random R	Zero cross	75.05 ± 0.29	0.8081 ± 0.0028
Random R	Shrink ($g = 0$)	76.33 ± 0.36	0.6678 ± 0.0020
Random C	Original	76.19 ± 0.24	0.6554 ± 0.0014
Random C	Half cross	74.91 ± 0.13	0.7085 ± 0.0014
Random C	Shrink ($g = 1/2$)	74.99 ± 0.19	0.6853 ± 0.0016
Random C	Zero cross	73.79 ± 0.22	0.8815 ± 0.0043
Random C	Shrink ($g = 0$)	74.69 ± 0.14	0.7094 ± 0.0019

Table 57: Every registered Qwen1.5B intervention contrast, first minus second: mean and nominal paired 95% interval over three seeds (two degrees of freedom), unadjusted. Zero-cross minus its matched shrink is primary within each family; the rotated family is primary overall. Accuracy differences are percentage points and NLL differences are nats per answer.

Family	Contrast	Accuracy difference [95% interval]	NLL difference [95% interval]
Random R	Zero cross − shrink ($g = 0$)	-1.28 $[-2.82, +0.26]$	$+0.1403$ $[+0.1359, +0.1447]$
Random R	Original − zero cross	$+2.43$ $[+1.51, +3.35]$	-0.1903 $[-0.1992, -0.1815]$
Random R	Half cross − shrink ($g = 1/2$)	-0.46 $[-1.18, +0.25]$	$+0.0188$ $[+0.0177, +0.0199]$
Random R	Original − half cross	$+1.09$ $[+0.47, +1.71]$	-0.0481 $[-0.0521, -0.0442]$
Random C	Zero cross − shrink ($g = 0$)	-0.90 $[-1.71, -0.09]$	$+0.1720$ $[+0.1646, +0.1795]$
Random C	Original − zero cross	$+2.40$ $[+1.38, +3.43]$	-0.2261 $[-0.2357, -0.2164]$
Random C	Half cross − shrink ($g = 1/2$)	-0.08 $[-0.69, +0.53]$	$+0.0232$ $[+0.0222, +0.0243]$
Random C	Original − half cross	$+1.28$ $[+0.44, +2.13]$	-0.0531 $[-0.0554, -0.0509]$

Removal raises choice loss beyond matched shrinking in both families, for both the zero-cross and half-cross pairs. The primary rotated accuracy interval includes zero; the secondary coefficient interval lies below zero. These findings do not establish preserved accuracy or equivalence. Zero-cross removal versus its matched shrink changes 47/52/47 predictions for rotations and 50/43/44 for coefficients. In the rotated family, respectively 32/25/31 examples are wrong under removal and right under shrinking, whereas 11/18/12 show the reverse. As with Llama, matching the remaining norm does not match distance from the trained update or isolate rank. The result is fixed-checkpoint sensitivity, not training-time necessity or unique importance against all perturbations.

B.20 QWEN3B CHECKPOINT INTERVENTIONS: RANDOM FRAMES AND LORA

Cross-region removal has a loss cost in both the random-frame and LoRA checkpoints. We evaluate five fixed variants per checkpoint: original, half cross, zero cross, and whole-update shrink controls matched to each intervention's remaining norm per matrix. The original random cohort contains six checkpoints (Random R, run label RAND_ROT128, primary; Random C, RAND_DIAG, secondary; seeds 17, 42 and 123); the separate LoRA cohort contains three rank-8 checkpoints at the same seeds. All native endpoints are at 842 updates. Each block includes its own same-path frozen evaluation, giving 31 and 16 evaluations, respectively. The two frozen per-example files are byte-identical and are not two independent references.

These studies are post hoc, after inspection of native outcomes. Each block's protocol fixes the population, variants and matched-shrink comparisons before its intervention evaluation. All variants score the same 1,221 public validation questions through a float32 dense path with autocast and TF32 disabled, batch size eight, and verified base restoration. Native bf16 results remain separate.

Table 58: Every original Qwen1.5B checkpoint intervention outcome. Correct counts are out of 1,221. All three seeds and all five variants are retained. The distributed records preserve per-example predictions and losses, protocol/source bindings, geometry and numerical-path diagnostics.

Family	Seed	Variant	Correct	Accuracy (%)	Choice NLL
Random R	17	Original	947	77.56	0.61490
Random R	17	Half cross	933	76.41	0.66466
Random R	17	Shrink ($g = 1/2$)	942	77.15	0.64536
Random R	17	Zero cross	913	74.77	0.80802
Random R	17	Shrink ($g = 0$)	934	76.49	0.66627
Random R	42	Original	945	77.40	0.61910
Random R	42	Half cross	929	76.09	0.66569
Random R	42	Shrink ($g = 1/2$)	935	76.58	0.64696
Random R	42	Zero cross	920	75.35	0.80540
Random R	42	Shrink ($g = 0$)	927	75.92	0.66710
Random R	123	Original	946	77.48	0.61944
Random R	123	Half cross	936	76.66	0.66752
Random R	123	Shrink ($g = 1/2$)	938	76.82	0.64911
Random R	123	Zero cross	916	75.02	0.81094
Random R	123	Shrink ($g = 0$)	935	76.58	0.67011
Random C	17	Original	932	76.33	0.65441
Random C	17	Half cross	915	74.94	0.70854
Random C	17	Shrink ($g = 1/2$)	913	74.77	0.68486
Random C	17	Zero cross	898	73.55	0.88373
Random C	17	Shrink ($g = 0$)	913	74.77	0.70904
Random C	42	Original	927	75.92	0.65480
Random C	42	Half cross	916	75.02	0.70714
Random C	42	Shrink ($g = 1/2$)	917	75.10	0.68397
Random C	42	Zero cross	903	73.96	0.87656
Random C	42	Shrink ($g = 0$)	910	74.53	0.70775
Random C	123	Original	932	76.33	0.65701
Random C	123	Half cross	913	74.77	0.70996
Random C	123	Shrink ($g = 1/2$)	917	75.10	0.68710
Random C	123	Zero cross	902	73.87	0.88413
Random C	123	Shrink ($g = 0$)	913	74.77	0.71150

We use the same $L = [0, 1536)$, $T = [1536, 2048)$ partition for both blocks. The registered numerical, reconstruction and remaining-norm gates pass; Numerical-path parity checks do not establish unchanged task behavior.

Removing cross terms versus matched shrinking reduces accuracy for both original random families and raises their choice NLL. LoRA's loss contrast is also positive, while its accuracy contrast is unresolved. The smaller observed LoRA loss cost is descriptive across different families, not an isolated effect of rank, rotation or parameter count. Tables 60 and 62 retain every registered pair. Matching remaining norm does not match distance from the trained endpoint or effective rank. These interventions establish fixed-checkpoint sensitivity, not training-time necessity, equivalence or recovery.

Table 59: Qwen3B original random checkpoints: all variants, mean±sample SD over three seeds. Random R is primary; Random C is secondary. All 31 evaluations use the same float32 path and 1,221 examples; frozen is evaluated once.

Family	Variant	Accuracy (%)	Choice NLL
Random R	Original	81.27 ± 0.24	0.53137 ± 0.00271
Random R	Half cross	80.62 ± 0.29	0.61845 ± 0.00659
Random R	Shrink ($g = 1/2$)	80.89 ± 0.19	0.57271 ± 0.00374
Random R	Zero cross	79.96 ± 0.13	0.92685 ± 0.01736
Random R	Shrink ($g = 0$)	80.81 ± 0.25	0.60465 ± 0.00464
Random C	Original	78.68 ± 0.29	0.59638 ± 0.00373
Random C	Half cross	78.68 ± 0.21	0.71525 ± 0.00239
Random C	Shrink ($g = 1/2$)	78.87 ± 0.22	0.65074 ± 0.00324
Random C	Zero cross	78.35 ± 0.17	1.09142 ± 0.01339
Random C	Shrink ($g = 0$)	78.76 ± 0.24	0.69345 ± 0.00370
Frozen	Same float32 path	76.82	3.01470

Table 60: Every registered Qwen3B original-random intervention pair: first minus second, nominal unadjusted paired 95% intervals. Accuracy differences are percentage points and choice NLL differences nats per answer.

Family	Contrast	Accuracy difference [95%]	Choice NLL difference [95%]
Random R	Zero cross − Shrink ($g = 0$)	-0.85 $[-1.16, -0.54]$	$+0.32220$ $[+0.28662, +0.35779]$
Random R	Original − Zero cross	$+1.31$ $[+0.90, +1.72]$	-0.39548 $[-0.43737, -0.35359]$
Random R	Half cross − Shrink ($g = 1/2$)	-0.27 $[-1.10, +0.55]$	$+0.04574$ $[+0.03727, +0.05422]$
Random R	Original − Half cross	$+0.66$ $[-0.23, +1.54]$	-0.08708 $[-0.09993, -0.07423]$
Random C	Zero cross − Shrink ($g = 0$)	-0.41 $[-0.61, -0.21]$	$+0.39797$ $[+0.35751, +0.43843]$
Random C	Original − Zero cross	$+0.33$ $[-0.21, +0.87]$	-0.49504 $[-0.53601, -0.45406]$
Random C	Half cross − Shrink ($g = 1/2$)	-0.19 $[-1.19, +0.81]$	$+0.06451$ $[+0.05474, +0.07428]$
Random C	Original − Half cross	-0.00 $[-0.73, +0.73]$	-0.11886 $[-0.12990, -0.10782]$

Table 61: Qwen3B LoRA checkpoints: all variants, mean±sample SD over three seeds. All 16 evaluations use the same float32 path and 1,221 examples. The frozen file matches the original-random block's reference exactly.

Family	Variant	Accuracy (%)	Choice NLL
LoRA r8	Original	82.83 ± 0.53	0.53668 ± 0.00201
LoRA r8	Half cross	83.54 ± 0.67	0.52857 ± 0.00389
LoRA r8	Shrink ($g = 1/2$)	83.27 ± 0.67	0.52599 ± 0.00469
LoRA r8	Zero cross	83.51 ± 0.34	0.55066 ± 0.00244
LoRA r8	Shrink ($g = 0$)	83.51 ± 0.59	0.52504 ± 0.00701
Frozen	Same float32 path	76.82	3.01470

Table 62: Every registered Qwen3B LoRA intervention pair: first minus second, nominal unadjusted paired 95% intervals. Accuracy differences are percentage points and choice NLL differences nats per answer.

Family	Contrast	Accuracy difference [95%]	Choice NLL difference [95%]
LoRA r8	Zero cross − Shrink ($g = 0$)	$+0.00$ $[-1.42, +1.42]$	$+0.02562$ $[+0.01228, +0.03895]$
LoRA r8	Original − Zero cross	-0.68 $[-1.64, +0.28]$	-0.01398 $[-0.01692, -0.01104]$
LoRA r8	Half cross − Shrink ($g = 1/2$)	$+0.27$ $[-0.44, +0.99]$	$+0.00258$ $[-0.00013, +0.00529]$
LoRA r8	Original − Half cross	-0.71 $[-1.13, -0.29]$	$+0.00811$ $[+0.00046, +0.01577]$

B.21 QWEN3B CHECKPOINT INTERVENTION: ADDITIONAL SEEDS

The additional checkpoints support the original intervention contrasts. The two new seeds contribute four random-frame checkpoints, each evaluated under all five variants, plus the same float32 frozen reference: 21 evaluations. Random R remains the primary family and random C the secondary family. Tables 63 and 64 report every new endpoint and contrast without a two-seed interval; Tables 65 and 66 give the separate post hoc five-seed view. Removing cross-region components increases choice NLL and lowers accuracy relative to matched shrinking in both families in that combined view.

The original three-seed results remain unchanged. The combined analysis is reconstructed from the exact original and new per-example exports, using paired training-seed differences and nominal, unadjusted 95% Student-t intervals with four degrees of freedom. Its population is ten checkpoints times five variants, with one shared frozen reference, not two independent frozen evaluations. Native bf16 outcomes remain separate from these common float32 treatment comparisons. Matched shrinking controls each matrix's remaining update norm; rank and distance from the trained update differ. This post hoc checkpoint sensitivity does not establish training-time necessity, an isolated rank effect or accuracy equivalence.

Table 63: All 20 new Qwen3B checkpoint-treatment endpoints, seeds 2026 and 2027. Correct counts have denominator 1,221; every treatment uses the same dense float32 path. The shared frozen reference appears in Table 65.

Family	Seed	Variant	Correct	Accuracy (%)	Choice NLL
Random R	2026	Original	990	81.08	0.53058
Random R	2026	Half cross	980	80.26	0.61692
Random R	2026	Shrink ($g = 1/2$)	982	80.43	0.57053
Random R	2026	Zero cross	977	80.02	0.93034
Random R	2026	Shrink ($g = 0$)	981	80.34	0.60237
Random R	2027	Original	989	81.00	0.53218
Random R	2027	Half cross	981	80.34	0.61666
Random R	2027	Shrink ($g = 1/2$)	982	80.43	0.57090
Random R	2027	Zero cross	975	79.85	0.92783
Random R	2027	Shrink ($g = 0$)	981	80.34	0.60282
Random C	2026	Original	963	78.87	0.59644
Random C	2026	Half cross	962	78.79	0.70766
Random C	2026	Shrink ($g = 1/2$)	962	78.79	0.64760
Random C	2026	Zero cross	956	78.30	1.07004
Random C	2026	Shrink ($g = 0$)	962	78.79	0.68886
Random C	2027	Original	961	78.71	0.59707
Random C	2027	Half cross	964	78.95	0.71638
Random C	2027	Shrink ($g = 1/2$)	965	79.03	0.65274
Random C	2027	Zero cross	956	78.30	1.08905
Random C	2027	Shrink ($g = 0$)	964	78.95	0.69624

Table 64: All eight intervention contrasts for the two additional seeds. Accuracy differences are percentage points; each metric reports seed 2026, seed 2027 and their mean. No interval is reported for this two-seed view.

Family	Contrast	Acc. 2026	Acc. 2027	Acc. mean	NLL 2026	NLL 2027	NLL mean
Random R	Zero cross − Shrink ($g = 0$)	-0.33	-0.49	-0.41	$+0.32797$	$+0.32501$	$+0.32649$
Random R	Original − Zero cross	$+1.06$	$+1.15$	$+1.11$	-0.39976	-0.39566	-0.39771
Random R	Half cross − Shrink ($g = 1/2$)	-0.16	-0.08	-0.12	$+0.04639$	$+0.04576$	$+0.04608$
Random R	Original − Half cross	$+0.82$	$+0.66$	$+0.74$	-0.08634	-0.08449	-0.08541
Random C	Zero cross − Shrink ($g = 0$)	-0.49	-0.66	-0.57	$+0.38117$	$+0.39282$	$+0.38699$
Random C	Original − Zero cross	$+0.57$	$+0.41$	$+0.49$	-0.47359	-0.49198	-0.48278
Random C	Half cross − Shrink ($g = 1/2$)	$+0.00$	-0.08	-0.04	$+0.06006$	$+0.06364$	$+0.06185$
Random C	Original − Half cross	$+0.08$	-0.25	-0.08	-0.11122	-0.11931	-0.11526

Table 65: Separate post hoc five-seed Qwen3B checkpoint-intervention supplement. Mean±sample SD across all five seeds. The single frozen row reuses the exact common float32 reference from the original and extension exports.

Family	Variant	Accuracy (%)	Choice NLL
Random R	Original	81.18 ± 0.21	0.53138 ± 0.00200
Random R	Half cross	80.49 ± 0.27	0.61779 ± 0.00475
Random R	Shrink ($g = 1/2$)	80.70 ± 0.29	0.57191 ± 0.00286
Random R	Zero cross	79.95 ± 0.11	0.92775 ± 0.01237
Random R	Shrink ($g = 0$)	80.62 ± 0.31	0.60383 ± 0.00347
Random C	Original	78.72 ± 0.22	0.59653 ± 0.00265
Random C	Half cross	78.76 ± 0.19	0.71396 ± 0.00394
Random C	Shrink ($g = 1/2$)	78.89 ± 0.18	0.65051 ± 0.00294
Random C	Zero cross	78.33 ± 0.12	1.08667 ± 0.01331
Random C	Shrink ($g = 0$)	78.80 ± 0.19	0.69309 ± 0.00372
Frozen	Reused float32 reference	76.82	3.01470

Table 66: All eight paired Qwen3B intervention contrasts in the post hoc five-seed view. Accuracy differences are percentage points; brackets are nominal 95% paired intervals ($df = 4$). Zero cross minus its matched shrinking control is the primary contrast within each family.

Family	Contrast	Δ accuracy (pp)	Δ choice NLL
Random R	Zero cross − Shrink ($g = 0$)	-0.67 $[-1.00, -0.35]$	$+0.32392$ $[+0.31094, +0.33690]$
Random R	Original − Zero cross	$+1.23$ $[+1.03, +1.43]$	-0.39637 $[-0.41136, -0.38138]$
Random R	Half cross − Shrink ($g = 1/2$)	-0.21 $[-0.52, +0.10]$	$+0.04588$ $[+0.04286, +0.04889]$
Random R	Original − Half cross	$+0.69$ $[+0.36, +1.01]$	-0.08641 $[-0.09116, -0.08166]$
Random C	Zero cross − Shrink ($g = 0$)	-0.48 $[-0.63, -0.32]$	$+0.39358$ $[+0.37666, +0.41050]$
Random C	Original − Zero cross	$+0.39$ $[+0.16, +0.63]$	-0.49013 $[-0.50869, -0.47158]$
Random C	Half cross − Shrink ($g = 1/2$)	-0.13 $[-0.50, +0.24]$	$+0.06345$ $[+0.05924, +0.06765]$
Random C	Original − Half cross	-0.03 $[-0.33, +0.27]$	-0.11742 $[-0.12324, -0.11160]$

B.22 ORIGINAL ENCODER OUTCOMES AND THE FROZEN-HEAD PROBE

Table 67 preserves the original inner-selection band results, Table 68 the native held-out interaction results, and Figure 5 plots the 18 interaction endpoints. The main tables report the subsequent held-out band outcomes with native loss; temperature-corrected outcomes are in Appendix C.10. The two evaluation splits are kept distinct.

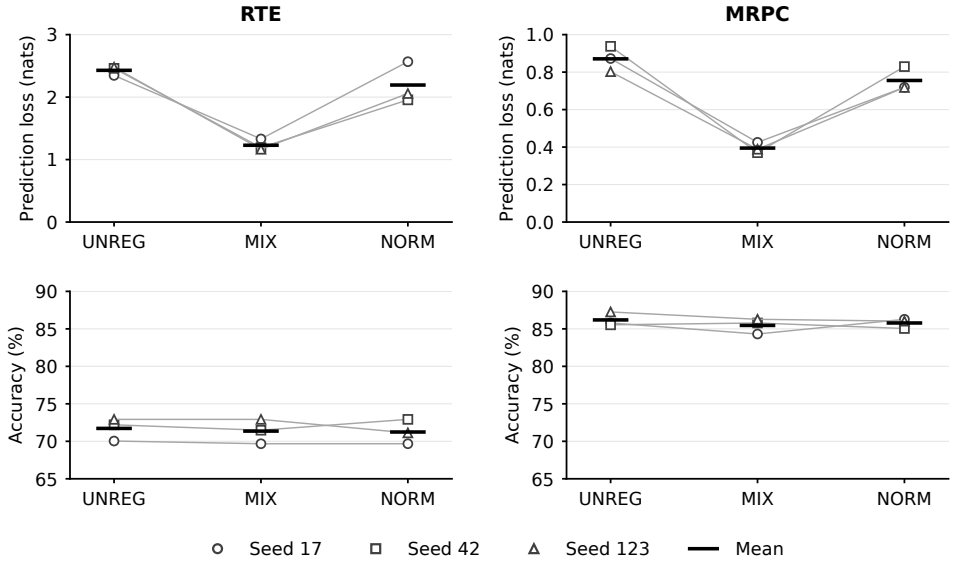


Figure 5: Held-out native prediction loss and accuracy at the fixed endpoints. Symbols identify seeds 17, 42 and 123; gray lines connect the same seed across conditions, and black bars mark means. MIX has lower native loss than both alternatives in every pair; its MIX–NORM loss advantage is unresolved after selection-fitted temperature correction (Appendix C.10). Table 68 gives the means and sample standard deviations.

On the fixed masked-token probe, the original backbone loss is 2.0956 nats. All adapted arms have higher mean probe loss, with MIX highest on both RTE and MRPC. Lower native task loss therefore does not establish retention of this separate pretraining-style behavior. The probe protocol is in Appendix B.2.

Table 67: RTE adaptation in leading, middle and tail directions. Accuracy and cross-entropy (lower is better) are fixed-endpoint inner-selection results, mean±sample SD over three paired seeds. Every condition trains the same 592,130-parameter classifier; adapter counts exclude it.

Trainable backbone update	Accuracy (%)	Task loss (nat)	Adapter params
Classifier only	59.77 ± 0.65	0.654 ± 0.002	0
Leading coefficients	71.75 ± 1.71	1.785 ± 0.083	12,288
Leading coefficients + rotations	72.22 ± 1.01	2.075 ± 0.118	405,504
Middle coefficients	71.02 ± 1.21	1.759 ± 0.027	12,288
Middle coefficients + rotations	72.22 ± 1.21	1.958 ± 0.044	405,504
Tail coefficients	70.01 ± 0.23	2.001 ± 0.141	12,288
Tail coefficients + rotations	71.02 ± 3.30	2.111 ± 0.268	405,504

Table 68: Interaction control, size control and unregularized training at the same fixed endpoints. All conditions use the scaled leading-plus-tail adapter without rotations. Entries are mean±sample SD over three seeds. Accuracy and prediction loss use held-out task examples; Cross is the pooled percentage of update energy connecting leading and tail directions, and ρ_F is relative update size. Table 12 gives per-seed size differences after calibration.

Task	Condition	Accuracy (%)	Task loss (nat)	Cross (%)	ρ_F
RTE	UNREG	71.72 ± 1.50	2.429 ± 0.071	43.7 ± 0.0	0.2751 ± 0.0041
RTE	MIX	71.36 ± 1.63	1.229 ± 0.089	10.0 ± 0.5	0.0736 ± 0.0046
RTE	NORM	71.24 ± 1.63	2.193 ± 0.327	46.2 ± 0.0	0.0759 ± 0.0016
MRPC	UNREG	86.19 ± 0.93	0.871 ± 0.068	42.4 ± 0.1	0.1802 ± 0.0009
MRPC	MIX	85.46 ± 1.02	0.395 ± 0.028	4.5 ± 0.3	0.0407 ± 0.0017
MRPC	NORM	85.78 ± 0.65	0.755 ± 0.064	46.2 ± 0.1	0.0423 ± 0.0005

C Follow-up analyses and sensitivity checks

These analyses characterize saved updates and examine uncertainty and calibration. The Qwen1.5B CommonsenseQA seed extension is reported here; Llama and Qwen3B extensions accompany their original cohorts in Appendix B. Each subsection states its cohort and evaluation split. Post hoc findings do not replace the original outcomes, and saved-update interventions do not identify a training mechanism. Appendices C.5, C.6 and C.7 are descriptive sensitivity checks, and Appendix C.9 lists the diagnostics needed to reuse these measurements.

C.1 Initialization and residual leading structure

The fixed choice $S_L = I_k$ gives equal inner coefficients to the leading directions, independent of their pretrained singular values. Choosing $S_L = \Sigma_L$ would instead weight this contribution by the pretrained spectrum. All practical conditions use the identity choice; the experiment does not compare these two designs.

For the prescribed identity rotations, $E = D = s_0 I$, $H = h_0 I_r$, and $S_L = I_k$, insertion adds

$$\Delta_{\text{init}} = s_0^2(U_L V_L^\top + h_0 U_T V_T^\top), \quad \|\Delta_{\text{init}}\|_F = s_0^2\sqrt{k + r h_0^2}, \quad p_{LL,\text{init}} = \frac{k}{k + r h_0^2}. \tag{43}$$

The cross blocks are initially zero. With $k = 512, r = 256$, the prescribed norm is approximately 0.22684 for MRPC/STS-B/SST-2, 0.0022684 for CoLA, and 0.0022628 for RTE. Leading energy is approximately 99.50% in the first four tasks and 99.995% in RTE. These are analytical values from Table 87, not recovered checkpoint measurements. These recipe calculations alone do not quantify the offset relative to each archived trained update. The fresh confirmation export separately supplies measured initial and learned deltas.

For an attained leading block, define its identity alignment

$$\eta_I = \frac{(\operatorname{tr} C_{LL})^2}{k \, \|C_{LL}\|_F^2}, \qquad 0 \le \eta_I \le 1 \quad (C_{LL} \neq 0). \tag{44}$$

It equals the fraction of leading-block Frobenius energy in the projection onto $\operatorname{span}(I_k)$. Under constant scalers and $S_L = I_k$, $C_{LL} = abI_k$ has $\eta_I = 1$ when $ab \neq 0$, even though cross mixing vanishes. This motivates reporting $C - C_{\text{init}}$, $C_{LL} - (\operatorname{tr} C_{LL}/k)I_k$, and centered scaler norms alongside total block energies. Subtracting the initialization and subtracting a fitted identity component answer different questions; neither can be reconstructed by subtracting squared norms. The fresh confirmation export supplies identity alignment and residual energies, analyzed in Appendix C.4; centered scaler vectors remain unavailable. The older archive lacks these trained measurements.

C.2 Learned placement after subtracting insertion

We evaluate $\Delta_{\text{learned}} = W_{\text{eff}}(T) - W_{\text{eff}}(0)$ directly, before squaring or normalizing. Subtracting initial energy from final energy would omit the cross term and is not this calculation. Every row is tied to the same fixed checkpoint and validation hash as the total-update report. The exported absolute energies of all 864 module records (18 runs, 48 modules each) independently reproduce pooled fractions, equal-module fractions and pooled relative norms to within 10^{-12}.

Table 69 distinguishes total and learned allocation; Table 70 lists every seed. After subtracting insertion, MIX's pooled leading share changes from 69.8% to 65.5% on MRPC and stays at 45.2% on RTE. Equal-module learned leading shares are 67.3% on RTE and 64.7% on MRPC. The task contrast therefore depends on whether modules are weighted by energy or equally. Initialization matters most for small final updates: the largest per-seed change in an equal-module block fraction is 6.6 points on RTE and 13.4 on MRPC. Learned displacement includes any cancellation of the inserted contribution.

Table 69: Total and learned-since-insertion energy percentages at fixed endpoints, mean±sample SD over three seeds. Cross combines LT and TL. Both normalizations use the same module energies; pooled weighting emphasizes modules with larger updates.

			Pooled energy			Equal module		
Task	Arm	Delta	LL	Cross	TT	LL	Cross	TT
RTE	UNREG	Total	45.9 ± 0.1	43.7 ± 0.0	10.4 ± 0.0	46.2 ± 0.1	43.5 ± 0.0	10.3 ± 0.1
RTE	UNREG	Learned	45.9 ± 0.1	43.7 ± 0.0	10.4 ± 0.0	46.2 ± 0.1	43.5 ± 0.0	10.3 ± 0.1
RTE	MIX	Total	45.2 ± 0.9	10.0 ± 0.5	44.8 ± 1.4	61.5 ± 2.1	14.9 ± 0.9	23.6 ± 1.8
RTE	MIX	Learned	45.2 ± 0.9	10.0 ± 0.5	44.8 ± 1.4	67.3 ± 3.0	10.0 ± 1.4	22.6 ± 2.0
RTE	NORM	Total	40.7 ± 0.0	46.2 ± 0.0	13.1 ± 0.0	40.5 ± 0.2	46.3 ± 0.1	13.2 ± 0.1
RTE	NORM	Learned	40.7 ± 0.0	46.2 ± 0.0	13.1 ± 0.0	40.5 ± 0.2	46.3 ± 0.1	13.2 ± 0.1
MRPC	UNREG	Total	48.0 ± 0.1	42.4 ± 0.1	9.6 ± 0.1	48.2 ± 0.0	42.2 ± 0.0	9.6 ± 0.0
MRPC	UNREG	Learned	47.9 ± 0.1	42.5 ± 0.0	9.7 ± 0.1	48.1 ± 0.0	42.4 ± 0.0	9.6 ± 0.0
MRPC	MIX	Total	69.8 ± 1.1	4.5 ± 0.3	25.8 ± 1.4	76.2 ± 0.5	5.5 ± 0.1	18.3 ± 0.6
MRPC	MIX	Learned	65.5 ± 1.1	5.1 ± 0.4	29.4 ± 1.5	64.7 ± 2.7	7.7 ± 0.8	27.6 ± 2.7
MRPC	NORM	Total	40.6 ± 0.2	46.2 ± 0.1	13.2 ± 0.1	40.5 ± 0.3	46.2 ± 0.1	13.3 ± 0.2
MRPC	NORM	Learned	41.3 ± 0.2	45.7 ± 0.1	13.0 ± 0.1	42.0 ± 0.2	45.1 ± 0.0	13.0 ± 0.2

C.3 Module norms under pooled matching

For module ℓ, let $r_{a,\ell} = \|\Delta_{a,\ell}\|_F / \|W_{\text{pre},\ell}\|_F$. Table 71 compares the same 48 modules within each seed. Its rescaled check compares $(r_{\text{NORM},\ell}/\rho_{\text{NORM}})/(r_{\text{MIX},\ell}/\rho_{\text{MIX}})$ with one, removing the residual global norm ratio descriptively without changing or retraining either arm. At most two modules per pair satisfy a 5% tolerance after this rescaling. Figure 6 shows where norms differ: RTE MIX concentrates its largest relative norms in output projections, notably layers 7 and 11, while many other modules have very small updates. The 10^{-6} threshold is a descriptive reporting threshold, not a training gate. Ratios against these near-zero updates can be arbitrarily large, so we show module norms directly. Absolute squared norms independently quantify energy concentration (Figure 7). For MRPC, the three largest modules carry 35.7–50.1% of MIX total energy versus 17.5–19.5% for NORM. These descriptive order statistics are accompanied by the complete cumulative curves; module counts are not additional experimental replicates.

Table 70: Every confirmation seed's learned-since-insertion block percentages. Fixed endpoints and validation identities agree with Table 14. All four blocks are retained; rounding may affect displayed sums.

			Pooled energy				Equal module			
Task	Arm	Seed	LL	LT	TL	TT	LL	LT	TL	TT
RTE	UNREG	17	45.92	22.48	21.17	10.43	46.09	22.36	21.20	10.35
RTE	UNREG	42	46.02	22.25	21.37	10.36	46.26	22.17	21.32	10.25
RTE	UNREG	123	45.87	22.41	21.29	10.43	46.12	22.29	21.27	10.32
RTE	MIX	17	44.41	4.53	4.87	46.19	69.14	4.71	3.87	22.27
RTE	MIX	42	45.09	4.69	5.40	44.82	63.85	6.40	4.91	24.83
RTE	MIX	123	46.17	5.29	5.08	43.47	68.96	5.84	4.37	20.83
RTE	NORM	17	40.63	23.24	22.95	13.17	40.35	23.25	23.11	13.29
RTE	NORM	42	40.67	23.12	23.09	13.13	40.39	23.13	23.15	13.32
RTE	NORM	123	40.72	23.33	22.82	13.13	40.76	23.28	22.83	13.14
MRPC	UNREG	17	47.77	21.85	20.67	9.71	48.02	21.72	20.67	9.59
MRPC	UNREG	42	47.97	21.86	20.57	9.60	48.09	21.80	20.57	9.54
MRPC	UNREG	123	47.87	21.70	20.74	9.69	48.07	21.58	20.75	9.60
MRPC	MIX	17	66.12	2.66	2.74	28.47	63.19	3.99	4.60	28.21
MRPC	MIX	42	66.23	2.51	2.72	28.54	67.79	3.27	4.23	24.72
MRPC	MIX	123	64.22	2.23	2.43	31.12	63.01	3.06	3.92	30.01
MRPC	NORM	17	41.08	22.82	22.97	13.13	41.80	22.68	22.35	13.17
MRPC	NORM	42	41.35	22.76	22.89	12.99	42.01	22.44	22.64	12.91
MRPC	NORM	123	41.48	22.55	23.06	12.91	42.12	22.34	22.73	12.81

Table 71: Signed pooled error $\delta_s = \rho_{\mathrm{NORM},s}/\rho_{\mathrm{MIX},s} - 1$ and module-relative norms. The two within-5% columns count matched module norms before and after global rescaling. The last column counts module norms below 10^{-6} for MIX/NORM. Modules describe heterogeneity within each seed.

Task	Seed	δ_s (%)	Median MIX	Median NORM	Within 5%	Rescaled	Small MIX/NORM
RTE	17	$+8.0$	0.0228	0.0788	1/48	0/48	14/0
RTE	42	-5.2	0.0209	0.0764	0/48	0/48	15/0
RTE	123	$+7.5$	0.0126	0.0746	0/48	0/48	12/0
MRPC	17	$+5.5$	0.0234	0.0430	4/48	2/48	0/0
MRPC	42	$+6.7$	0.0277	0.0426	1/48	1/48	1/0
MRPC	123	-0.1	0.0273	0.0420	2/48	2/48	1/0

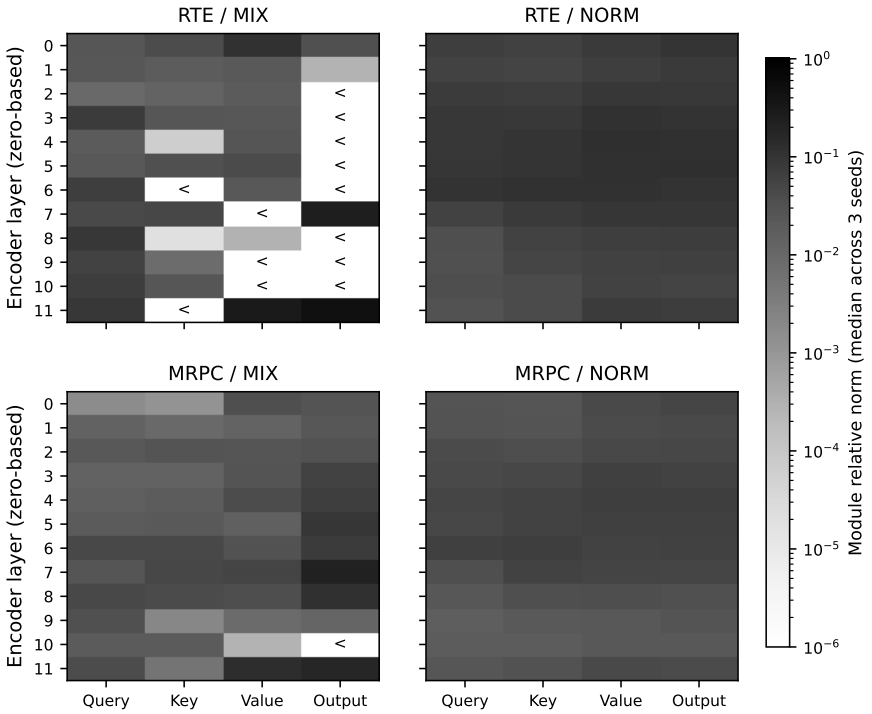


Figure 6: Module-relative update norms, median across three confirmation seeds for each module. All panels use the same logarithmic scale; "$<$" marks medians below the displayed 10^{-6} floor. This heatmap displays medians only; it does not show per-module seed dispersion. No zero-update energy fractions are inferred from these norms.

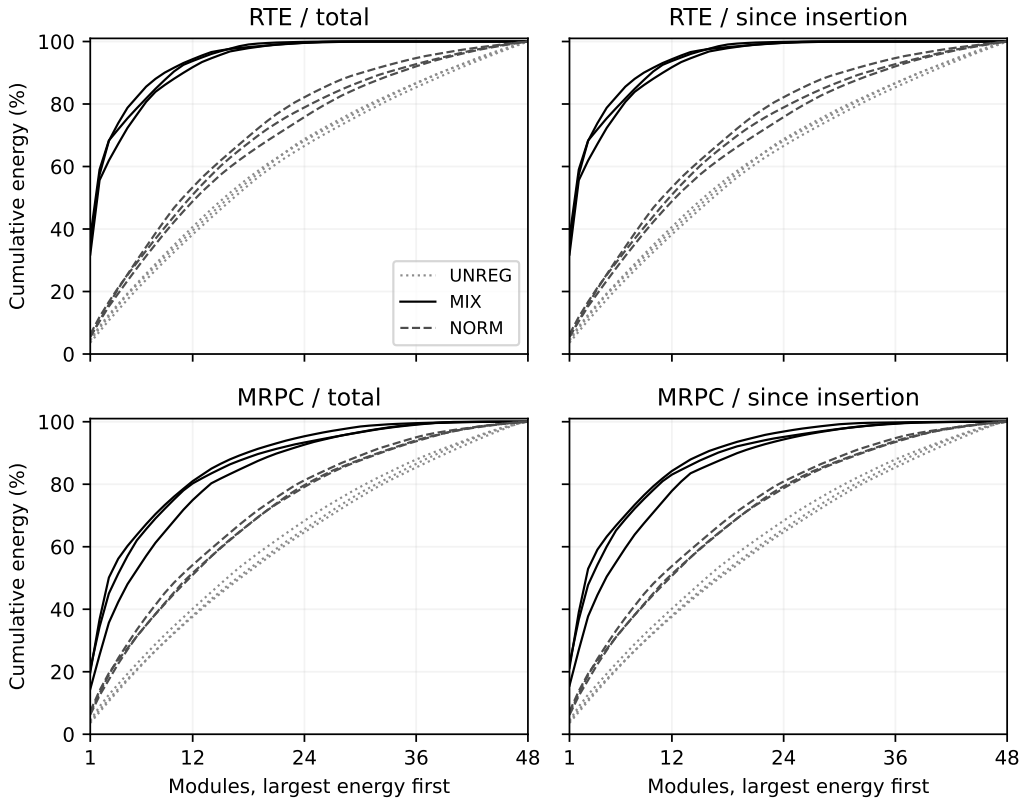


Figure 7: Cumulative absolute squared update norm across all 48 modules, sorted separately within each run. Each curve is one confirmation seed. The left panels include insertion; the right panels subtract it before computing energies. Steeper early accumulation indicates more concentrated adaptation. Sorting describes concentration, not correspondence of module identities across arms.

C.4 IDENTITY ALIGNMENT OF THE LEARNED LEADING BLOCK

For each module, project C_{LL} onto its own identity matrix as in (44). Table 72 reports the fraction of summed leading energy explained by these module-wise projections, $\eta_{I,\mathrm{pool}} = 1 - \sum_\ell \|C_{LL,\ell} - (\operatorname{tr} C_{LL,\ell}/k) I_k\|_F^2 / \sum_\ell \|C_{LL,\ell}\|_F^2$. This permits a different fitted scalar in each module; it does not fit one scalar to the whole network. We also report the equal-module mean of η_I and the non-identity leading residual divided by total delta energy. All quantities use exported checkpoint diagnostics, not a fit to the rounded block fractions.

MIX's learned leading energy is mostly aligned with the identity component. Its non-identity leading residual is about 6.9% of total learned energy on both tasks. Learned cross shares are 10.0% and 5.1%, so the endpoint retains departures from the constant-scaler limit, which has exactly zero cross energy. Identity alignment describes the leading block; centered-scaler measurements and random-projector controls would distinguish the factor mechanisms that produce it.

Table 72: Identity alignment and residual leading structure (percentages, mean±sample SD, three seeds). The first two columns weight modules by leading-block energy; the third weights modules equally. The final column uses all learned delta energy as denominator.

Task	Arm	Total $\eta_{I,\mathrm{pool}}$	Learned $\eta_{I,\mathrm{pool}}$	Learned $\eta_{I,\mathrm{eqm}}$	Non-identity LL / learned total
RTE	UNREG	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	45.94 ± 0.08
RTE	MIX	84.67 ± 1.04	84.67 ± 1.05	79.67 ± 2.01	6.94 ± 0.60
RTE	NORM	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	40.67 ± 0.04
MRPC	UNREG	0.81 ± 0.07	0.29 ± 0.04	0.29 ± 0.05	47.73 ± 0.10
MRPC	MIX	91.38 ± 0.45	89.50 ± 0.61	79.23 ± 3.45	6.89 ± 0.51
MRPC	NORM	0.05 ± 0.01	2.76 ± 0.13	5.94 ± 0.37	40.16 ± 0.21

C.5 SENSITIVITY TO MODULES WITH SMALL UPDATES

This descriptive sensitivity analysis was specified after inspecting the original outcomes. For each task and seed, retain the same named modules in MIX and NORM only if both total-update relative norms are at least the stated floor. Apply that identical subset to total and learned diagnostics, then average module cross fractions. The two floors, 10^{-6} and 10^{-3}, are reporting choices, not training gates or preregistered hypotheses. Counts and all seed differences are retained in the analysis data; the seed remains the unit of replication.

Table 73 shows that the gap survives both restrictions. We recompute each retained module's fraction directly, including small nonzero updates. The restriction conditions on attained norms, a post-treatment variable. It is a descriptive check of averaging sensitivity; matching module energy during training would be a different intervention.

Table 73: Equal-module cross shares and paired NORM–MIX differences (percentage points), mean±sample SD over three seeds. Counts are retained modules in seed order 17, 42, 123; each seed starts with 48.

Task	Norm floor	Delta	Counts	MIX	NORM	NORM–MIX
RTE	$1e-06$	Total	34,33,36	14.68 ± 1.80	46.72 ± 0.13	32.04 ± 1.71
RTE	$1e-06$	Learned	34,33,36	14.06 ± 2.21	46.72 ± 0.13	32.66 ± 2.10
RTE	0.001	Total	30,32,30	15.19 ± 1.84	46.76 ± 0.18	31.57 ± 1.83
RTE	0.001	Learned	30,32,30	15.19 ± 1.86	46.76 ± 0.18	31.57 ± 1.84
MRPC	$1e-06$	Total	48,47,47	5.46 ± 0.25	46.24 ± 0.04	40.78 ± 0.21
MRPC	$1e-06$	Learned	48,47,47	7.79 ± 0.74	45.14 ± 0.09	37.35 ± 0.83
MRPC	0.001	Total	47,44,46	5.63 ± 0.38	46.26 ± 0.03	40.63 ± 0.38
MRPC	0.001	Learned	47,44,46	8.08 ± 0.75	45.23 ± 0.09	37.15 ± 0.84

C.6 HELD-ASIDE LOSS AND CONFIDENCE CALIBRATION

Task negative log likelihood (NLL) is the cross-entropy column in Table 16. UNREG/MIX/NORM mean NLL is $2.429/1.229/2.193$ nats on RTE and $0.871/0.395/0.755$ on MRPC. MIX has lower loss in every paired seed, and its three-seed range is below both comparator ranges within each task. Table 74 reports all contrasts. We identified this ordering after inspecting the completed evaluation.

The subsequent export includes the required inner-selection logits. We fit a positive temperature there, freeze it, and apply it to the held-out logits (Guo et al., 2017). Appendix C.10 reports native and corrected losses. Temperature preserves the predicted class; the correction therefore tests probability scale without changing accuracy. The raw values above remain the original outcomes of the complete regularization comparison.

Table 74: Paired held-aside task NLL differences in nats, left minus right. Sample SD uses the $n-1$ divisor, $n=3$. Intervals follow Appendix B.5; negative values favor the left arm on this loss only.

Task	Contrast	Seed 17	Seed 42	Seed 123	Mean±SD	Nominal 95% CI
RTE	MIX–NORM	-1.236	-0.764	-0.892	-0.964 ± 0.244	$[-1.571, -0.357]$
RTE	MIX–UNREG	-1.017	-1.269	-1.313	-1.199 ± 0.160	$[-1.596, -0.803]$
RTE	NORM–UNREG	+0.220	-0.505	-0.421	-0.235 ± 0.396	$[-1.219, 0.749]$
MRPC	MIX–NORM	-0.293	-0.459	-0.330	-0.361 ± 0.087	$[-0.578, -0.144]$
MRPC	MIX–UNREG	-0.447	-0.568	-0.414	-0.476 ± 0.081	$[-0.677, -0.276]$
MRPC	NORM–UNREG	-0.154	-0.108	-0.084	-0.115 ± 0.035	$[-0.203, -0.028]$

C.7 DESCRIPTIVE SENSITIVITY ALONG THE MAGNITUDE FRONTIER

For each task, we fit $p^{(\%)}_{\text{cross}} = a + b \log \rho_F$ by ordinary least squares to all seven NORM calibration doses (seed 31415). Table 75 reports the slope, maximum fitted residual and illustrative change $|b| \log(1.08)$. The latter is about 0.22 points, compared with the observed 36–42 point confirmation contrast. At signed norm error δ_s, the corresponding fitted change is $b \log(1+\delta_s)$; all six values are in the analysis export. This is an interpolation of one observed frontier per task. It describes sensitivity within the measured range rather than bounding an exactly matched or unobserved experiment.

Table 75: Single-calibration-seed NORM frontier. Cross shares and fitted changes are percentage points. The norm range is computed within each task, not across tasks. The illustrative shift is a fitted sensitivity, not a causal bound.

Task	Norm max/min	Cross range	b (per log norm)	8% shift	Max residual
RTE	12.10	43.66–50.91	-2.905	0.224	0.878
MRPC	8.47	42.41–48.60	-2.900	0.223	0.301

C.8 POWER SENSITIVITY UNDER EXPLICIT VARIANCE ASSUMPTIONS

The nominal intervals describe uncertainty; their half-widths are not minimum detectable effects. To make the small-seed limitation quantitative, consider a hypothetical two-sided paired t test with $n=3$, independent Gaussian seed differences, significance level 0.05, and fixed true difference SD σ. Its power at a mean difference δ is

$$\Pr\left\{ |T_{2,\sqrt{3}\delta/\sigma}| > t_{0.975,2} \right\}. \tag{45}$$

Here $T_{\nu,\lambda}$ is noncentral Student t. Solving for 80% power gives $|\delta| \simeq 3.26\sigma$. Table 76 substitutes each observed paired sample SD as an uncertain variance scenario, including all task-score, task-NLL and probe contrasts. These are post hoc design sensitivities, not achieved power, effects excluded by the experiment, or an equivalence margin. At three seeds the variance itself is poorly estimated; all entries scale linearly with the assumed σ and condition on the fixed evaluation examples.

Table 76: Effect magnitude for 80% power in the stated hypothetical Gaussian paired test, using the observed SD as a variance scenario. This is a planning calculation, not a new test of the reported contrasts. pp: percentage points.

Task	Contrast	Metric	Paired SD	80%-power effect
RTE	MIX–NORM	Accuracy (pp)	1.628	5.313
RTE	MIX–NORM	Task NLL (nat)	0.244	0.798
RTE	MIX–NORM	Probe CE (nat)	0.262	0.854
RTE	MIX–UNREG	Accuracy (pp)	0.361	1.178
RTE	MIX–UNREG	Task NLL (nat)	0.160	0.521
RTE	MIX–UNREG	Probe CE (nat)	0.313	1.021
RTE	NORM–UNREG	Accuracy (pp)	1.268	4.138
RTE	NORM–UNREG	Task NLL (nat)	0.396	1.293
RTE	NORM–UNREG	Probe CE (nat)	0.099	0.324
MRPC	MIX–NORM	Accuracy (pp)	1.436	4.688
MRPC	MIX–NORM	F1 (pp)	0.976	3.185
MRPC	MIX–NORM	Task NLL (nat)	0.087	0.285
MRPC	MIX–NORM	Probe CE (nat)	0.249	0.812
MRPC	MIX–UNREG	Accuracy (pp)	0.884	2.884
MRPC	MIX–UNREG	F1 (pp)	0.581	1.895
MRPC	MIX–UNREG	Task NLL (nat)	0.081	0.264
MRPC	MIX–UNREG	Probe CE (nat)	0.234	0.764
MRPC	NORM–UNREG	Accuracy (pp)	0.861	2.810
MRPC	NORM–UNREG	F1 (pp)	0.685	2.237
MRPC	NORM–UNREG	Task NLL (nat)	0.035	0.115
MRPC	NORM–UNREG	Probe CE (nat)	0.113	0.369

C.9 A REUSABLE SPECTRAL REPORT

Use a fixed pretrained checkpoint, adapted module list, and declared cutoff for all methods. The primary cutoff here follows the adapter tail dimension, not a measured spectral gap. For each saved effective matrix, record both $\Delta_{\text{total}} = W_{\text{eff}} - W_0$ and $\Delta_{\text{learned}} = W_{\text{eff}} - W_{\text{eff}}(0)$; report zero-update fractions as undefined. The report comprises absolute Frobenius block energies, their fractions, ρ_F, operator norm, stable rank, and the dimension-only fractions. Add leading identity alignment η_I and absolute non-identity residual energy under both delta definitions. Export each module's norm and absolute block energies, the cumulative energy curve, and threshold-count curves; a cumulative energy curve alone cannot recover relative-norm thresholds. We report 10^{-6} for small-norm counts and both $10^{-6}, 10^{-3}$ for sensitivity subsets. Future comparisons must declare thresholds before inspection and retain the underlying values. Use the aggregation specified by each table: the focused headline pools squared energies, while equal-module summaries normalize before averaging. This report is diagnostic, not a preservation pass/fail criterion.

Let P, P' be rank-k left projectors before and after adaptation, and $m = \min(k, d-k)$. Supplement the largest-angle sine with

$$d_{\text{ch}} = \frac{\|P - P'\|_F}{\sqrt{2m}}, \qquad o = \frac{\text{tr}(PP')}{k}, \qquad d_{\text{ch}}^2 = \frac{k}{m}(1-o). \tag{46}$$

Here d_{ch}^2 averages squared sines over the m potentially nonzero angles, including any zeros within them; o is normalized overlap, with random-independent-subspace expectation k/d. Apply the same definitions on the right. Report the boundary values σ_k, σ_{k+1}, their gap, and a prespecified cutoff sweep, such as $k \in \{16, 64, 128, 256, 512\}$. Singular-subspace perturbation bounds require separation conditions (Wedin, 1972); a norm/gap quotient is not an unconditional prediction of drift. A boundary tie is flagged, not presented as uniquely defined ranked stability.

Pair these measurements with validation and untouched-test performance, and one frozen-pretraining-head behavior probe on the same checkpoints when retention is studied. Store per-seed values at both a fixed training step and the selected checkpoint; log diagnostics separately from optimization costs. Full ranked-subspace diagnostics require dense SVDs, $O(d^3)$ for square matrices; a truncated

or randomized approximation must report its rank, residual tolerance and convergence checks, particularly near cutoff ties. The archived study supplies total-delta allocation and maximum-angle drift, but not the full initialization-corrected or chordal report. Those fields require checkpoint evaluation, not imputation from the current ratios.

C.10 Selection-fitted temperature: encoder follow-up

We use saved inner-selection logits to fit one positive temperature per endpoint, then evaluate the unchanged temperature on the held-out logits. The training-time interaction experiment uses 277 RTE and 408 MRPC examples; the band/head experiment uses the same 277-example RTE validation set. All source predictions and native losses reproduce before correction. Table 77 reports three-seed means. The two experiments have separate training recipes and their outcomes are not pooled.

For the training-time interaction experiment, corrected MIX–NORM loss differences are -0.0023 nats on RTE, with nominal paired interval $[-0.0543, 0.0496]$, and -0.0092 on MRPC, with interval $[-0.0736, 0.0552]$. The native advantage is largely removable by this scalar correction; the remaining contrasts are unresolved. This does not decompose the training effect into independent calibration and representation components.

For the band experiment, mean fitted temperatures range from 7.91 to 9.48, versus 0.905 for head-only. All six native band-minus-head loss differences are positive; all six corrected differences are negative, ranging from -0.142 to -0.098 nats, with nominal seed intervals excluding zero. None of the nine location/rotation loss contrasts excludes zero before or after correction. Accuracy is unchanged. These are three-seed comparisons on a fixed evaluation set, not equivalence tests or multiplicity-adjusted conclusions.

Table 77: Encoder held-out outcomes before and after selection-fitted temperature scaling. Entries are means over seeds 17, 42 and 123. Accuracy is unchanged by the positive scalar; both loss columns are in nats.

Study	Condition	Accuracy (%)	Native NLL	Corrected NLL
RTE interaction	UNREG	71.72	2.429	0.565
RTE interaction	MIX	71.36	1.229	0.551
RTE interaction	NORM	71.24	2.193	0.553
MRPC interaction	UNREG	86.19	0.871	0.335
MRPC interaction	MIX	85.46	0.395	0.325
MRPC interaction	NORM	85.78	0.755	0.334
RTE band	Head only	58.12	0.680	0.682
RTE band	Leading coefficients	72.68	1.783	0.555
RTE band	Leading + rotations	74.61	1.932	0.540
RTE band	Middle coefficients	69.92	1.726	0.564
RTE band	Middle + rotations	71.72	2.047	0.550
RTE band	Tail coefficients	69.19	2.007	0.570
RTE band	Tail + rotations	68.35	2.250	0.584

C.11 Qwen1.5B CommonsenseQA seed extension and RTE contrasts

The original Qwen2.5-1.5B CommonsenseQA spectral-band cohort uses seeds 17, 42 and 123. After observing its outcomes, we registered seeds 2027 and 2028 for all six spectral arms. We report the original three, added two and combined five separately; the post hoc extension does not replace the original results. The separate Llama and Qwen3B extensions use seeds 2026 and 2027 (Appendices B.13 and B.15).

L, M and T denote leading, middle and tail bands; C and R denote coefficients and added rotations. A contrast subtracts the second condition from the first. Accuracy differences are percentage points; loss differences are nats in the task's own prediction units. Seed intervals use paired differences and a Student-t critical value with $n - 1$ degrees of freedom. They are nominal, exploratory and unadjusted for multiple comparisons. Shared arms and test examples do not form independent

replications. Tables 78 and 79 give the CommonsenseQA band contrasts; Table 82 gives the RTE held-out contrasts.

Table 78: CommonsenseQA: paired differences in accuracy (percentage points). Each cell is the mean difference and nominal 95% seed interval. The original cohort remains primary.

Contrast	Original 3 seeds	Post hoc added 2	Post hoc combined 5
LC−MC	-0.19 $[-1.37, +0.98]$	-0.08 $[-1.12, +0.96]$	-0.15 $[-0.58, +0.28]$
LC−TC	-0.52 $[-1.90, +0.87]$	-0.82 $[-1.86, +0.22]$	-0.64 $[-1.17, -0.10]$
MC−TC	-0.33 $[-0.68, +0.02]$	-0.74 $[-2.82, +1.34]$	-0.49 $[-0.83, -0.15]$
LR−MR	$+0.03$ $[-1.44, +1.50]$	-0.08 $[-3.20, +3.04]$	-0.02 $[-0.58, +0.55]$
LR−TR	-0.44 $[-0.95, +0.08]$	-0.37 $[-5.05, +4.31]$	-0.41 $[-0.78, -0.04]$
MR−TR	-0.46 $[-1.51, +0.58]$	-0.29 $[-1.85, +1.27]$	-0.39 $[-0.80, +0.01]$
LR−LC	$+1.20$ $[-0.27, +2.67]$	$+1.15$ $[+0.11, +2.19]$	$+1.18$ $[+0.65, +1.71]$
MR−MC	$+0.98$ $[-0.29, +2.25]$	$+1.15$ $[+0.11, +2.19]$	$+1.05$ $[+0.58, +1.52]$
TR−TC	$+1.12$ $[+0.81, +1.43]$	$+0.70$ $[-3.99, +5.38]$	$+0.95$ $[+0.50, +1.40]$

Table 79: CommonsenseQA: paired differences in native NLL (nats). Each cell is the mean difference and nominal 95% seed interval. The original cohort remains primary.

Contrast	Original 3 seeds	Post hoc added 2	Post hoc combined 5
LC−MC	$+0.00285$ $[-0.00156, +0.00726]$	$+0.00309$ $[-0.00036, +0.00653]$	$+0.00294$ $[+0.00136, +0.00453]$
LC−TC	$+0.01583$ $[+0.01400, +0.01766]$	$+0.01531$ $[+0.00706, +0.02357]$	$+0.01562$ $[+0.01469, +0.01655]$
MC−TC	$+0.01298$ $[+0.00995, +0.01601]$	$+0.01222$ $[+0.00741, +0.01703]$	$+0.01268$ $[+0.01144, +0.01391]$
LR−MR	$+0.00256$ $[+0.00172, +0.00340]$	$+0.00174$ $[-0.00023, +0.00370]$	$+0.00223$ $[+0.00158, +0.00288]$
LR−TR	$+0.01296$ $[+0.01013, +0.01579]$	$+0.01171$ $[+0.01023, +0.01319]$	$+0.01246$ $[+0.01114, +0.01378]$
MR−TR	$+0.01040$ $[+0.00770, +0.01310]$	$+0.00997$ $[+0.00949, +0.01046]$	$+0.01023$ $[+0.00923, +0.01123]$
LR−LC	-0.03936 $[-0.04385, -0.03487]$	-0.04146 $[-0.05926, -0.02365]$	-0.04020 $[-0.04266, -0.03774]$
MR−MC	-0.03908 $[-0.04237, -0.03579]$	-0.04010 $[-0.05643, -0.02378]$	-0.03949 $[-0.04125, -0.03772]$
TR−TC	-0.03649 $[-0.03893, -0.03406]$	-0.03785 $[-0.04888, -0.02683]$	-0.03704 $[-0.03851, -0.03556]$

We also report evaluation-example uncertainty at the observed seeds. For each contrast, 10,000 paired bootstrap draws (random seed 20260920) resample the same example indices for both arms and every seed. Choice NLL averages examples. The percentile intervals in Tables 80 and 81 vary examples, not training seeds, and are not a substitute for seed replication or a joint uncertainty interval. For

Table 80: Original three-seed cohorts: paired example-bootstrap differences and 95% percentile intervals.

Task	Contrast	Accuracy (points)	Native NLL (nats)
CSQA	LC−MC	-0.19 $[-0.71, +0.33]$	$+0.00285$ $[+0.00045, +0.00523]$
CSQA	LC−TC	-0.52 $[-1.06, +0.03]$	$+0.01583$ $[+0.01247, +0.01931]$
CSQA	MC−TC	-0.33 $[-0.85, +0.22]$	$+0.01298$ $[+0.00994, +0.01614]$
CSQA	LR−MR	$+0.03$ $[-0.38, +0.41]$	$+0.00256$ $[+0.00040, +0.00477]$
CSQA	LR−TR	-0.44 $[-0.93, +0.05]$	$+0.01296$ $[+0.00951, +0.01637]$
CSQA	MR−TR	-0.46 $[-0.90, -0.03]$	$+0.01040$ $[+0.00708, +0.01375]$
CSQA	LR−LC	$+1.20$ $[+0.57, +1.83]$	-0.03936 $[-0.04461, -0.03412]$
CSQA	MR−MC	$+0.98$ $[+0.38, +1.61]$	-0.03908 $[-0.04406, -0.03412]$
CSQA	TR−TC	$+1.12$ $[+0.55, +1.69]$	-0.03649 $[-0.04134, -0.03168]$

RTE leading-minus-middle rotation accuracy, all three observed differences are identical. Its seed interval consequently has zero width; this does not remove uncertainty from the finite evaluation set. It is not evidence that the effect is known exactly.

Table 81: Combined five-seed cohorts: paired example-bootstrap differences and 95% percentile intervals. This view includes the post hoc extension.

Task	Contrast	Accuracy (points)	Native NLL (nats)
CSQA	LC−MC	-0.15 $[-0.54, +0.23]$	$+0.00294$ $[+0.00107, +0.00484]$
CSQA	LC−TC	-0.64 $[-1.06, -0.21]$	$+0.01562$ $[+0.01306, +0.01827]$
CSQA	MC−TC	-0.49 $[-0.90, -0.08]$	$+0.01268$ $[+0.01029, +0.01503]$
CSQA	LR−MR	-0.02 $[-0.33, +0.28]$	$+0.00223$ $[+0.00057, +0.00392]$
CSQA	LR−TR	-0.41 $[-0.79, -0.02]$	$+0.01246$ $[+0.00977, +0.01513]$
CSQA	MR−TR	-0.39 $[-0.75, -0.03]$	$+0.01023$ $[+0.00774, +0.01278]$
CSQA	LR−LC	$+1.18$ $[+0.69, +1.67]$	-0.04020 $[-0.04415, -0.03610]$
CSQA	MR−MC	$+1.05$ $[+0.57, +1.52]$	-0.03949 $[-0.04338, -0.03567]$
CSQA	TR−TC	$+0.95$ $[+0.49, +1.41]$	-0.03704 $[-0.04090, -0.03311]$

Table 82: RTE held-out band contrasts over three seeds. Each cell shows a paired mean difference and nominal 95% seed interval. Native and temperature-corrected NLL are kept distinct. L/M/T and D/R use the same shorthand as above; R rotates 64 directions here.

Contrast	Accuracy (points)	Native NLL (nats)	Corrected NLL (nats)
LC−MC	$+2.77$ $[-2.64, +8.17]$	$+0.057$ $[-0.080, +0.193]$	-0.009 $[-0.029, +0.012]$
LC−TC	$+3.49$ $[-3.00, +9.98]$	-0.225 $[-0.779, +0.330]$	-0.015 $[-0.052, +0.023]$
MC−TC	$+0.72$ $[-7.83, +9.28]$	-0.281 $[-0.934, +0.371]$	-0.006 $[-0.049, +0.037]$
LR−MR	$+2.89$ $[+2.89, +2.89]$	-0.115 $[-0.735, +0.505]$	-0.010 $[-0.053, +0.033]$
LR−TR	$+6.26$ $[-2.90, +15.42]$	-0.318 $[-1.050, +0.413]$	-0.044 $[-0.128, +0.041]$
MR−TR	$+3.37$ $[-5.79, +12.53]$	-0.204 $[-0.611, +0.204]$	-0.034 $[-0.116, +0.048]$
LR−LC	$+1.93$ $[-0.81, +4.67]$	$+0.149$ $[-0.435, +0.734]$	-0.015 $[-0.054, +0.025]$
MR−MC	$+1.81$ $[-0.89, +4.50]$	$+0.321$ $[-0.184, +0.826]$	-0.014 $[-0.049, +0.021]$
TR−TC	-0.84 $[-3.10, +1.41]$	$+0.243$ $[-0.170, +0.655]$	$+0.014$ $[-0.062, +0.091]$
LC−head	$+14.56$ $[+10.94, +18.19]$	$+1.103$ $[+0.948, +1.258]$	-0.127 $[-0.154, -0.099]$
LR−head	$+16.49$ $[+15.12, +17.86]$	$+1.252$ $[+0.720, +1.785]$	-0.142 $[-0.184, -0.099]$
MC−head	$+11.79$ $[+9.72, +13.86]$	$+1.046$ $[+0.820, +1.273]$	-0.118 $[-0.138, -0.098]$
MR−head	$+13.60$ $[+12.23, +14.97]$	$+1.367$ $[+1.043, +1.691]$	-0.132 $[-0.147, -0.117]$
TC−head	$+11.07$ $[+4.34, +17.80]$	$+1.328$ $[+0.902, +1.754]$	-0.112 $[-0.138, -0.087]$
TR−head	$+10.23$ $[+1.90, +18.56]$	$+1.571$ $[+0.869, +2.273]$	-0.098 $[-0.194, -0.002]$

C.12 COMMONSENSEQA CONFIDENCE CALIBRATION

Temperature correction is fitted on inner-selection examples without changing checkpoints or predicted choices. On CommonsenseQA, it lowers the frozen reference's NLL substantially; trained-state losses change much less. Table 83 reports the accepted original cohort and random-frame control.

Table 83: CommonsenseQA choice NLL before and after selection-fitted temperature scaling. Trained entries are means over the original three paired seeds; the frozen reference is scored once. Random-frame controls were added after the spectral-band study. No prediction changes under this correction.

Condition	Native NLL	Corrected NLL
Frozen reference	1.9129	0.8275
Leading coefficients	0.6651	0.6649
Leading + rotations	0.6258	0.6253
Middle coefficients	0.6623	0.6621
Middle + rotations	0.6232	0.6228
Tail coefficients	0.6493	0.6493
Tail + rotations	0.6128	0.6126
Random frame coefficients	0.6562	0.6561
Random frame + rotations	0.6176	0.6172

C.13 MODULE-MAGNITUDE TRANSFERS AND NUMERICAL PRECISION

For each paired MIX and NORM endpoint, we rescale each module's saved update to the other endpoint's Frobenius norm while retaining its direction. Both ambient scalers are multiplied by the square root of that module's positive rescaling factor; the encoder's source classifier is unchanged. These transformed states were never produced by optimization. Their results test a property of the saved endpoints, not the causal effect of allocating magnitudes that way during training.

The float64 audit reconstructs updates from their actual factors. All 576 encoder module updates have stable, nonzero directions at the declared numerical tolerance, including very small MIX updates. We therefore evaluate all 12 complete encoder transfers; a small update is not treated as deletion. Some reciprocal transfers require rescaling factors as large as 6.4×10^{14}, which limits their interpretation. Table 84 reports all transfers. Per-module cross share is invariant to positive scalar rescaling by construction; its preservation is not an independent empirical finding. Pooled shares can change as modules receive different weights.

Table 84: Endpoint module-magnitude transfers, three-seed means. Direction and Magnitudes name the source arms; equal names are native endpoints. Each endpoint receives its own selection-fitted temperature. Accuracy is percent; task losses use their respective units.

Task	Direction	Magnitudes	Accuracy	Native NLL	Corrected NLL
RTE	NORM	NORM	71.24	2.1934	0.5533
RTE	NORM	MIX	60.41	5.5889	0.6581
RTE	MIX	MIX	71.36	1.2292	0.5510
RTE	MIX	NORM	49.46	3.1319	0.6947
MRPC	NORM	NORM	85.78	0.7555	0.3338
MRPC	NORM	MIX	84.40	1.8297	0.3921
MRPC	MIX	MIX	85.46	0.3946	0.3246
MRPC	MIX	NORM	71.49	0.7803	0.5232

For the encoder, transplanting magnitudes increases corrected loss relative to each source endpoint. Matching the final module norms therefore does not reproduce the opposite endpoint's predictions. It does not determine what would happen if those magnitudes were imposed during training: the update directions, optimization histories and encoder classifier heads also differ.

C.14 Prediction disagreement between penalty arms

This post hoc, CPU-only analysis aligns saved encoder predictions by example identifier and reproduces all 39 exported endpoint accuracies (21 band/head and 18 interaction endpoints) before comparing decisions in the interaction study. For two arms evaluated on the same examples, the accuracy difference is the difference of the two one-arm-only correct counts, whereas disagreement is their sum, so offsetting gains and losses cancel in the score but not in the predictions. Cross-arm rows in Table 85 average the three matched seeds; within-arm rows compare the three seeds of one arm and share checkpoints with the cross-arm rows, so they are descriptive references, not a test of excess disagreement. On RTE seed 17, MIX and NORM both answer 193 of 277 examples correctly and disagree on 44, each alone correct on 22. The analysis supplement contains the script and per-pair counts.

Table 85: Accuracy difference and prediction disagreement on the held-aside encoder splits (RTE $n = 277$, MRPC $n = 408$). Cross-arm rows average seeds 17, 42 and 123 paired by seed; within-arm rows average the ordered seed pairs (17,42), (17,123) and (42,123), so their signed accuracy means are not treatment effects. Accuracy differences are percentage points (first minus second); disagreement is the percentage of examples on which the two models predict different labels, with its range across the three pairs.

Task	Comparison	Acc. diff. (pp)	Disagree (%)	Range
RTE	MIX−NORM (same seed)	+0.12	15.76	14.80–16.61
RTE	MIX−UNREG (same seed)	−0.36	17.69	13.72–20.94
RTE	NORM−UNREG (same seed)	−0.48	12.52	11.19–13.36
MRPC	MIX−NORM (same seed)	−0.33	8.01	7.35–8.58
MRPC	MIX−UNREG (same seed)	−0.74	9.40	9.07–9.80
MRPC	NORM−UNREG (same seed)	−0.41	6.78	6.13–7.84
MRPC	MIX, seed pairs	−1.31	7.52	6.37–8.33
MRPC	NORM, seed pairs	+0.16	6.37	5.64–6.86
MRPC	UNREG, seed pairs	−0.98	6.86	6.62–7.11
RTE	MIX, seed pairs	−2.17	16.85	15.52–17.69
RTE	NORM, seed pairs	−0.96	10.83	9.75–12.64
RTE	UNREG, seed pairs	−1.93	12.27	11.55–13.00

D Computational cost of the spectral instruments

D.1 Construction and arithmetic

We distinguish trainable capacity from frozen storage, setup cost, and the computation required to apply the adapter. Let a square layer have width d, selected tail dimension r, and rotated sub-block dimension $q \leq r$. Let b count the token vectors processed by that layer in one forward call. All costs below are per adapted matrix and exclude the base model's own storage and forward computation, the task head, and hardware-dependent constants. Appendix D.2 exposes the archived timing records.

SVD initialization and frozen storage. A dense SVD of a $d \times d$ matrix has conventional $O(d^3)$ arithmetic cost. For an $m \times n$ matrix with $p = \min(m, n)$, the corresponding economy-SVD cost is $O(mnp)$ and its two bases contain $(m + n)p$ entries. This is a one-time setup cost per distinct weight decomposition, not a decomposition at every optimization step. The practical implementation computes the SVD in float32 before casting stored factors to the adapter dtype, so initialization workspace can exceed persistent basis storage.

A tail-only implementation needs $2dr$ basis entries to apply its update. The implemented leading-plus-tail construction retains the complete bases, totaling $2d^2$ entries, plus $O(d)$ fixed coefficients and any orthogonal-map state. Freezing these entries avoids gradients and optimizer moments for them; it does not remove their memory cost. Setting the leading coefficients to zero in a full-basis implementation does not itself discard those bases or shorten its matrix products. Diagnostics that retain an additional full reference must also be included in memory accounting.

Trainable entries versus effective degrees of freedom. Fixed-tail coefficients require r trainable entries. Two dense $q \times q$ rotation parameters add $2q^2$ stored trainable entries; their orthogonal outputs have only $q(q-1)$ combined manifold degrees of freedom. Ambient input/output scalers add $2d$ entries. Thus the scaled unrotated and partially rotated forms store $r + 2d$ and $r + 2d + 2q^2$ trainable entries, respectively, before the task head. A directly trained unrestricted tail core instead has r^2 entries. These counts describe different parameterizations, not equal-capacity interventions. For example, $d = 768, r = 256$ gives 1,792 trainable entries per unrotated coordinate-scaled matrix but 1,179,648 frozen basis entries. The GLUE score comparison is not used to infer total training-memory savings.

Forward computation. Applying a diagonal tail update through its thin factors costs $O(bdr)$; ambient scaling adds only $O(bd)$. If partial rotations are materialized into the basis first, the two basis products add $O(dq^2)$ per reconstruction, in addition to the orthogonal-map construction cost. Dense $q \times q$ maps can incur $O(q^3)$ arithmetic. A dense tail core applied between thin basis products instead costs $O(bdr + br^2)$.

The practical forward path applies leading and tail basis products separately; together they span the complete bases. Its adapter branch therefore costs $O(bd^2)$ even when $q = 0$, plus rotation construction when enabled. This is additional to the $O(bd^2)$ base projection. A rank-ℓ LoRA branch costs $O(bd\ell)$, so a smaller trainable count need not mean a cheaper forward pass. Backpropagation also differentiates through the adapter operations, and end-to-end peak memory depends on retained activations, dtype, optimizer state, and checkpointing. The arithmetic counts alone do not establish a measured throughput ratio.

At two bytes per entry, the full frozen bases cost 2.25, 4, and 64 MiB per matrix at widths $d = 768, 1024, 4096$, respectively. Each is twice the storage of the corresponding base projection, and the two basis products add approximately twice its matrix-product arithmetic before rotations or penalties. Across the 48 RoBERTa-base projections, 56.6M basis entries are about 45% of a 125M-parameter backbone's weight storage at equal dtype; the fresh float32 runs require twice the two-byte figures. No measured usable-width limit is established beyond the tested backbones. At larger widths this quadratic storage and cubic SVD setup favor thin selected-band implementations where the update family permits them, or measuring conventional adapters after training. A strict tail-only branch at $r = d/3$ reduces basis storage by threefold; zeroing a leading core in the existing full-basis path does not realize that saving.

Regularization and diagnostics. Directly forming $M_E = U_L^\top E U_T$ and $M_D = V_L^\top D V_T$ costs $O(dkr)$ for a leading dimension $k = d - r$; evaluating their squared Frobenius penalties then costs $O(kr)$. The fresh runs instead cache the two leading projectors and use the equivalent commutator formula (13). This requires $O(d^2k)$ setup and $O(d^2)$ extra storage per layer, followed by $O(d^2)$ work per step. Raw mixing is measured in every arm; gradients through it are computed only when it is penalized. Neither form requires an SVD of the adapted weight. By contrast, exact ranked-subspace diagnostics use a new dense SVD of $W_0 + \Delta$ and can cost $O(d^3)$ per measured layer/checkpoint. Diagnostic time should be reported separately from optimization time and should not be charged only to the spectral methods in a controlled comparison.

Merging and deployment. At evaluation with dropout disabled, the additive update can be materialized and merged into W_0. A thin diagonal-tail construction costs $O(d^2r)$ to materialize, whereas the current full-basis product costs $O(d^3)$, plus the rotation and scaler operations; dense materialization requires $O(d^2)$ output storage. After merging and removing the adapter branch, the linear layer has the base model's shape and inference-time matrix multiplication. This removes the extra branch at deployment, while retaining the preceding SVD, training and merge costs. Algebraic merge equivalence does not guarantee identical finite-precision predictions. In multi-adapter serving, frozen pretrained bases can be shared across tasks; their quadratic storage need not be duplicated per adapter. Keeping branches permits task switching with additional arithmetic, whereas caching separately merged weights requires a dense weight copy per task. The 192-fold branch-arithmetic comparison is not a 192-fold trainable-adapter storage ratio.

D.2 ARCHIVED TIMING RECORDS

The archived experiments use public benchmarks, pretrained models, and the PyTorch/Hugging Face ecosystem. The documented hardware includes NVIDIA A100-SXM4-40GB and RTX-3090, but the exported summaries do not assign a verified device to every run. Table 86 exposes the available timing and memory fields. In the inspected runner, time spans the training call, including its validation and checkpoint work; peak memory is CUDA maximum allocated memory measured after training, following a pre-training reset. It is not total process memory, reserved memory, or SVD setup peak. The table therefore supports cost disclosure, not a controlled speed or memory comparison across methods. Token counts and matched LoRA timings are unavailable.

Table 86: Logged training-call minutes and peak allocated GiB (2^{30} bytes) for all 27 archived runs. These are raw resource observations with unverified per-run hardware attribution, not throughput benchmarks. Setup, post-training spectral diagnostics, and total project compute are not reconstructed by these fields.

Task	Condition	Seed	Minutes	Peak allocated GiB
RTE	A	17	12.06	4.408
RTE	A	42	10.44	7.790
RTE	B	17	16.12	4.423
RTE	B	42	11.62	7.787
RTE	C	17	17.46	4.422
RTE	C	42	11.52	7.790
MRPC	A	17	4.96	2.399
MRPC	A	42	2.97	3.737
MRPC	B	17	6.79	2.415
MRPC	B	42	3.89	3.728
MRPC	C	17	6.84	2.418
MRPC	C	42	4.03	3.728
CoLA	A	17	29.22	1.610
CoLA	A	42	32.33	1.599
CoLA	B	17	41.12	1.649
CoLA	B	42	39.62	1.644
CoLA	C	17	42.91	1.650
CoLA	C	42	39.00	1.650
STS-B	A	17	15.85	2.647
STS-B	A	42	18.54	2.637
STS-B	B	17	21.94	2.662
STS-B	B	42	24.14	2.648
STS-B	C	17	22.22	2.663
STS-B	C	42	23.94	2.648
SST-2	A	42	122.96	1.888
SST-2	B	42	164.46	1.922
SST-2	C	42	151.57	1.924

E HISTORICAL EXPERIMENT RECORDS

These earlier experiments retain their original recipes, selection rules and numerical records. They provide historical context and do not supply controlled comparisons for the primary studies. Their metrics are defined in Appendix B.1. Appendix E.4 describes the archived regularization study and Appendix E.6 its block allocation; broader reported GLUE and generation results follow in Appendices F.1 and F.2.

E.1 Historical backbones, tasks, and placement

The historical language-understanding evaluation uses RoBERTa-base and RoBERTa-large on a six-task GLUE subset: CoLA, SST-2, MRPC, STS-B, QNLI, and RTE (Wang et al., 2019). The archived regularization study uses RoBERTa-base on five of these tasks, excluding QNLI. Neither covers MNLI or QQP. Table 91 reports Matthews correlation for CoLA, accuracy for SST-2/MRPC/QNLI/RTE, and Pearson correlation for STS-B. The archived regularization study instead reports MRPC F1. These historical evaluations use official validation splits; the new fixed-endpoint study uses an inner split and evaluates official validation only after freezing its confirmation checkpoints (Appendix B.4). Backbone parameters are frozen; adapters and the task head train jointly. The adapted matrices are the query, key, value, and attention-output projections; feed-forward projections remain frozen.

E.2 Historical training configuration

The historical recipe uses AdamW with separate adapter and head learning rates, linear decay, a 6% warmup fraction, batch size 64, and maximum sequence length 256. Table 87 gives the recorded RoBERTa-base settings. The fresh controlled experiment uses its own common recipe, given in Appendix B.2. Imported benchmark methods retain their source protocols.

Table 87: Historical RoBERTa-base training settings. The scaler value initializes both ambient diagonal scalers. The new fixed-endpoint study has a separate protocol (Appendix B.4).

Task	Epochs	Head LR	Adapter LR	Scaler	Coefficients
CoLA	80	$5 \cdot 10^{-3}$	$3 \cdot 10^{-2}$	10^{-2}	10^{-1}
SST-2	40	10^{-2}	$4 \cdot 10^{-2}$	10^{-1}	10^{-1}
MRPC	30	10^{-3}	$5 \cdot 10^{-2}$	10^{-1}	10^{-1}
STS-B	60	$5 \cdot 10^{-3}$	10^{-2}	10^{-1}	10^{-1}
QNLI	25	10^{-3}	$2 \cdot 10^{-2}$	10^{-1}	10^{-1}
RTE	90	$5 \cdot 10^{-4}$	10^{-2}	10^{-2}	10^{-2}

Adapter dropout is zero and the nominal adapter scaling coefficient is one. The broad GLUE configurations use $k_{\mathrm{val}} = 256$, with $k_{\mathrm{vec}} = 0$ for UILinLoRA and 64 for UIOrthoLoRA. The mixing study uses 256 and zero, respectively. Compact rotations, when enabled, use an orthogonal parameterization; only the selected sub-block is rotated. With nonzero initialized scalers and coefficients, the scaled leading-plus-tail adapter need not have zero initial delta. The measured displacement from W_0 includes that initial contribution.

E.3 Historical seeds and checkpoint selection

The broader spectral evaluation protocol reports six seeds $\{42, 123, 2021, 17, 31415, 1054\}$; Table 91 retains its reported means and across-seed standard deviations. The imported comparison uses five runs. The mixing analysis uses the nine task/seed pairs with archived paired summary and module-level records: seeds 42 and 17 for RTE, MRPC, CoLA, and STS-B, and 42 for SST-2. These records permit independent reaggregation of the regularization comparison, whereas the broad GLUE table is a report of the original evaluation, not a newly reconstructed six-seed experiment.

The archived ablation runner selects the best validation checkpoint by the reported task metric, then reloads it for scoring and spectral measurements. Table 89 records each selected step; RTE seed-42 A/B/C select steps 312, 3198, and 1755. These are selection-set outcomes. Table 88 gives sample SD for the two-seed settings and individual values for the one-seed setting. The original controlled studies instead fix training endpoints before held-out evaluation.

E.4 Archived regularization study: design and summary

The archived regularization study uses RoBERTa-base (Liu et al., 2019) on RTE, MRPC, CoLA, STS-B, and SST-2. Adapters modify the query, key, value, and attention-output projections in each of its 12 layers, giving 48 adapted matrices. The configuration adapts 256 coefficients with $k_{\mathrm{vec}} = 0$, holding rotations out of the comparison. The analysis uses the scaled leading-plus-tail form (7). The

three conditions are A: no penalty, B: $\lambda_E = 10^{-3}, \lambda_D = 0$, and C: $\lambda_E = \lambda_D = 10^{-3}$. The backbone is frozen and the task head trains jointly with the adapter.

The analysis comprises seeds 42 and 17 for RTE, MRPC, CoLA, and STS-B, and seed 42 for SST-2: nine paired task/seed comparisons. Spectral quantities are measured at each run's validation-selected checkpoint, averaged over its 48 matrices, then averaged over seeds within a task. Runs, rather than layers, are the units of replication. Checkpoint selection is shared as a rule, not necessarily as an optimizer step. Appendix E gives the training settings and complete three-condition results; Appendix B.1 defines the metrics.

The diagnostic cutoff is $k = 512$ in each 768×768 matrix, leaving the 256 adapted tail directions. Table 88 reports task score, relative norm ρ_F, and cross energy. Appendix E.5 retains q_{off} and both largest-principal-angle sines. The latter saturate near one under A: a single boundary direction can cause this, so they are supplementary indicators, not measures of how much the entire leading subspace moved.

Table 88: All three conditions: A unregularized, B left-only, C two-sided. Entries are mean±sample SD across n seeds; $n = 1$ is a single run, with no estimated dispersion. Scores and cross fractions are multiplied by 100. Spectral metrics first average 48 modules within each selected checkpoint. Scores use MRPC F1 here, unlike Table 91. Per-seed values and selection steps appear in Appendix E.5.

Task	Condition	n	Task score	ρ_F	Cross energy (%)
RTE	A	2	77.08 ± 1.28	0.3787 ± 0.0142	43.41 ± 0.15
RTE	B	2	79.06 ± 1.02	0.2847 ± 0.0212	43.55 ± 0.43
RTE	C	2	76.71 ± 0.77	0.0626 ± 0.0004	27.45 ± 1.93
MRPC	A	2	92.15 ± 0.45	0.3174 ± 0.0571	43.25 ± 0.14
MRPC	B	2	92.40 ± 0.06	0.1958 ± 0.0552	40.32 ± 2.13
MRPC	C	2	91.32 ± 0.04	0.0430 ± 0.0109	10.08 ± 3.47
CoLA	A	2	64.09 ± 1.40	0.3076 ± 0.0048	43.75 ± 0.10
CoLA	B	2	64.30 ± 1.11	0.2057 ± 0.0041	42.38 ± 0.23
CoLA	C	2	63.14 ± 1.22	0.0334 ± 0.0208	21.09 ± 5.63
STS-B	A	2	90.80 ± 0.27	0.3820 ± 0.0030	43.46 ± 0.01
STS-B	B	2	90.48 ± 0.13	0.2506 ± 0.0702	41.96 ± 0.23
STS-B	C	2	90.58 ± 0.39	0.0679 ± 0.0128	13.71 ± 1.36
SST-2	A	1	94.72	0.2820	43.33
SST-2	B	1	94.38	0.2676	40.19
SST-2	C	1	93.69	0.0557	17.21

E.5 Complete mixing results

The accompanying `scripts/build_mixing_tables.py` generates Tables 88 and 89 from the included per-task/seed CSV summaries. These records retain checkpoint steps and metric columns. The analysis includes the paired A/B/C runs and excludes separate smoke tests and C-only reruns. This makes the aggregation independently checkable without rerunning training.

E.6 Archived block allocation and factor asymmetry

The aggregate off-tail ratio combines direct leading adaptation and cross interaction. To separate them, we reconstruct the four energy fractions from the saved per-matrix diagnostics, square before averaging, and retain equal weighting of modules within each run. Figure 8 shows the resulting leading, cross, and tail allocation. Appendix E.7 gives all four blocks and accounts for the small finite-precision difference between the recorded coordinate and ambient norms.

Without regularization, cross blocks contain approximately 43% of update energy, the leading block 46%, and the tail 10%. Under C, cross share is 10–27%, tail share 22–34%, and leading share 44–64%; the latter rises in four task means. The retained leading block agrees with the structure in Proposition 2. These are fractions of the total update, which includes insertion. Appendix C.1 calculates the

Table 89: Every archived mixing run and its validation-selected step. A: no penalty; B: left-only; C: two-sided. Scores are in native metric units (MRPC F1). Spectral values are means over 48 modules of that run: μ_E, ν_D and drift use operator norms, while ρ_F, q_{off} are Frobenius norm ratios. Drift is the largest-angle sine, often near its ceiling.

Task	Cond.	Seed	Step	Score	μ_E	ν_D	ρ_F	q_{off}	Drift_U	Drift_V
RTE	A	17	6318	0.7798	1.392	1.464	0.389	0.948	1.000	1.000
RTE	A	42	312	0.7617	1.376	1.420	0.369	0.947	1.000	1.000
RTE	B	17	3510	0.7978	0.331	2.173	0.300	0.934	0.999	0.999
RTE	B	42	3198	0.7834	0.225	2.051	0.270	0.933	1.000	1.000
RTE	C	17	3588	0.7726	0.208	0.201	0.062	0.823	0.534	0.523
RTE	C	42	1755	0.7617	0.234	0.207	0.063	0.853	0.511	0.500
MRPC	A	17	1725	0.9247	1.347	1.405	0.358	0.947	1.000	1.000
MRPC	A	42	1102	0.9183	1.217	1.284	0.277	0.948	0.998	0.999
MRPC	B	17	2300	0.9236	0.242	1.714	0.235	0.941	1.000	1.000
MRPC	B	42	696	0.9244	0.276	1.361	0.157	0.955	0.966	0.966
MRPC	C	17	1495	0.9129	0.226	0.182	0.035	0.865	0.606	0.601
MRPC	C	42	1334	0.9134	0.234	0.193	0.051	0.848	0.728	0.723
CoLA	A	17	13400	0.6508	1.248	1.325	0.304	0.946	0.999	0.999
CoLA	A	42	4556	0.6310	1.250	1.346	0.311	0.946	0.999	0.999
CoLA	B	17	4556	0.6351	0.288	1.631	0.209	0.937	0.983	0.983
CoLA	B	42	4556	0.6508	0.285	1.605	0.203	0.939	0.995	0.995
CoLA	C	17	14472	0.6401	0.153	0.165	0.048	0.860	0.649	0.643
CoLA	C	42	6432	0.6228	0.166	0.123	0.019	0.900	0.468	0.460
STS-B	A	17	8820	0.9099	1.355	1.450	0.384	0.947	1.000	1.000
STS-B	A	42	9720	0.9061	1.395	1.443	0.380	0.947	1.000	0.999
STS-B	B	17	9360	0.9039	0.231	1.778	0.300	0.936	1.000	1.000
STS-B	B	42	2340	0.9057	0.319	1.665	0.201	0.943	1.000	1.000
STS-B	C	17	8820	0.9030	0.200	0.256	0.059	0.816	0.712	0.709
STS-B	C	42	9540	0.9085	0.203	0.240	0.077	0.795	0.786	0.784
SST-2	A	42	71570	0.9472	1.219	1.255	0.282	0.948	1.000	1.000
SST-2	B	42	50520	0.9438	0.185	1.604	0.268	0.943	1.000	1.000
SST-2	C	42	46310	0.9369	0.134	0.170	0.056	0.850	0.679	0.672

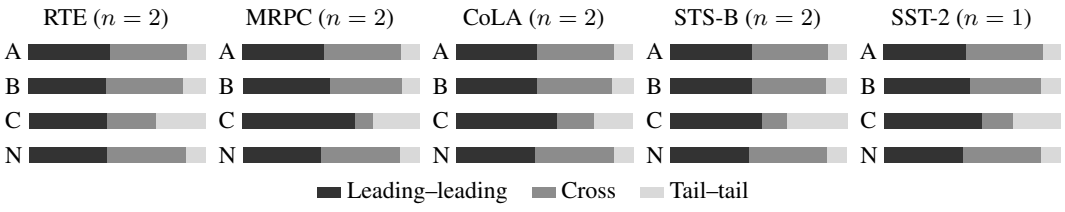


Figure 8: Leading/cross/tail allocation, averaged over modules then n seeds. A: no penalty; B: left-only; C: two-sided. N is the dimension-only reference $(4/9, 4/9, 1/9)$, not a training run. C reduces cross share but retains a leading block; Table 88 supplies magnitude and dispersion, and Appendix E.7 supplies each seed's four fractions.

prescribed initial contribution, and the fresh study separately measures learned displacement and leading identity alignment.

Table 90 in Appendix E.7 gives each run's mean block fractions; the supplementary archive contains all module records.

One-sided control and factor asymmetry. B has the highest mean score on RTE, MRPC and CoLA while its cross shares remain about 40–44% (Table 88). The score ordering therefore differs from the interaction ordering. In all nine A/B pairs, μ_E falls and ν_D rises. Reciprocal rescaling of the two factors can also produce this asymmetry, as shown in Appendix A.9; the effective update is needed to interpret the factor measurements.

Interpreting normalized geometry. The archived regularization study changes normalized block allocation. Uniformly rescaling a fixed trained matrix would preserve those fractions. A size penalty during training acts through the adapter factors and can change them, which motivates the separate controlled experiment in Section 5.2. The archived measurements describe the selected endpoints.

An isotropically oriented perturbation has expected leading/cross/tail fractions k^2/d^2, $2kr/d^2$, and r^2/d^2 (Appendix A.6). Figure 8 makes this dimension-only reference explicit. A lies nearby, but agreement of three aggregate fractions is not a test of isotropy or evidence that pretrained structure is unused. Equal-dimensional random-frame controls and measurements of other adapters are needed to distinguish those interpretations.

E.7 Reconstructing four-block allocation

The 27 run summaries are accompanied by 1,296 module-level records: 48 adapted matrices for each A/B/C run. Their saved Frobenius leakage ratios use $\|\Delta\|_F$ in the denominator, whereas the off-tail ratio uses $\|C\|_F$. Write these recorded quantities as ℓ_{ab} for $ab \in \{LL, LT, TL\}$ and q for the off-tail ratio. To obtain an allocation normalized in the recorded coordinate frame, we compute per matrix

$$\widehat{p}_{ab} = q^2 \frac{\ell_{ab}^2}{\ell_{LL}^2 + \ell_{LT}^2 + \ell_{TL}^2}, \qquad \widehat{p}_{TT} = 1 - q^2. \tag{47}$$

For exact orthogonal bases these agree with the fractions defined in Section 4. With saved finite-precision bases, the inferred discrepancy $\left|(\ell_{LL}^2 + \ell_{LT}^2 + \ell_{TL}^2)/q^2 - 1\right|$ is at most $2.64 \cdot 10^{-4}$ across the archive. The normalization makes the four recorded fractions sum to one without assuming perfect numerical orthogonality. Diagnostic denominators also include a 10^{-12} numerical stabilizer. We square each matrix's ratios before averaging, never a task-mean ratio.

The analysis script reconciles every shared summary metric with the mean of its 48 saved module values before calculating these fractions. It excludes smoke tests and unmatched C-only reruns. Cross share falls in 410 of 432 matched A/C module pairs, q_{off} in 332, and left/right drift in 260/256; both drift run means fall in all nine pairs. These are within-run heterogeneity counts, not independent replication or uniform layerwise stabilization.

Table 90: Four-block energy percentages for every archived seed, averaged over its 48 modules. Averaging seeds gives Figure 8. The middle columns sum to cross share; rows sum to 100% up to rounding. SST-2 has only seed 42.

Task	Condition	Seed	Leading–leading	Leading–tail	Tail–leading	Tail–tail
RTE	A	17	46.61	22.18	21.11	10.10
RTE	A	42	46.17	22.05	21.47	10.31
RTE	B	17	43.37	19.86	23.99	12.77
RTE	B	42	43.89	20.08	23.18	12.86
RTE	C	17	43.45	13.09	13.00	30.46
RTE	C	42	45.16	14.72	14.10	26.02
MRPC	A	17	46.41	22.20	21.15	10.24
MRPC	A	42	46.72	22.10	21.04	10.13
MRPC	B	17	46.81	21.77	20.05	11.37
MRPC	B	42	52.45	23.99	14.82	8.74
MRPC	C	17	62.62	6.77	5.77	24.85
MRPC	C	42	64.68	3.89	3.74	27.69
CoLA	A	17	45.89	22.58	21.10	10.43
CoLA	A	42	45.59	22.39	21.43	10.59
CoLA	B	17	45.25	21.18	21.37	12.20
CoLA	B	42	45.98	21.54	20.69	11.80
CoLA	C	17	57.57	9.96	7.15	25.33
CoLA	C	42	56.39	14.78	10.29	18.54
STS-B	A	17	46.24	22.32	21.15	10.29
STS-B	A	42	46.28	22.27	21.18	10.27
STS-B	B	17	45.48	20.35	21.78	12.39
STS-B	B	42	47.21	21.90	19.90	10.99
STS-B	C	17	52.98	8.74	5.93	32.35
STS-B	C	42	51.55	7.75	5.01	35.69
SST-2	A	42	46.52	22.38	20.95	10.15
SST-2	B	42	48.76	21.51	18.68	11.05
SST-2	C	42	55.74	9.26	7.94	27.06

F SUPPORTING DOWNSTREAM EVALUATIONS

F.1 HISTORICAL GLUE RESULTS

These historical six-task scores provide context for the adapter constructions. Imported baselines and spectral runs use separate protocols; this table is not a controlled ranking. All original numeric entries are preserved.

Table 91: Historical RoBERTa results on six GLUE tasks: reported central values±SD, multiplied by 100, with source rounding retained. Imported reference runs (Albert et al., 2025) and spectral runs use separate tuning protocols. MRPC is accuracy throughout. Params are source-reported for imported runs; † marks adapter-only counts derived from the stated spectral configurations rather than verified checkpoint inventories. Appendix D accounts for frozen storage and computation.

Method	SST-2	MRPC Acc.	CoLA	QNLI	RTE	STS-B	Params
RoBERTa-base							
Imported reference runs (RandLoRA protocol; five runs)							
VeRA-1024	91.9±0.4	88.4±1.2	59.9±2.2	90.5±0.4	74.9±1.5	90.4±0.2	0.26M
LoRA-4	94.4±0.5	87.3±0.2	58.4±0.8	92.7±0.2	71.5±1.2	90.5±0.1	0.7M
RandLoRA-64	92.2±0.3	88.0±1.5	59.4±2.1	91.3±0.4	74.7±1.9	90.3±0.2	0.7M
Spectral instruments (reported protocol; six seeds)							
UIOrthoLoRA	94.8±0.3	92.0±1.0	64.4±1.0	92.4±0.1	78.7±1.0	90.7±0.1	0.479M†
UILinLoRA	93.1±1.0	90.0±1.0	65.8±1.0	92.4±0.2	76.2±1.0	90.6±0.1	0.086M†
RoBERTa-large							
Imported reference runs (RandLoRA protocol; five runs)							
VeRA-256	95.8±0.3	89.3±1.2	65.3±1.1	94.1±0.3	81.6±0.8	91.8±0.1	0.26M
LoRA-4	95.5±0.2	87.2±0.7	64.7±1.2	94.5±0.1	83.6±0.4	91.8±0.1	1.8M
RandLoRA-100	95.5±0.3	90.1±0.4	67.4±0.3	94.1±0.3	84.5±0.3	91.4±0.6	1.8M
Spectral instruments (reported protocol; six seeds)							
UIOrthoLoRA	96.3±0.3	92.5±0.2	67.4±0.5	95.0±0.1	86.0±0.2	92.1±0.2	1.008M†
UILinLoRA	96.3±0.2	93.4±0.7	68.7±1.0	94.7±0.1	86.0±1.1	92.0±0.2	0.221M†

F.2 GENERATION

Table 92 reports historical GPT-2 Medium (Radford et al., 2019) results on E2E data-to-text generation (Novikova et al., 2017). The fixed-core and partially rotated constructions attain BLEU 68.7 and 68.8. These runs have no spectral-block measurements or interaction-regularization comparison. The imported LoRA and FourierFT entries average three final-epoch runs (Gao et al., 2024). Training, decoding and uncertainty conventions differ across sources, so these values provide context rather than a controlled comparison.

Table 92: Supporting GPT-2 Medium E2E results. Non-UI values are from Gao et al. (2024), including their imported full-fine-tuning and adapter results; L and H denote Lin-style and Houlsby-style adapters. Uncertainties retain source conventions. Trainable counts exclude frozen-basis storage; these scores do not measure spectral mixing.

Method	Trainable	BLEU	NIST	METEOR	ROUGE-L	CIDEr
Full FT	354.92M	68.2	8.62	46.2	71.0	2.47
Adapter^L	0.37M	66.3	8.41	45.0	69.8	2.40
Adapter^L	11.09M	68.9	8.71	46.1	71.3	2.47
Adapter^H	11.09M	$67.3_{\pm.6}$	$8.5_{\pm.07}$	$46.0_{\pm.2}$	$70.7_{\pm.2}$	$2.44_{\pm.01}$
LoRA	0.35M	$68.9_{\pm.3}$	$8.76_{\pm.06}$	$46.6_{\pm.1}$	$71.5_{\pm.1}$	$2.53_{\pm.03}$
FourierFT	0.048M	$69.1_{\pm.1}$	$8.82_{\pm.05}$	$47.0_{\pm.3}$	$71.8_{\pm.1}$	$2.51_{\pm.02}$
UIOrthoLoRA	0.35M	$68.8_{\pm.1}$	$8.729_{\pm.06}$	$46.71_{\pm.3}$	$71.6_{\pm.02}$	$2.47_{\pm.01}$
UILinLoRA	0.078M	$68.7_{\pm.1}$	$8.732_{\pm.1}$	$46.52_{\pm.3}$	$71.45_{\pm.01}$	$2.46_{\pm.02}$

G SCOPE OF FUNCTIONAL-RETENTION EVIDENCE

Exploratory retention measurements cover 40 model–method–capacity–task conditions across Gemma-3-12B-it and Llama-3.2-3B-it on TriviaQA and HotpotQA. The records do not establish the full run inventory, selection rule, matched adaptation quality or checkpoint-level geometry. These measurements are excluded from our claims; that decision was made after their mixed outcomes were observed. They remain in the analysis archive. A test of retention would require adaptation and retention measurements on the same checkpoints, with comparable tuning and multiple seeds.