DO LORA MERGE DIAGNOSTICS SCORE THE DEPLOYED MODEL?

Anonymous authors
Paper under double-blind review

ABSTRACT

A LoRA merge diagnostic can fail at three distinct stages: the update it scores, the curvature it estimates, and the endpoint it is intended to predict. We analyze these mismatches while holding candidate construction and evaluation roles explicit. An exact factor-covariance identity separates factor aggregation from scalar-equivalent correction and rank truncation. For sequence-balanced Fisher estimation, probe-count correction removes aggregation bias but retains cross-probe variance; categorical marginalization removes label-sampling variance. A five-adapter Pythia-1.4B case study exhibits a $4.78\times$ deployed-to-reference norm ratio and task-dependent local-quadratic miscalibration. Controlled studies use twenty independently trained banks per panel, two architectures, and a fixed twenty-four-candidate library. Replacing uniform logit drift with specialist-probability weighting lowers mean worst-task KL regret from 0.3981 to 0.0238 in a nonlinear classifier and from 0.0820 to 0.0202 in causal attention. This preservation gain does not establish a native-accuracy gain: weighted-minus-uniform attention accuracy is -0.058 percentage points, with paired 95% interval $[-0.615, 0.505]$. Selection for native retention using selection labels improves accuracy over KL selection by 0.743 percentage points $[0.243, 1.316]$. The results support matching diagnostics to deployed updates and preservation geometry, while evaluating native utility separately.

1 INTRODUCTION

A diagnostic of LoRA mergeability should describe the model that will actually run. Low-Rank Adaptation (LoRA) represents a specialist's layer update as a product BA (Hu et al., 2022). Combining the factors separately produces a product of sums, whereas task arithmetic combines the realized updates through a sum of products (Ilharco et al., 2023). A diagnostic evaluated on the latter can therefore describe a different model from the deployed merge. We characterize this discrepancy through an exact factor-covariance identity and distinguish its scalar correction from the error introduced by rank truncation.

The deployed update is only the first part of the problem. Specialist-relative Kullback–Leibler (KL) divergence has zero gradient at the specialist, and its local quadratic uses the model Fisher. That Fisher must share the endpoint's sequence and position distribution. Averaging sampled token scores before taking an outer product changes this target. Probe-count normalization restores the expectation, but the corrected estimator still contains cross-products whose zero mean does not imply zero variance. Estimator bias, estimator variance, and finite-displacement approximation error consequently require separate checks.

A further distinction arises when a diagnostic selects among candidates. Accurate prediction of KL magnitude and identification of its minimum are different requirements. Even scores computed from the same candidate logits can disagree: uniform output drift and specialist-relative KL use different output geometry. High correlation across candidates can coexist with consistently selecting the wrong minimum, especially when large-displacement candidates dominate the ordering. We therefore compare selectors on a fixed candidate library and evaluate their decisions on an independent endpoint sample.

Our analysis connects these three stages. The deployment decomposition identifies cross-factor covariance, scalar-equivalent correction, and a separate rank-truncation residual. The estimator

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.

analysis separates sequence weighting, cross-probe variance, and label-sampling variance. A finite-logit control then replaces uniform geometry with the specialist's categorical covariance, allowing us to distinguish output weighting from network linearization and the curvature remaining along the logit path. The contribution is this matched analysis of diagnostics, rather than a new Fisher identity or a claim that factor/product mismatch was previously unknown.

The empirical evidence has two scopes. A five-specialist Pythia-1.4B bank demonstrates deployment mismatch and local miscalibration. Three controlled panels provide bank-level mechanism tests: two nonlinear-classifier implementations and one causal-attention model, each with twenty independently trained banks conditional on its own fixed base and teachers. Probability weighting reduces preservation regret by 94.0% in classifier panel B and by 0.0617 $[0.0308, 0.0946]$ in attention panel C. The attention comparison does not establish an accuracy gain. We retain both findings because preserving specialist distributions and recovering specialist competence are distinct objectives, and a useful merge evaluation should measure both.

2 Related Work

Weight averaging, task arithmetic, Fisher merging, TIES, and DARE combine models through different scales, coordinate weights, and interference rules (Wortsman et al., 2022; Ilharco et al., 2023; Matena & Raffel, 2022; Yadav et al., 2023; Yu et al., 2024). RegMean and AdaMerging use activation statistics or calibration objectives (Jin et al., 2023; Yang et al., 2024). We study whether a diagnostic matches the resulting deployed candidate and the endpoint used to judge it.

Factor/product mismatch is recognized in PEFT's distinct `linear`, `cat`, and `svd` operators (Mangrulkar & Paul, 2024) and in product-aligned federated aggregation (Jeon et al., 2026). KnOTS aligns realized updates in a shared SVD basis (Stoica et al., 2025); Compress then Merge enforces the rank bottleneck before combination (He et al., 2026); Crowded in B-Space calibrates shared output-side directions (Tang & Yang, 2026). Our decomposition separates scalar and covariance terms from the subsequent rank residual. Native-rank KnOTS controls test whether truncation explains its behavior in the controlled experiments.

GeoMerge distinguishes symmetry-aware geometry from symmetry-aware averaging (da Silva et al., 2026). For curvature, the model Fisher differs from observed-label gradient outer products (Kunstner et al., 2019), while categorical marginalization and Fisher estimator variance are established subjects (Martens, 2020; Soen & Sun, 2021). We use these distinctions to align score sampling with a sequence-balanced behavioral target.

Function-space merging is also established. Output-space projection uses a squared-output calibration objective and a basis-dependent diagnostic (Evans et al., 2026); Fisher–Rao merging targets predictive behavior through geometric averaging (Wang et al., 2026). Multi-teacher distillation learns merge coefficients using KL objectives (Dalili & Mahdavi, 2025), and SimMerge selects operators from similarity signals (Bolton et al., 2026). Our uniform-logit baseline is a controlled metric comparison, not a reproduction of these methods. Holding candidates and data roles fixed isolates how output-coordinate covariance affects preservation selection, calibration, and native utility.

3 The Deployed Update

Adapter i changes layer ℓ by $s_{\ell,i}B_{\ell,i}A_{\ell,i}$. Stack these products into $u_i \in \mathbb{R}^p$, and let $U = [u_1, \ldots, u_K]$. The product is invariant under $(B, A) \mapsto (BS, S^{-1}A)$ for invertible S, whereas factor cosines are not. A candidate deploys v after every mask, scale, multiplication, and rank map. Relative to declared coefficients α, its deployment residual is

$$e = v - U\alpha. \tag{1}$$

A residual can be consistent with a merger's objective while invalidating the substitution of $U\alpha$ for its actual update.

3.1 FACTOR COVARIANCE, SCALAR EQUIVALENCE, AND RANK

For one layer with common scale s, write $w(\alpha) = s\sum_i \alpha_i B_i A_i$, $S_\alpha = \sum_i \alpha_i$, and

$$v_{\mathrm{fac}}(\alpha) = s\Big(\sum_i \alpha_i B_i\Big)\Big(\sum_i \alpha_i A_i\Big). \tag{2}$$

The map is homogeneous of degree two: $v_{\mathrm{fac}}(c\alpha) = c^2 v_{\mathrm{fac}}(\alpha)$, while $w(c\alpha) = cw(\alpha)$.

Proposition 1 (Factor-covariance residual)**.** *For compatible factors and real coefficients,*

$$v_{\mathrm{fac}}(\alpha) = S_\alpha w(\alpha) - s\sum_{i<j} \alpha_i\alpha_j (B_i - B_j)(A_i - A_j). \tag{3}$$

For $S_\alpha \neq 0$, coefficient correction $v_{\mathrm{corr}} = v_{\mathrm{fac}}/S_\alpha$ is exact when either factor set is common. For nonnegative coefficients with $S_\alpha > 0$, $p_i = \alpha_i/S_\alpha$, $\bar{A} = \sum_i p_i A_i$, and $\bar{B} = \sum_i p_i B_i$ give

$$\|v_{\mathrm{corr}} - w\|_F \le |s| S_\alpha \sqrt{\sum_i p_i \left\|B_i - \bar{B}\right\|_F^2} \sqrt{\sum_i p_i \left\|A_i - \bar{A}\right\|_{\mathrm{op}}^2}. \tag{4}$$

The proof pairs the two cross-products for each unordered adapter pair; weighted Cauchy–Schwarz then bounds their covariance (Appendix A). Small absolute covariance does not imply small relative error when w nearly cancels. Direct product construction avoids unstable division near $S_\alpha = 0$.

Uniform coefficients require a scalar comparator. For $\alpha = \lambda\mathbf{1}$, $\lambda > 0$,

$$v_{\mathrm{corr}}(\lambda\mathbf{1}) = K\lambda v_{\mathrm{fac}}(\mathbf{1}/K) = v_{\mathrm{fac}}\left(\sqrt{\lambda/K}\,\mathbf{1}\right). \tag{5}$$

Correction is therefore a point on the factor-scaling ray, not an additional candidate direction. Unit factor arithmetic is exactly K^2 times factor averaging, regardless of initialization; its approximately K-fold relation to the product sum requires near-common factors.

For K rank-r adapters, exact concatenation represents each layer of w at rank at most Kr. Rank-matched deployment instead uses the truncated SVD $\mathrm{Proj}_r(w)$, with

$$\|\mathrm{Proj}_r(w) - w\|_F^2 = \sum_{j>r} \sigma_j(w)^2. \tag{6}$$

The rank bottleneck is separate from aggregation. PEFT's documented `linear` rule uses square-root weights on both factors, while `cat` represents the product sum and `svd` compresses it (Mangrulkar & Paul, 2024). The coefficient correction above refers to Equation (2); nonuniform square-root weighting generally has no equivalent scalar correction.

3.2 THE BEHAVIORAL TARGET AND ITS LOCAL APPROXIMATION

Let Q_t draw a task sequence and then a uniformly selected eligible position. Prefixes are fixed by that source sequence; sampled labels do not alter later prefixes. A prefix $x \sim Q_t$ defines the specialist-relative target

$$\mathcal{D}_t(v) = \mathbb{E}_{x\sim Q_t}\mathrm{KL}(p_{\theta_0+u_t}(\cdot \mid x) \,\|\, p_{\theta_0+v}(\cdot \mid x)). \tag{7}$$

For a specialist-sampled label, define $s_t(x,y) = \nabla_u \log p_{\theta_0+u}(y \mid x)|_{u=u_t}$ and $\mathcal{F}_t = \mathbb{E}[s_t s_t^\top]$. With $h = v - u_t$, $b_t = \alpha - \mathbf{e}_t$ (where $\mathbf{e}_t$ is the t-th standard basis vector, so that $Ue_t = u_t$), and $G_t = U^\top \mathcal{F}_t U$, the standard integral identity and algebraic decomposition are

$$\mathcal{D}_t(v) = \int_0^1 (1-a) h^\top H_t^{\mathrm{KL}}(a) h \, da, \tag{8}$$

$$q_t(v) := \tfrac{1}{2} h^\top \mathcal{F}_t h = \tfrac{1}{2} b_t^\top G_t b_t + b_t^\top U^\top \mathcal{F}_t e + \tfrac{1}{2}\|e\|_{\mathcal{F}_t}^2. \tag{9}$$

Here $H_t^{\mathrm{KL}}(a) = \nabla^2 \mathcal{D}_t(u_t + ah)$. Under Assumption 4, a Lipschitz Hessian along the entire segment gives $|\mathcal{D}_t - q_t| \le L_t^{\mathrm{KL}} \|h\|^3/6$. The endpoint residual does not estimate that constant. A supervised gradient belongs to a downstream-risk expansion, not to KL, whose derivative vanishes at the specialist. Appendix A states the regularity conditions and complete argument.

4 MATCH THE FISHER AND THE OUTPUT METRIC

4.1 AGGREGATION BIAS AND VARIANCE ARE DIFFERENT ERRORS

Condition on a sequence with T_z eligible positions, and draw independent specialist labels. Let $\widetilde{s} = T_z^{-1} \sum_\tau s_\tau$ and $\mathcal{F}_\tau = \mathbb{E}[s_\tau s_\tau^\top \mid z]$. Zero conditional score means give the following identity.

Proposition 2 (Sequence-balanced target)**.**

$$\mathbb{E}[\widetilde{s}\widetilde{s}^\top \mid z] = T_z^{-2} \sum_\tau \mathcal{F}_\tau, \qquad \mathbb{E}[T_z \widetilde{s}\widetilde{s}^\top \mid z] = T_z^{-1} \sum_\tau \mathcal{F}_\tau. \tag{10}$$

Thus $\sqrt{T_z}\widetilde{s}$ restores sequence-balanced weighting. For only m_z uniformly sampled positions, the correction is $\sqrt{m_z}$ instead. Define the uncorrected covariance as $\widetilde{\mathcal{F}}_t := \mathbb{E}_z[\mathbb{E}[\widetilde{s}\widetilde{s}^\top \mid z]]$, where z follows the task-t sequence distribution. If $T_{\min} \le T_z \le T_{\max}$, the following PSD sandwich holds: $T_{\min}\widetilde{\mathcal{F}}_t \preceq \mathcal{F}_t \preceq T_{\max}\widetilde{\mathcal{F}}_t$. Variable lengths need not induce a single multiplicative bias.

For centered independent score projections $a_j = s_j^\top h$, define $Y_{\text{sum}} = m^{-1}(\sum_j a_j)^2$ and $Y_{\text{tok}} = m^{-1}\sum_j a_j^2$. They have the same mean but different variances.

Proposition 3 (Cross-probe variance)**.** *Conditionally on the sequence and sampled positions, if $v_j = \mathbb{E}[a_j^2]$ and fourth moments are finite, then*

$$\operatorname{Var}(Y_{\text{sum}}) - \operatorname{Var}(Y_{\text{tok}}) = \frac{4}{m^2} \sum_{i<j} v_i v_j. \tag{11}$$

For identically distributed probes, these variances are $\mu_4/m + (2 - 3/m)v^2$ and $(\mu_4 - v^2)/m$, respectively.

The cross-products have zero covariance with the squared terms, giving the identity. For $v > 0$, the summed estimator retains limiting variance $2v^2$. Half-quadratic estimators divide both variances by four. Longer aggregation does not replace independent replicates.

For specialist probabilities p_t and directional logit derivative $j_h = D_u \ell_{\theta_0+u}|_{u_t}[h]$, categorical marginalization gives

$$q_t^{\text{marg}}(v) = \tfrac{1}{2}\mathbb{E}_{Q_t}\left[\sum_y p_{t,y} j_{h,y}^2 - \Big(\sum_y p_{t,y} j_{h,y}\Big)^2\right] = q_t(v). \tag{12}$$

This standard Fisher identity (Martens, 2020) removes label-sampling variance, not random-position or source-sequence variation. Its input is a derivative at the specialist, not a candidate forward pass. Comparisons at fixed probes and fixed differentiation cost are consequently different experiments.

4.2 AN INTERMEDIATE FINITE-LOGIT DIAGNOSTIC

For V output classes, let $\delta(x) = \ell_v(x) - \ell_t(x)$, $C = I - \mathbf{1}\mathbf{1}^\top/V$, and $G(p) = \operatorname{diag}(p) - pp^\top$. Uniform centered output drift and its specialist-weighted counterpart are

$$q_t^{\text{out}}(v) = \mathbb{E}_{Q_t}\frac{\|C\delta\|^2}{V}, \qquad q_t^{\text{wlog}}(v) = \tfrac{1}{2}\mathbb{E}_{Q_t}\delta^\top G(p_t)\delta. \tag{13}$$

Both execute the same candidate on the same contexts. The latter replaces uniform coordinate covariance by the categorical covariance and includes the probability-weighted mean subtraction. It is a mechanism control, not a claim of cheaper evaluation than direct KL.

The distinction is operational because the finite-logit KL has an exact form. For $p_a(y) \propto p_t(y)e^{a\delta_y}$,

$$D_x = \log \mathbb{E}_{p_t} e^{\delta} - \mathbb{E}_{p_t}\delta = \int_0^1 (1-a)\operatorname{Var}_{p_a}(\delta)\, da. \tag{14}$$

Writing $r_\ell = \delta - j_h$ separates the two approximation steps:

$$q_t^{\text{wlog}} - q_t = \mathbb{E}_{Q_t}\left[j_h^\top G(p_t) r_\ell + \tfrac{1}{2} r_\ell^\top G(p_t) r_\ell\right], \tag{15}$$

$$\mathcal{D}_t - q_t^{\text{wlog}} = \mathbb{E}_{Q_t}\int_0^1 (1-a)\delta^\top\{G(p_a) - G(p_t)\}\delta\, da. \tag{16}$$

The first term is finite network-linearization error. The second remains even with exact candidate logits because the categorical covariance changes along the logit path. A weighted score can select a better candidate while still miscalibrating its KL magnitude.

5 Experimental Design

5.1 Pretrained-model case study and controlled panels

The language-model case study uses five rank-16 Pythia-1.4B specialists on GSM8K, HumanEval, MBPP, Alpaca, and HH-RLHF (Biderman et al., 2023; Cobbe et al., 2021; Chen et al., 2021; Austin et al., 2021; Taori et al., 2023; Bai et al., 2022). Training uses at most 500 examples, one epoch, attention modules `query_key_value` and `dense`, LoRA scale two, zero dropout, maximum length 256, and AdamW learning rate 2×10^{-4}. The evaluated factor-coordinate candidates are averaging, unit arithmetic, TIES, DARE, and Fisher merging. Appendix B gives the full recipe and data windows. HumanEval geometry reuses its 64 specialist-training examples and is therefore an in-sample control; the other four tasks form the primary diagnostic summary. The Pythia local-score and KL matrix uses geometry contexts and is reported as point estimates; it does not provide the independent endpoint evaluation used in the controlled panels.

Controlled panel A isolates estimator and operator mechanisms; panel B independently reimplements the evaluation with a newly generated base and teacher collection to test output weighting and native-rank alignment. Each panel contains twenty independently trained adapter banks conditional on its own fixed teachers and base. The predictive model is

$$p_W(y \mid x) = \mathrm{softmax}(2W_2 \tanh(W_1 x))_y, \quad W_1 \in \mathbb{R}^{48\times 32},\ W_2 \in \mathbb{R}^{32\times 48}. \tag{17}$$

Five nonlinear teachers define labels on task-specific Gaussian inputs. Each specialist trains rank-16 updates on 512 examples for 160 AdamW steps. Each bank has 16 geometry, 48 selection, and 96 endpoint sequences per task, with 8, 16, or 32 positions that share a sequence-level latent variable. The three role streams are independent. Panel B changes the base, teachers, and implementation, but not the architecture; it therefore tests robustness within Equation (17).

Panel C changes the computation to residual causal attention. With rows of X representing sequence positions, its logits are

$$\ell_W(X) = 2\{X + A_W(X)XW_V^\top\}W_O^\top, \quad A_W(X) = \mathrm{softmax}\left(\frac{XW_Q^\top W_K X^\top}{\sqrt{32}} + M_{\text{causal}}\right). \tag{18}$$

All weight matrices are 32×32. LoRA adapts W_Q and W_O; W_K and W_V are frozen. Twenty banks each train five specialists on 512 sequences of lengths $4, 8, 16$, using 160 optimizer steps. Geometry, selection, and endpoint pools contain 16, 48, and 96 independent sequences per task. The data-generating teachers use the same attention architecture. Panel C tests a different nonlinearity and dependence on preceding inputs; it remains a controlled sequence classifier. Appendix G specifies its complete construction.

5.2 Candidate support, rank, and data access

Let $F = v_{\text{fac}}(\mathbf{1}/5)$ and $P = \mathrm{Proj}_{16}(\bar{u})$, $\bar{u} = \sum_i u_i/5$. The shared twenty-four-candidate library contains four scales $\gamma \in \{0.5, 1, 2, 5\}$ of each ray, two TIES densities in each coordinate space, factor and product DARE, factor and product Fisher, and eight KnOTS–TIES–SVD16 settings. The KnOTS grid crosses densities $0.2, 0.5$ with the four scales. Corrected unit arithmetic is already $5F$; $25F$ and the untruncated exact sum $5\bar{u}$ are separate stress controls rather than members of the rank-matched library.

KnOTS aligns horizontal concatenations of realized products in a shared left-singular basis before TIES merging (Stoica et al., 2025). The final SVD16 is a deployment adaptation. Panel B also evaluates the eight corresponding reconstructions without truncation, using the same selection budget. Native-rank variants remain outside the rank-matched library. Appendix C.2 specifies the support threshold, mask conventions, scale placement, and Fisher coordinates. Candidate construction grids are fixed before selection, and duplicate realized updates do not increase the candidate count.

Table 1: Pythia-1.4B deployment mismatch. Norms use stacked realized layer updates; $e = v - U\alpha$. All operators in this table act on factors. Independently rounded norms need not reproduce the separately rounded ratio.

Candidate	S_α	$\|U\alpha\|$	$\|v\|$	$\|e\|$	$\|e\| / \|U\alpha\|$
Factor average	1	1.92	1.83	0.45	0.24
Factor Fisher	1	1.92	3.07	2.07	1.08
Factor DARE	1	1.92	3.76	3.32	1.73
Factor TIES	1	1.92	4.16	3.39	1.77
Unit factor arithmetic	5	9.59	45.85	36.60	3.82

5.3 Estimation and decisions

For bank b, task t, and candidate c, scores use the selection sample, while $D_{b,t,c}$ uses the independent endpoint sample. A selector minimizes its maximum task score; its regret is

$$R_b(q) = \max_t D_{b,t,\widehat{c}_b(q)} - \min_c \max_t D_{b,t,c}. \tag{19}$$

Within-task Spearman correlation ranks candidates before averaging tasks within a bank. Top-1 measures selection of an endpoint minimizer, with uniform credit for exact ties. Wilson intervals have a binomial interpretation for binary bank-level hits; that interpretation does not extend to fractional tie credit. Correlation and regret intervals use $10{,}000$ paired bootstrap resamples of complete bank tuples. Top-1 uses Wilson score intervals. Panels are analyzed separately, conditional on their fixed bases and teachers, rather than pooled into an architecture-population interval.

Native accuracy is a co-primary endpoint, not a consequence inferred from KL. Let $S_{b,t,c}$ and $S^{\mathrm{sp}}_{b,t}$ be candidate and specialist native accuracies, and define $r_{b,t,c} = S_{b,t,c}/S^{\mathrm{sp}}_{b,t}$ for positive denominators. We report absolute accuracy, mean retention, and worst-task retention. Panel C also selects

$$\widehat{c}^{\,\mathrm{nat}}_b \in \arg\max_c \frac{1}{5}\sum_t \widehat{r}^{\,\mathrm{sel}}_{b,t,c}. \tag{20}$$

This objective control uses selection labels; label-free preservation scores do not. Neither procedure observes endpoint labels. Fixed full-scale product Fisher provides a no-selection comparator, while an endpoint-label native oracle measures the ceiling of the fixed candidate library.

For sequence contributions Q_z and D_z, the calibration SE is the sample SD of $Q_z - D_z$ divided by $\sqrt{n}$, thereby retaining the observed covariance. Panel A's estimator experiment uses the factor average direction, four fixed geometry sequences per task, and 256 probe replicates at counts $8, 32, 128$. All estimators share positions; the sampled-score estimators also share labels. The reference is the full-position marginalized Fisher for that sequence, not endpoint KL. Appendix A.4 separates sequence, position, and label uncertainty.

6 Results

6.1 Deployment correction and compression answer different questions

The Pythia operators deploy materially different products from their nominal references (Table 1). Unit arithmetic has norm 45.85 against 9.59, while four candidates have residual norms greater than their reference norm. Near-shared right factors explain the approximate fivefold product-reference discrepancy; the exact 25-fold relation between factor averaging and unit arithmetic instead follows from degree-two homogeneity.

The local quadratic also changes calibration direction across tasks. Every ordinary MBPP candidate has q/D between 3.443 and 5.603, whereas every corresponding GSM8K, Alpaca, and HH-RLHF candidate is underpredicted. The minimum MBPP additive overprediction is 0.977, substantially larger than its smallest candidate KL gap, 0.002. We therefore emphasize magnitude error rather than the fragile four-candidate rank reversal. Removing unit arithmetic lowers the four-task mean within-task correlation from 0.673 to 0.343. The full geometry-context local-score and behavioral-KL matrix appears in Appendix Table 6.

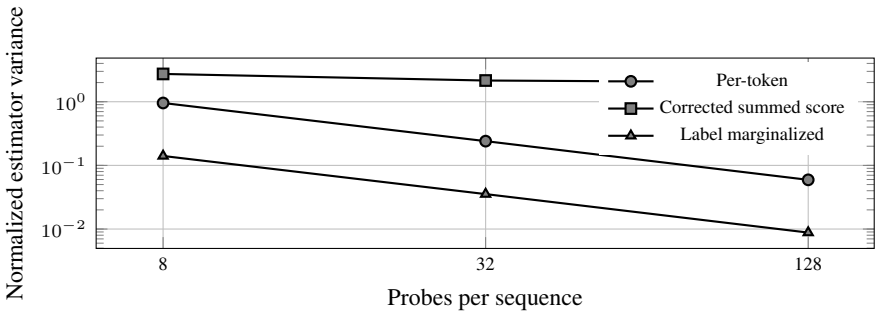


Figure 1: Panel A estimator comparison: 400 fixed sequence–direction pairs across twenty banks, with 256 replicates per pair. Variance is normalized by the squared exact finite-sequence Fisher target. Positions are shared across all estimators and labels across the two sampled-score estimators.

Panel A directly evaluates the correction controls. Mean-task KL drops from 49.386 for $25F$ to 7.137 for corrected arithmetic $5F$, but $5F$ remains a large-displacement stress point rather than a new useful direction. At the same nominal product sum, rank-16 truncation increases mean-task KL over the exact product by 0.0921 $[0.0283, 0.1604]$. Enumerating all specialist subsets gives deployed-to-reference ratios 1.790, 2.579, 3.365, and 4.150 for $K = 2, 3, 4, 5$; the factor-ray multiplier is nevertheless exactly K^2. Appendix Table 8 reports paired stress intervals and distinguishes across-bank from within-bank subset variability.

Removing the final rank truncation does not explain KnOTS's classifier result. In panel B, validation-selected KnOTS–TIES–SVD16 has regret 0.6740 $[0.5861, 0.7693]$ against the rank-matched oracle; the native-rank version has 0.6804 $[0.6026, 0.7649]$. Holding the selected density and scale fixed, native-minus-truncated worst-task KL is 0.0146 $[-0.0140, 0.0458]$. Hence this panel does not show that truncation causes the baseline's poor endpoint. It remains a comparison on this classifier and teacher family, not a refutation of KnOTS on its published vision or language benchmarks.

6.2 Estimator validation does not imply endpoint calibration

The trained-model score experiment confirms the variance distinction (Figure 1). Increasing probes from 8 to 128 reduces per-token normalized variance from 0.9556 to 0.0593. The corrected summed score falls only from 2.7334 to 2.0193, consistent with its cross-probe plateau. Label marginalization reduces the corresponding values from 0.1406 to 0.0088. All relative-bias intervals include zero; the complete intervals are in Appendix Table 10.

Panel C reproduces the variance separation on attention-model scores: at 128 probes, normalized variance is 0.0822 for per-token estimation, 2.0766 for corrected summation, and 0.0073 for marginalization. Appendix H gives the complete probe sweep. Paired sequence SEs are computed from the actual $Q_z - D_z$ contributions; neither panel supplies a fixed correlation to this calculation. These estimator comparisons do not remove finite-displacement error.

6.3 Output geometry changes selection and calibration

In panel A (Appendix Table 12), centered output drift and taskwise norm coincide on all twenty decisions, with mean regret 0.3606, versus 0.0045 for both the per-token and marginalized quadratics. The replication preserves the failure but identifies its candidate support (Table 2). Norm selects P in all twenty banks; uniform output drift selects P in nineteen and F in one. The endpoint optimum is instead $0.5P$ in eleven banks, $0.5F$ in eight, and factor Fisher in one. The failure is therefore not selection of the extreme $5\times$ ray endpoint. It is selection of an ordinary-scale merge when a smaller displacement better preserves the specialists' predictive distributions.

Changing only the output-coordinate covariance resolves most of this selection error. Specialist-weighted drift chooses $0.5F$ in every bank and reduces mean regret by 0.3743 $[0.3041, 0.4447]$ relative to uniform drift, a 94.0% reduction in their mean regrets. Its paired regret difference from the marginalized Fisher is 0.00003 $[-0.0155, 0.0174]$, so this panel does not separate those two decision

Table 2: Independent implementation panel B: twenty banks, a new base and teacher collection, and the same twenty-four rank-matched candidates. Correlation and regret intervals resample paired banks; top-1 intervals are Wilson score intervals. The direct KL row uses selection contexts, not endpoint contexts.

Panel B diagnostic	Within-task ρ [95% CI]	Top-1 [95% CI]	Regret [95% CI]
Taskwise norm	.965 [.963, .967]	.00 [.00, .161]	.4019 [.3394, .4653]
Marginalized Fisher	.990 [.989, .991]	.50 [.299, .701]	.0237 [.0114, .0373]
Uniform output drift	.964 [.962, .967]	.00 [.00, .161]	.3981 [.3307, .4643]
Weighted output drift	.987 [.986, .988]	.40 [.219, .613]	.0238 [.0120, .0377]
Selection-split KL	.997 [.997, .997]	.80 [.584, .919]	.0077 [.0001, .0207]

rules. Direct selection-split KL has lower mean regret but still misses four endpoint minima. Finite selection samples remain consequential even when score and target are identical functionals.

The scale comparison makes the metric mechanism explicit (Appendix Table 13). Uniform drift prefers P to $0.5F$, while the specialist-weighted score reverses their mean worst-task ordering. Both use the same candidate logits. Thus, in this comparison, an extra forward pass is not the missing ingredient; the difference is how the displacement is weighted and centered in output space.

The weighted score substantially reduces the selection error without becoming a calibrated KL estimate. At its selected candidate's worst endpoint task, panel B has mean $D = 1.749$, local $q = 1.065$, and finite-logit $q^{\mathrm{wlog}} = 0.655$. The network and logit remainders are -0.410 and 1.094, respectively. Appendix Table 14 gives their paired intervals. A covariance-interpolation control in Appendix F traces the change from full-scale to half-scale choices using the same logits.

6.4 Shrinkage exposes a conflict between preservation and utility

The panel-B result favors doing less, not recovering most specialist competence. Its base and specialists achieve 30.88% and 51.93% accuracy; the KL selector reaches only 33.77%. The gain is 2.89 of the 21.05 percentage points separating those reference means. Full-scale product Fisher instead has the highest point estimate of worst-task retention among the evaluated policies, 53.01%, and mean retention 65.34%, compared with 52.36% and 65.00% for KL selection. A paired contrast is needed to establish superiority; marginal interval overlap alone does not test that difference (Appendix Table 16). The implication nevertheless changes the objective: low preservation regret alone does not identify the most useful merge.

This tension follows from the target rather than a contradiction in the estimator. Along a ray γv, the local target is $\frac{1}{2}(\gamma v - u_t)^\top \mathcal{F}_t(\gamma v - u_t)$. When $v^\top \mathcal{F}_t v > 0$, its unconstrained minimizer is

$$\gamma_t^* = \frac{v^\top \mathcal{F}_t u_t}{v^\top \mathcal{F}_t v}. \tag{21}$$

The minimax compromise can therefore prefer attenuation; no native-label term rewards a correct decision that differs from the specialist distribution. Half-scale selection is not evidence of useful transfer merely because it reduces KL. We retain it as a valid preservation outcome and test native utility alongside it, with the labeled selector in Equation (20) separating a change of objective from a change of output metric.

6.5 A second architecture preserves the metric effect, not an accuracy gain

The causal attention experiment reproduces the preservation advantage without reproducing panel B's universal half-scale choice (Table 3). Uniform drift selects P in twelve banks and F in eight. Specialist weighting selects factor Fisher in fourteen, a half-scale ray in five, and F in one. Its paired regret reduction is $0.0617\,[0.0308, 0.0946]$ relative to uniform drift, whereas its paired difference from marginalized Fisher is $-0.0026\,[-0.0166, 0.0067]$. The latter comparison does not distinguish the two decision rules. Direct selection-split KL also has residual regret because its empirical minimum need not match the independent endpoint minimum.

Table 3: Causal attention, twenty independently trained banks and the same twenty-four candidate constructions. Correlation and regret use paired bank-bootstrap intervals; top-1 uses Wilson intervals. Results are conditional on panel C's fixed teachers, base, and source generator, separately from panels A and B.

Panel C diagnostic	Within-task ρ [95% CI]	Top-1 [95% CI]	KL regret [95% CI]
Taskwise norm	.955 [.952, .959]	.05 [.009, .236]	.0890 [.0650, .1154]
Marginalized Fisher	.981 [.978, .983]	.45 [.258, .658]	.0228 [.0105, .0380]
Uniform output drift	.948 [.944, .952]	.05 [.009, .236]	.0820 [.0551, .1112]
Weighted output drift	.978 [.976, .981]	.45 [.258, .658]	.0202 [.0105, .0318]
Selection-split KL	.991 [.990, .992]	.50 [.299, .701]	.0251 [.0070, .0524]

Table 4: Panel C co-primary endpoints. Accuracy and retention are percentages; KL regret is in nats per position under sequence-balanced weighting. Native selection uses selection labels, while the final row uses endpoint labels solely to bound the library's attainable mean retention. Each bracket resamples complete bank tuples.

Panel C policy	Accuracy [95% CI]	Mean retention [95% CI]	KL regret [95% CI]
Common base	48.22 [47.81, 48.63]	—	—
Specialists	76.01 [75.67, 76.32]	—	—
Uniform output drift	53.01 [52.53, 53.54]	69.66 [68.98, 70.38]	.0820 [.0551, .1112]
Weighted output drift	52.96 [52.32, 53.60]	69.60 [68.76, 70.47]	.0202 [.0105, .0318]
Selection-split KL	53.17 [52.60, 53.76]	69.88 [69.10, 70.70]	.0251 [.0070, .0524]
Native-retention selection	53.92 [53.42, 54.40]	70.86 [70.27, 71.45]	.0400 [.0214, .0610]
Product Fisher, fixed	54.04 [53.54, 54.52]	71.01 [70.40, 71.61]	.0637 [.0393, .0905]
Native-retention oracle	54.24 [53.80, 54.67]	71.27 [70.72, 71.81]	.0507 [.0297, .0754]

Native accuracy does not follow the KL improvement (Table 4). Weighted-minus-uniform accuracy is $-0.058\,[-0.615, 0.505]$ points. Native-retention selection instead improves accuracy over KL selection by $0.743\,[0.243, 1.316]$ points.

Its paired KL-regret increase is $0.0149\,[-0.0061, 0.0367]$. The objectives therefore differ: native-retention selection improves accuracy, while the corresponding KL-regret difference is not distinguished from zero by its paired interval.

Fixed full-scale product Fisher has the highest reported mean-accuracy point estimate among the evaluated label-free selection policies and fixed comparators in both panel B (33.93%) and panel C (54.04%); in panel C it also exceeds the native-label selector's 53.92%. The retrospective native oracle reaches only 71.27% mean retention, so the fixed candidate library still limits the native retention attainable by all evaluated selectors.

7 Discussion, Conclusion, and Limitations

LoRA merge diagnostics depend on the realized update, the approximation to its behavior, and the endpoint used for selection. Factor correction, rank truncation, Fisher estimation, and output geometry address distinct errors; improving one does not resolve the others. Across two controlled architectures, specialist-weighted output drift lowers preservation regret relative to uniform drift. The classifier decomposition nevertheless shows substantial KL miscalibration. Its accuracy effect in causal attention is not distinguished from zero, whereas selection for native retention improves accuracy over KL selection. Attenuation can therefore preserve specialist distributions without recovering specialist competence. The bank intervals describe variation within each controlled panel and do not quantify uncertainty across pretrained architectures. These findings are conditional on fixed teachers, bases, and a twenty-four-candidate library; its native oracle recovers only 21.60% of the base-to-specialist gain in panel C. The pretrained evidence is one Pythia bank with near-floor code and mathematics performance, and HumanEval geometry reuses training examples. The ten-bank Pythia study in Appendix I is an unexecuted transfer protocol. Broader checkpoints, stronger specialists, and richer candidate libraries are needed to establish transfer. Within the evaluated settings, diagnostics should reconstruct deployed products and report native utility alongside specialist-preservation endpoints.

Reproducibility Statement

The appendices specify the Pythia recipe, controlled generators, merge operators, candidate library, sampling streams, statistical units, and paired uncertainty calculations. They also report complete bank decisions, sequence-level second moments, estimator controls, rank interventions, and negative native outcomes. The Pythia independent-bank appendix is explicitly identified as a transfer protocol and excluded from completed empirical claims.

AI Use Statement

Generative AI assisted with mathematical exposition, analysis of controlled-model implementations, literature comparison, and manuscript revision. The authors reviewed all AI-assisted content and independently checked the mathematical arguments and implementation-related conclusions. The authors remain responsible for the claims, proofs, protocols, and reported numerical results.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. In *arXiv preprint arXiv:2204.05862*, 2022. URL https://arxiv.org/abs/2204.05862.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2304.01373.

Oliver Bolton, Aakanksha, Arash Ahmadian, Sara Hooker, Marzieh Fadaee, and Beyza Ermis. SimMerge: Learning to Select Merge Operators from Similarity Signals. *arXiv preprint arXiv:2601.09473*, 2026. URL https://arxiv.org/abs/2601.09473.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Rei Nakano, et al. Training verifiers to solve math word problems. In *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Marvin F. da Silva, Mohammed Adnan, Felix Dangel, and Sageev Oore. Generalizing the geometry of model merging through Fréchet averages. In *arXiv preprint arXiv:2604.27155*, 2026. URL https://arxiv.org/abs/2604.27155.

Seyed Arshan Dalili and Mehrdad Mahdavi. Model Merging via Multi-Teacher Knowledge Distillation. *arXiv preprint arXiv:2512.21288*, 2025. URL https://arxiv.org/abs/2512.21288.

Bethan Evans, Benjamin Etheridge, Stephen Roberts, and Jared Tanner. Model merging by output-space projection. *arXiv preprint arXiv:2605.29101*, 2026. URL https://arxiv.org/abs/2605.29101.

Zhengbao He, Ruiqi Ding, Zhehao Huang, Ruikai Yang, Tao Li, and Xiaolin Huang. Compress then merge: From multiple LoRAs into one low-rank adapter. *arXiv preprint arXiv:2606.03723*, 2026. URL https://arxiv.org/abs/2606.03723. Accepted to ICML 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Juseok Jeon, Ramy E. Ali, Doyun Kwon, Myungbeom Her, Jinhwi Kim, and Jinhyun So. FedPA-LoRA: Product-aligned framework for mitigating aggregation and initialization errors in heterogeneous federated LoRA, 2026. URL https://arxiv.org/abs/2608.15381.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FCnohuR6AnM.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical Fisher approximation for natural gradient descent, 2019. URL https://arxiv.org/abs/1905.12558.

Sourab Mangrulkar and Sayak Paul. PEFT welcomes new merging methods. Hugging Face technical documentation, February 2024. URL https://huggingface.co/blog/peft_merging. Published February 19, 2024.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://jmlr.org/papers/v21/17-678.html.

Michael S. Matena and Colin A. Raffel. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2111.09832.

Alexander Soen and Ke Sun. On the variance of the Fisher information for deep learning, 2021. URL https://arxiv.org/abs/2107.04205.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the KnOTS. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.19735.

Yixuan Tang and Yi Yang. Crowded in B-space: Calibrating shared directions for LoRA merging. *arXiv preprint arXiv:2604.16826*, 2026. URL https://arxiv.org/abs/2604.16826.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following LLaMA model. *Project report*, 2023. URL https://crfm.stanford.edu/2023/03/13/alpaca.html.

Jiayu Wang, Zuojun Ye, and Wenpeng Yin. Functionality-oriented LLM merging on the Fisher–Rao manifold. *arXiv preprint arXiv:2603.04972*, 2026. URL https://arxiv.org/abs/2603.04972.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2203.05482.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.01708.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nHkXwEi1s4.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2311.03099.

A Derivations and Statistical Units

A.1 Factor covariance, scalar-equivalent correction, and recompression

Expanding the factor operator and subtracting its product reference gives

$$e_{\text{fac}} = s\sum_i(\alpha_i^2-\alpha_i)B_iA_i + s\sum_{i\neq j}\alpha_i\alpha_jB_iA_j. \tag{22}$$

Subtracting $S_\alpha w$ instead pairs the two off-diagonal terms:

$$v_{\text{fac}} - S_\alpha w = s\sum_{i<j}\alpha_i\alpha_j(B_iA_j + B_jA_i - B_iA_i - B_jA_j) = -s\sum_{i<j}\alpha_i\alpha_j(B_i-B_j)(A_i-A_j).$$

This proves Proposition 1, including signed coefficients. Division is valid only for $S_\alpha \neq 0$. For nonnegative coefficients, substitution of $\alpha_i = S_\alpha p_i$ gives

$$v_{\text{corr}} - w = -sS_\alpha\sum_i p_i(B_i-\bar{B})(A_i-\bar{A}).$$

Use the triangle inequality, $\|XY\|_F \le \|X\|_F\|Y\|_{\text{op}}$, and weighted Cauchy–Schwarz to obtain Equation (4). If either factor set is constant, every pairwise term vanishes. For signed coefficients, the corresponding bound is

$$\|v_{\text{corr}} - w\|_F \le \frac{|s|}{|S_\alpha|}\sum_{i<j}|\alpha_i\alpha_j|\,\|B_i-B_j\|_F\,\|A_i-A_j\|_{\text{op}}. \tag{23}$$

Each identity holds layerwise. Stacking layers combines squared Frobenius errors by summation. Independent gauge changes preserve the individual products but can change a factor merge, so its actual supplied factors remain part of the operator definition.

A common coefficient multiplier appears in both factors, proving degree-two homogeneity for positive, negative, and zero multipliers. Setting $\alpha_i = 1/K$ and multiplying all coefficients by K gives $v_{\text{fac}}(\mathbf{1}) = K^2 v_{\text{fac}}(\mathbf{1}/K)$. For $\alpha = \lambda\mathbf{1}$, division by $K\lambda$ then proves Equation (5). This comparison is distinct from approximate agreement with the product sum under near-common factors.

Exact product aggregation can be implemented without a dense SVD. Define

$$B_{\text{cat}} = [\alpha_1 s_1 B_1 \;\cdots\; \alpha_K s_K B_K], \qquad A_{\text{cat}} = [A_1^\top \;\cdots\; A_K^\top]^\top.$$

Their product is w. Thin QR factorizations $B_{\text{cat}} = Q_B R_B$ and $A_{\text{cat}}^\top = Q_A R_A$ yield $w = Q_B(R_B R_A^\top)Q_A^\top$. An SVD of the small core gives the nonzero singular values and their ambient directions. Orthogonality proves the tail formula in Equation (6). If a right factor with r rows is exactly shared, w already has rank at most r. Otherwise,

$$\|v_{\text{corr}} - \text{Proj}_r(w)\|_F \le \|v_{\text{corr}} - w\|_F + \left(\sum_{j>r}\sigma_j(w)^2\right)^{1/2}.$$

For positive output scale s_{out} and retained product $L\Sigma R^\top$, serialize $B_{\text{out}} = L\Sigma^{1/2}/\sqrt{s_{\text{out}}}$ and $A_{\text{out}} = \Sigma^{1/2}R^\top/\sqrt{s_{\text{out}}}$. Reconstructing $s_{\text{out}}B_{\text{out}}A_{\text{out}}$ then returns the specified product, with no extra scaling.

A.2 KL in parameter space and finite-logit space

Assumption 4 (Parameter-path regularity)**.** *Along $u_t + ah$, $a \in [0,1]$, support is fixed, differentiation passes through the behavioral expectation, and the score has zero specialist expectation. Behavioral KL is twice continuously differentiable. For the cubic remainder, its Hessian satisfies $\left\|H_t^{\text{KL}}(a) - \mathcal{F}_t\right\|_{\text{op}} \le L_t^{\text{KL}} a\,\|h\|$ throughout the segment.*

Let $f(a) = \mathcal{D}_t(u_t + ah)$. The score identity gives $f(0) = f'(0) = 0$, and the information identity gives $f''(0) = h^\top\mathcal{F}_t h$. Integrating twice proves Equation (8). Subtracting $f''(0)/2$ and applying the Hessian condition gives

$$|\mathcal{D}_t(v) - q_t(v)| \le L_t^{\text{KL}}\|h\|^3\int_0^1 a(1-a)\,da = L_t^{\text{KL}}\|h\|^3/6. \tag{24}$$

Substitute $h = Ub_t + e$ and expand to obtain Equation (9). For a supervised risk with gradient g_t, initial Hessian within ξ_t of the Fisher, and Hessian Lipschitz constant L_t^R along the same segment, the corresponding bound is

$$|\mathcal{R}_t(v) - \mathcal{R}_t(u_t) - g_t^\top h - q_t(v)| \le \xi_t \|h\|^2 /2 + L_t^R \|h\|^3 /6.$$

This risk statement does not identify the model Fisher with the gold-label gradient covariance. The parameter path is in ambient update coordinates and need not be rank constrained between its endpoints.

For the finite-logit identities, no parameter-path approximation is needed. At a fixed input, set $p = \mathrm{softmax}(\ell_t)$, $\delta = \ell_v - \ell_t$, and $\psi(a) = \log \sum_y p_y e^{a\delta_y}$. Softmax normalization gives

$$\log \frac{p_y}{p_{v,y}} = -\delta_y + \psi(1), \quad \psi'(a) = \mathbb{E}_{p_a}\delta, \quad \psi''(a) = \mathrm{Var}_{p_a}(\delta).$$

Since $\psi(0) = 0$, the same integral identity gives $D_x = \psi(1) - \psi'(0) = \int_0^1 (1-a)\psi''(a)\, da$. The second derivative is $\delta^\top G(p_a)\delta$. Subtracting half its value at zero proves Equation (16). Finally, inserting $\delta = j_h + r_\ell$ into $\delta^\top G(p)\delta/2$ proves Equation (15). These expressions are invariant to a common candidate-logit offset because $G(p)\mathbf{1} = 0$ and the offset cancels from $\psi(1) - \psi'(0)$.

Uniform centering and probability weighting should not be conflated. With $\bar\delta_U = V^{-1}\sum_y \delta_y$,

$$\sum_y p_y(\delta_y - \bar\delta_U)^2 = \mathrm{Var}_p(\delta) + (\mathbb{E}_p\delta - \bar\delta_U)^2.$$

The weighted diagnostic in Equation (13) uses the categorical variance, including its weighted mean subtraction. Its intervention is replacement of the entire coordinate covariance, not weighting an otherwise unchanged squared norm.

A.3 Score aggregation, importance sampling, and variance

Conditional independence and zero score means imply

$$\mathbb{E}[\widetilde{ss}^\top \mid z] = T_z^{-2} \sum_{\tau,\sigma} \mathbb{E}[s_\tau s_\sigma^\top \mid z] = T_z^{-2} \sum_\tau \mathcal{F}_\tau(z).$$

Multiplying by T_z before the sequence expectation gives the sequence-balanced Fisher. With $A_z = T_z^{-1}\sum_\tau \mathcal{F}_\tau \succeq 0$, the inequalities $T_{\min}/T_z \le 1 \le T_{\max}/T_z$ prove the PSD sandwich. For uniformly sampled positions, the same argument replaces T_z by the aggregated probe count m_z, including sampling with replacement provided each repeated occurrence receives an independent label.

For a positive position distribution π_z fixed before label draws, define

$$g_z = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \frac{s_t(x_{\tau_j}, y_j)}{\sqrt{T_z \pi_z(\tau_j)}}. \tag{25}$$

Each summand has zero mean and second moment $\sum_\tau \pi_z(\tau)\mathcal{F}_\tau(z)/(T_z\pi_z(\tau)) = T_z^{-1}\sum_\tau \mathcal{F}_\tau(z)$. The factor $m^{-1/2}$ therefore gives the same second moment for the sum. Empty sequences are excluded before sampling, and every eligible position must have positive probability. Differentiation of the weighted log-probability sum produces g_z in one reverse-mode pass; probabilities defining the sampling scheme are not differentiated.

To prove Proposition 3, write $Z = Y_{\mathrm{sum}} - Y_{\mathrm{tok}} = 2m^{-1}\sum_{i<j} a_i a_j$. Every covariance between a squared term and a pair product contains an independent centered factor. Two distinct pair products likewise have zero covariance. Thus $\mathrm{Cov}(Y_{\mathrm{tok}}, Z) = 0$ and

$$\mathrm{Var}(Z) = 4m^{-2} \sum_{i<j} \mathbb{E}[a_i^2]\mathbb{E}[a_j^2].$$

Adding $\mathrm{Var}(Y_{\mathrm{tok}})$ proves the variance gap. For identically distributed draws, the gap is $2(m-1)v^2/m$ and $\mathrm{Var}(Y_{\mathrm{tok}}) = (\mu_4 - v^2)/m$. Conditional expectations coincide sequence by sequence, so averaging over sequences preserves the nonnegative variance difference under finite moments.

A.4 MARGINALIZATION AND PAIRED UNCERTAINTY

The directional score for label y is $a_y = j_{h,y} - \sum_k p_k j_{h,k}$. Therefore $\mathbb{E}_p a_y = 0$ and $\mathbb{E}_p a_y^2 = \sum_y p_y j_{h,y}^2 - (\sum_y p_y j_{h,y})^2$, proving Equation (12). For a sampled quadratic contribution Z and sampled positions V_z, the law of total variance separates

$$\mathrm{Var}(Z) = \mathrm{Var}_z(\mathbb{E}[Z \mid z]) + \mathbb{E}_z \mathrm{Var}_{V_z}(\mathbb{E}[Z \mid z, V_z] \mid z) + \mathbb{E}_{z,V_z} \mathrm{Var}(Z \mid z, V_z). \tag{26}$$

Marginalization removes the last, label-sampling term. Enumerating positions removes the middle term; neither operation increases the number of independent source sequences.

For n independent sequence contributions and $\Delta_z = Q_z - D_z$, the paired SE is

$$\widehat{\mathrm{SE}}(\bar{\Delta}) = \sqrt{\frac{\sum_z (\Delta_z - \bar{\Delta})^2}{n(n-1)}}. \tag{27}$$

Equivalently, with centered sums S_{QQ}, S_{DD}, S_{QD}, its square is $(S_{QQ} + S_{DD} - 2S_{QD})/[n(n-1)]$. The same calculation applies to a difference between two candidates. These SEs estimate sequence-sampling variation; bank intervals instead resample independently trained banks with all candidates and tasks retained. Finite-sample Gaussian coverage is not assumed for the per-sequence SEs.

For a binary endpoint-minimizer hit count k among n banks, the Wilson interval is

$$\frac{\widehat{p} + z^2/(2n) \;\pm\; z\sqrt{\widehat{p}(1-\widehat{p})/n + z^2/(4n^2)}}{1 + z^2/n}, \quad \widehat{p} = k/n,\ z = 1.959964. \tag{28}$$

This avoids the degenerate $[0, 0]$ percentile bootstrap for an all-zero hit vector. Bank means use $10{,}000$ bootstrap draws, seed 820571, and percentile endpoints at 2.5% and 97.5%. Paired differences are formed before resampling. All such intervals are conditional on the panel's fixed task distributions and common base.

B PYTHIA RECIPE AND ENDPOINT MATRIX

The checkpoint is `EleutherAI/pythia-1.4b`. LoRA rank is 16, parameter $\alpha_{\mathrm{LoRA}} = 32$, and scale two. Only attention `query_key_value` and `dense` are adapted, with no trainable biases or dropout. The backbone is float16. Training uses one full-sequence-loss epoch, batches of two, four-step gradient accumulation, AdamW learning rate 2×10^{-4}, and weight decay 10^{-3}. An incomplete final accumulation group receives no extra optimizer step. Resetting the seed before each specialist initialization shares its initial right factors.

GSM8K uses question/answer formatting; HumanEval concatenates prompt and canonical solution; MBPP concatenates task text and code; Alpaca uses instruction/response headers; HH-RLHF uses the chosen conversation. The entire formatted sequence is trained, with a 256-token truncation and padding excluded. Splits are shuffled with seed zero. The sources are `gsm8k` (`main`), `openai_humaneval`, `google-research-datasets/mbpp` (`full`), `tatsu-lab/alpaca`, and `Anthropic/hh-rlhf`.

Table 5: Pythia source windows after seed-zero shuffling. The geometry distribution defines the local diagnostic measurement. HumanEval geometry is in-sample and excluded from the main four-task summary. MBPP uses 45 examples in each validation half; source counts cap the requested allocations.

Task	Training	Geometry	Selection	Endpoint
GSM8K	train $[0, 500)$	test $[200, 712)$	test $[100, 200)$	test $[0, 100)$
HumanEval	test $[0, 64)$	test $[0, 64)$	test $[114, 164)$	test $[64, 114)$
MBPP	train up to 500	test up to 512	validation second half	validation first half
Alpaca	train $[0, 500)$	train $[700, 1212)$	train $[600, 700)$	train $[500, 600)$
HH-RLHF	train $[0, 500)$	test $[200, 712)$	test $[100, 200)$	test $[0, 100)$

Factor averaging and arithmetic separately average or sum both factors before reconstructing sBA. Factor TIES retains the largest-magnitude 20% of each tensor, including cutoff ties, elects the sign by retained-sign frequency, and averages agreeing nonzero coordinates. DARE drops coordinates with

probability 0.7, rescales survivors by $1/0.3$, then averages factors. The fixed mask seed is part of the candidate. Fisher averages factor coordinates using their per-token model-Fisher weights and denominator stabilizer 10^{-8}. Its factor-coordinate merging metric is distinct from the ambient Fisher used to score the deployed displacement.

Table 6: Complete Pythia local-score and behavioral-KL matrix. The local score uses 128 sequence-balanced score draws per task. Entries are geometry-context means, not independent-bank averages. HumanEval remains an in-sample control.

Task	Quantity	Average	Arithmetic	TIES	DARE	Fisher
GSM8K	q	.540	118.827	.434	.569	.309
	D	.725	6.222	.634	.724	.354
HumanEval	q	.061	25.819	.080	.060	.085
	D	.039	7.644	.049	.037	.057
MBPP	q	1.828	124.662	2.017	1.812	1.377
	D	.398	7.200	.360	.405	.400
Alpaca	q	.201	44.094	.192	.201	.169
	D	.544	5.809	.467	.559	.505
HH-RLHF	q	.145	46.736	.117	.153	.129
	D	.230	5.791	.207	.232	.232

For the four ordinary candidates, MBPP ratios range from $1.377/.400 = 3.4425$ to $2.017/.360 = 5.6028$, displayed as 3.443 and 5.603 under round-half-up. The remaining KL order has gaps 0.038, 0.002, and 0.005. Display precision and stochastic uncertainty are distinct; the smallest half-gap is not an estimate of Monte Carlo SE. The main text consequently does not base its conclusion on this small ordering.

For completeness, the taskwise local-score correlations are $(.900, 1.000, .100, .872, .821)$ with arithmetic and $(.800, 1.000, -.800, .738, .632)$ without it. Excluding HumanEval gives means .673 and .343. On GSM8K, factor-average KL along path fractions $(.1, .3, .5, .7, .9, 1)$ is $(.008, .063, .161, .319, .564, .725)$; Fisher gives $(.005, .036, .094, .173, .281, .354)$. Their local scores are the endpoint quadratics .540 and .309 multiplied by the squared path fraction.

B.1 Centroid and repeated-randomization controls

The RMS norm ranking follows from

$$K^{-1} \sum_t \|v - u_t\|^2 = \|v - \bar{u}\|^2 + K^{-1} \sum_t \|u_t - \bar{u}\|^2 . \tag{29}$$

The cross-term vanishes after summing $u_t - \bar{u}$. For a uniform reference $w = S\bar{u}$, polarization recovers

$$\|v - \bar{u}\|^2 = \|v\|^2 + \|w\|^2 / S^2 - (\|v\|^2 + \|w\|^2 - \|v - w\|^2)/S. \tag{30}$$

Applying Equation 30 to Table 1 gives centroid distances $.45, 2.07, 3.32, 3.39, 43.99$ for averaging, Fisher, DARE, TIES, and arithmetic, respectively. The RMS rule selects averaging with worst-task KL .725, versus the Fisher minimum .505, hence regret .220. This does not recover the individual taskwise norms or their maximum.

The separate thirty-randomization Pythia summaries give pooled correlation .661 and regret .0118 for the deployed quadratic, .653 and .0158 for its supervised-gradient variant, and fixed-policy regrets .0044, .0669, and .0861 for Fisher, TIES, and averaging. Their dependence follows the original training-and-geometry randomization scheme; they are not independent-task or independent-architecture intervals. Only these target-defined summaries enter the comparison. Pythia native capability is near floor: MBPP pass@1 is zero, the highest GSM8K exact match is .040, and HumanEval pass@1 is at most .016.

C Controlled-Panel Construction

C.1 Panel A and the independently implemented panel B

Both panels use Equation (17). A common base has Gaussian entries with variance $1/32$ in W_{01} and $1/48$ in W_{02}. Each task teacher has weights $W_{0\ell} + 2B^*_{t\ell}A^*_{t\ell}$: right-factor entries have inverse-input-dimension variance, and left-factor entries have variance $.30^2/16$. Teacher labels are maximizing logits. Task means have distribution $\mathcal{N}(0, .60^2 I)$ and coordinate scales are $d_{tj} = \exp(a_{tj})$, $a_{tj} \sim \mathrm{Unif}[-.5, .5]$.

Panel A uses primary seed $M = 672190$ for its base, teacher factors, means, and scales. Panel B is a separate implementation with $M = 672191$. To specify panel B's draw order completely, PCG64 first draws W_{01} then W_{02}; for each task in order, it draws $A^*_{t1}, B^*_{t1}, A^*_{t2}, B^*_{t2}$; it next draws the full 5×32 mean array and finally the scale array. Each panel holds its realized base, teachers, and task distributions fixed across its twenty banks.

For zero-indexed bank b and task t, the training-input seed is $M + 200000 + 100b + t$. Each specialist receives 512 independent Gaussian inputs with its task mean and diagonal covariance. Adapter initialization uses seed $M + 300000 + b$, drawing the first and second layer right factors in order; those right factors are shared within a bank and left factors start at zero. Base weights stay frozen. LoRA rank is 16 and scale two. AdamW uses learning rate .004, weight decay .001, betas $(.9, .999)$, and epsilon 10^{-8}, with 160 steps and 64 training indices sampled with replacement per specialist. The minibatch seed is $M + 400000$, and each step draws a 100×64 index array. Trainable blocks and optimizer moments are disjoint; the joint loss is their mean. There is no dropout, clipping, or schedule.

A source sequence has $T \sim \mathrm{Unif}\{8, 16, 32\}$, shared latent $z \sim \mathcal{N}(0, I)$, and positions

$$x_\tau = \mu_t + \mathrm{diag}(d_t)(.5z + \sqrt{.75}\epsilon_\tau), \qquad \epsilon_\tau \overset{\text{iid}}{\sim} \mathcal{N}(0, I). \tag{31}$$

Geometry, selection, and endpoint seeds are $M + 100000 + 1000b + 100t + j$, with $j = 1, 2, 3$. Their sequence counts are 16, 48, and 96. Positions are averaged before sequences. Shared within-sequence latents do not count as additional independent observations. All methods use identical inputs, masks, and teacher labels within a bank and task.

Both implementations use PyTorch 2.10.0 on CPU with four threads and float32 training. Panel A uses float32 forwards and float64 SVDs and scalar aggregation. Panel B evaluates analytic derivatives and forward probabilities in float64 after training to isolate the metric comparisons from softmax roundoff. Input streams and trained adapter factors are first represented in float32, then promoted for these calculations; the initial Gaussian base draws are retained in float64 for panel B evaluation. The panel-B derivative is

$$j_h = 2h_2 \tanh(W_1 x) + 2W_2 \,\mathrm{diag}(1 - \tanh^2(W_1 x))h_1 x.$$

The two panels have separate statistical summaries; their bank outcomes are not pooled into one forty-bank interval.

C.2 Library definitions and native-rank alignment

The rank-matched library contains the following fixed entries.

The factor/product TIES controls retain the largest absolute coordinates, including cutoff ties; sign election sums retained signs and merging averages agreeing nonzero values. Panel B retains $\lceil \rho d \rceil$ of d coordinates before including cutoff ties. Zero elected signs produce zero. Density is applied separately to each layer or factor, unlike the aligned concatenated mask described below. DARE uses a fixed PCG64 mask, seed $M + 500000 + b$, and inverse-survival multiplier $1/.3$. Panel B draws each layer's A mask then B mask, followed by the first and second product masks. Product outputs are truncated only after aggregation.

For KnOTS, horizontally concatenate the five layer products and compute

$$[\Delta W_1 \ \cdots \ \Delta W_5] = L\Sigma[R_1 \ \cdots \ R_5], \qquad C_i = \Sigma R_i. \tag{32}$$

The shared-left-basis convention and singular-value support threshold 10^{-5} follow the source implementation (Stoica et al., 2025). Aligned coordinates from both layers are concatenated before each

Table 7: Twenty-four rank-matched candidates in each controlled panel. Corrected unit arithmetic equals $5F$ exactly. Native-rank KnOTS variants, the exact product sum, and $25F$ are external controls.

Family	Definition	Count
Factor ray	$\gamma F, \gamma \in \{.5, 1, 2, 5\}$.	4
Product ray	$\gamma \operatorname{Proj}_{16}(\bar{u})$ on the same grid.	4
Factor TIES	Density .2 or .5 separately on factor tensors.	2
Product TIES	Same densities on products, then SVD16.	2
Factor / product DARE	Drop .7, rescale survivors, then average in the specified coordinates.	2
Factor / product Fisher	Coordinate model-Fisher weighting in the respective space.	2
KnOTS–TIES–SVD16	Shared-left-SVD alignment; two densities crossed with four output scales.	8

specialist's magnitude mask. TIES elects the sign from retained values, fills zero consensus with the global majority sign, and averages agreeing entries. Panel B resolves a zero global majority as positive. Reconstruct each layer as $L\widehat{C}$ before scaling. SVD16 retains its sixteen leading singular values; the native control omits only this final map. Native rank is at most 32 per layer in the controlled architecture. Density and scale are selected on the same selection-split KL budget for both versions. A matched-rank contrast also freezes the density and scale selected for SVD16 before evaluating its native counterpart.

Factor and product Fisher merging use the model-label-integrated diagonal on geometry sequences in their own coordinates: $\theta_{\text{merge}} = (\sum_i f_i \odot \theta_i)/(\sum_i f_i + 10^{-8})$. Factor products are reconstructed after weighting; product-coordinate weighting is followed by SVD16. In panel B, with $a = \tanh(W_1 x)$ and $p = \operatorname{softmax}(2W_2 a)$, the ambient diagonal contributions are

$$F_{W_2,oj} = 4p_o(1-p_o)a_j^2, \quad F_{W_1,jk} = 4(1-a_j^2)^2 \operatorname{Var}_{Y\sim p}(W_{2,Yj})x_k^2.$$

Chain-rule transformations yield the factor diagonals; an ambient diagonal is not replaced by a factor diagonal. Sequence and position weighting is the same as for behavioral KL.

C.3 Estimator replicates and retained statistics

Panel A fixes the factor-average direction for the first four geometry sequences in each bank and task. Each unit has 256 replicates, using seed $M + 700000 + 1000b + 100t + i$. A replicate draws 128 positions with replacement and fresh labels, with the 8- and 32-probe arms using stream prefixes. Marginalization shares positions and integrates labels. Relative bias is $\bar{\widehat{q}}/q_{\text{ref}} - 1$ and normalized variance is the replicate sample variance divided by q_{ref}^2. Zero references are excluded from relative summaries. The 400 unit summaries are averaged within banks before computing intervals.

The panel-A sampled-score selectors use 32 positions and four label replicates per sequence; the marginalized selector uses all eligible positions. Panel B's metric comparisons all use exact categorical reductions and all eligible positions. No derivative-pass, latency, or memory saving is inferred from these probe comparisons. All endpoint records retain the sequence identity and paired local, KL, output, and native contributions. Panel B additionally retains $\mathbb{E}_p\delta$ and $\log \mathbb{E}_p e^{\delta}$ for the decomposition.

D Detailed Operator and Estimator Results

Panel A evaluates stress controls at their actual deployed endpoints, with the five task KLs averaged within each bank. The corrected-minus-unit contrast is $-42.249\,[-43.163, -41.303]$. The truncated-minus-exact contrast is $0.0921\,[0.0283, 0.1604]$.

D.1 Adapter count and subset variability

For each trained bank, all $\binom{5}{K}$ specialist subsets are evaluated at unit nominal product coefficients. Ratios use the individual subset's product norm before averaging subsets. Table 9 gives the sample SD of twenty bank-level means. Within-bank subset SD instead averages to $(.0224, .0194, .0067)$ for $K = 2$, $(.0291, .0150, .0070)$ for $K = 3$, and $(.0299, .0106, .0060)$ for $K = 4$, in the order factor

Table 8: Stress interventions in the controlled classifier. Each observation is the mean of five task-level endpoint KLs within one of twenty banks. The exact product sum is evaluated without rank truncation and is not a rank-16 candidate in the selector library.

Stress control	Mean-task KL [95% interval]	Across-bank SD
Unit factor arithmetic ($25F$)	$49.386\,[48.251, 50.472]$	2.614
Corrected unit arithmetic ($5F$)	$7.137\,[6.928, 7.343]$	0.487
Exact product sum ($5\bar{u}$)	$8.646\,[8.476, 8.813]$	0.399
Rank-16 product sum ($5P$)	$8.738\,[8.538, 8.938]$	0.470

ratio, corrected residual, and SVD residual. For $K = 5$ there is only one subset and no within-bank subset SD.

Table 9: Adapter-count intervention using trained controlled-model factors. Values are means $\pm$ sample SD of the twenty bank-level subset averages. For $K = 5$, the SD is across banks, not across the single subset inside a bank. Factor-ray homogeneity remains exactly K^2 for every row.

K	Subsets/bank	$\|v_{\text{fac}}\| \,/\, \|w\|$	$\|v_{\text{corr}} - w\| \,/\, \|w\|$	$\|\text{Proj}_{16}(w) - w\| \,/\, \|w\|$
2	10	1.790 ± 0.007	0.327 ± 0.009	0.130 ± 0.005
3	10	2.579 ± 0.016	0.381 ± 0.012	0.166 ± 0.006
4	5	3.365 ± 0.026	0.404 ± 0.014	0.183 ± 0.007
5	1	4.150 ± 0.038	0.416 ± 0.016	0.193 ± 0.008

D.2 Estimator bias and variance

Table 10 reports the complete panel-A comparison underlying Figure 1. All intervals retain the pairing of estimators within a bank.

Table 10: Bias and variance on trained controlled-model scores. Intervals resample complete bank summaries, preserving all three estimators' pairings. The target is the finite-sequence local Fisher, not behavioral KL or teacher-label accuracy.

Estimator	Probes	Relative bias [95% interval]	Normalized variance [95% interval]
Per-token	8	$0.0038\,[-0.0015, 0.0090]$	$0.9556\,[0.9081, 1.0072]$
Per-token	32	$0.0012\,[-0.0011, 0.0037]$	$0.2404\,[0.2284, 0.2530]$
Per-token	128	$0.0004\,[-0.0008, 0.0018]$	$0.0593\,[0.0565, 0.0622]$
Corrected summed score	8	$-0.0011\,[-0.0099, 0.0071]$	$2.7334\,[2.6253, 2.8427]$
Corrected summed score	32	$0.0030\,[-0.0039, 0.0101]$	$2.1543\,[2.1061, 2.2042]$
Corrected summed score	128	$-0.0061\,[-0.0145, 0.0026]$	$2.0193\,[1.9697, 2.0742]$
Label marginalized	8	$-0.0014\,[-0.0038, 0.0011]$	$0.1406\,[0.1337, 0.1478]$
Label marginalized	32	$0.0002\,[-0.0006, 0.0011]$	$0.0354\,[0.0336, 0.0372]$
Label marginalized	128	$0.0004\,[-0.0001, 0.0008]$	$0.0088\,[0.0084, 0.0093]$

D.3 Paired second moments

Table 11 gives panel A, bank 1, task 3, with $n = 96$. Define $S_{QQ} = \sum_z (Q_z - \bar{Q})^2$, $S_{DD} = \sum_z (D_z - \bar{D})^2$, and $S_{QD} = \sum_z (Q_z - \bar{Q})(D_z - \bar{D})$. The marginal SEs are $\sqrt{S_{QQ}/[n(n-1)]}$ and $\sqrt{S_{DD}/[n(n-1)]}$; the paired SE is Equation (27).

E Selection, Metric Decomposition, and Native Performance

The panel-A results in Table 12 use the fixed twenty-four-candidate library. Their main negative control is that high within-task correlation does not ensure the correct minimum. The per-token and marginalized selectors coincide on all twenty banks; corrected summed-score selection differs once.

Table 11: Centered second-moment sums for controlled bank 1, task 3. All rows use the same ninety-six source sequences. These quantities recover the marginal and paired standard errors in the diagnostic comparison in Section 6.

Candidate	S_{QQ}	S_{DD}	S_{QD}
Factor average	19.926264	11.396021	5.714389
Factor TIES (0.5)	221.291781	65.396263	42.495600
Factor DARE	154.619319	44.389826	33.885310
Factor Fisher	24.556086	10.802705	7.081508
Product average	25.243788	14.903407	7.656765
Product TIES (0.5)	136.581995	49.829243	19.457092
Product Fisher	37.411664	17.074955	12.123725
KnOTS–TIES (0.2), $\gamma = .5$	55.019044	18.281254	8.757353
KnOTS–TIES (0.5), $\gamma = .5$	35.317424	14.672544	7.329340

Table 12: Panel A controlled-classifier selection over 20 independent banks and 24 candidates. Each bank contributes its mean of five within-task correlations and one minimax decision. Brackets are 95% paired bank-bootstrap intervals for correlation and regret, and binomial score intervals for top-1. KnOTS selection searches its eight configurations using selection-split KL; its regret is against the full-library oracle.

Diagnostic or policy	Within-task ρ	Top-1	Worst-task regret
Taskwise displacement norm	0.967 [0.965, 0.969]	0.000 [0.000, 0.161]	0.3606 [0.3305, 0.3933]
Per-token Fisher quadratic	0.990 [0.989, 0.991]	0.600 [0.387, 0.781]	0.0045 [0.0013, 0.0085]
Corrected summed-score quadratic	0.987 [0.985, 0.988]	0.600 [0.387, 0.781]	0.0171 [0.0021, 0.0434]
Label-marginalized quadratic	0.990 [0.989, 0.991]	0.600 [0.387, 0.781]	0.0045 [0.0013, 0.0085]
Centered output drift	0.966 [0.964, 0.968]	0.000 [0.000, 0.161]	0.3606 [0.3305, 0.3933]
Selection-split KL	0.997 [0.996, 0.997]	0.600 [0.387, 0.781]	0.0031 [0.0008, 0.0058]
Product Fisher, fixed	—	0.000 [0.000, 0.161]	0.3298 [0.2873, 0.3734]
KnOTS–TIES–SVD16, selected	—	0.000 [0.000, 0.161]	0.5700 [0.5263, 0.6195]

The paired excess regret of corrected summation is .0125 [.0000, .0376], while the uniform-output-minus-marginalized difference is .3560 [.3247, .3898].

E.1 Library and scale sensitivity

Panel A's full-library correlation and the correlation after deleting all four scale-five entries are .967/.944 for norm, .990/.983 for per-token Fisher, .987/.978 for corrected summation, .990/.984 for marginalization, .966/.943 for uniform output, and .997/.995 for selection-split KL. The reduced support contains twenty candidates. The uniform-output failure therefore persists without relying on the largest scale to inflate rank association.

Deleting complete DARE, Fisher, KnOTS, or TIES families leaves the original norm/marginalized/output regrets unchanged at .3606/.0045/.3606. Deleting the factor ray gives .3486/.0088/.3486; deleting the product ray gives .2017/.0000/.2017. Each comparison recomputes both the selector and oracle after removal. The per-bank candidate support, rather than a removed full-library optimum, defines its regret.

E.2 Finite-logit decomposition in panel B

For each policy, choose its candidate using only the selection split. For the decomposition, identify that candidate's largest task KL on its endpoint split and average the paired quantities on that task. This post-selection decomposition diagnoses the chosen endpoint; it does not provide the selector with endpoint information. The exact equality $D = \log \mathbb{E}_p e^{\delta} - \mathbb{E}_p \delta$ applies at each position before sequence and bank aggregation.

The uniform policy selects P in nineteen banks and F in one; the weighted policy always selects $0.5F$. The local Fisher selects $0.5F$, $0.5P$, and factor Fisher in fifteen, two, and three banks. Direct selection-split KL selects those candidates in nine, nine, and two banks, respectively. The endpoint

Table 13: Panel B metric intervention. Each entry first takes the maximum task score within a bank and then averages twenty banks. Uniform and specialist-weighted columns evaluate identical logits. Selection identities are reported separately because an ordering of these means is not a bankwise selection rule.

Candidate	Selection $\max_t q_t^{\mathrm{out}}$	Selection $\max_t q_t^{\mathrm{wlog}}$	Endpoint $\max_t D_t$
$0.5F$	10.996	0.694	1.749
F	10.225	0.956	1.930
$0.5P$	10.466	0.755	1.744
P	9.736	1.186	2.127

Table 14: Panel B decomposition at the selected candidate's worst endpoint task. Each interval resamples twenty paired bank summaries. Local and finite-logit quantities use the same specialist probabilities, but different directions. The log-normalizer decomposition holds before aggregation.

Quantity at worst endpoint task	Uniform-output selection	Weighted-output selection
Local Fisher q	$1.483\,[1.424, 1.545]$	$1.065\,[1.037, 1.092]$
Finite-logit q^{wlog}	$1.156\,[1.102, 1.213]$	$0.655\,[0.640, 0.669]$
Behavioral KL D	$2.123\,[2.055, 2.188]$	$1.749\,[1.723, 1.774]$
Network remainder $q^{\mathrm{wlog}} - q$	$-0.328\,[-0.372, -0.280]$	$-0.410\,[-0.438, -0.380]$
Logit remainder $D - q^{\mathrm{wlog}}$	$0.967\,[0.919, 1.011]$	$1.094\,[1.068, 1.119]$
$\mathbb{E}_p\delta$	$-4.155\,[-4.296, -4.005]$	$-5.435\,[-5.540, -5.334]$
$\log \mathbb{E}_p e^{\delta}$	$-2.032\,[-2.217, -1.845]$	$-3.687\,[-3.772, -3.604]$

optimum selects $0.5P$, $0.5F$, and factor Fisher in eleven, eight, and one bank. All comparisons use the full frozen library, not a retrospectively optimized scale grid.

E.3 Native-rank KnOTS controls

Native-rank and SVD16 KnOTS each search eight density–scale settings by minimum worst-task selection KL. Their reference floor is the original rank-matched library's best endpoint. A second comparison fixes the SVD16-selected setting and changes only rank. Table 15 separates the two contrasts. The density .5 and output scale .5 are selected in nineteen banks by each selector, although the exceptional bank need not be the same. These comparisons retain the native method's aligned coordinate construction rather than replacing it by a heuristic factor average.

Table 15: Panel B native-rank alignment controls. Regret is relative to the same twenty-four-candidate rank-matched oracle. KL differences are paired within a bank. Native outputs have rank at most thirty-two per layer, whereas the matched competitors have rank at most sixteen.

KnOTS control	Mean [95% paired-bank interval]
SVD16 selected: regret	$0.6740\,[0.5861, 0.7693]$
Native rank selected: regret	$0.6804\,[0.6026, 0.7649]$
Native minus SVD16, settings reselected: KL	$0.0065\,[-0.0185, 0.0338]$
Native minus SVD16, setting fixed: KL	$0.0146\,[-0.0140, 0.0458]$

E.4 Native correctness and the candidate-library ceiling

Controlled-model native accuracy is agreement with the fixed task teacher's maximizing label. Positions are averaged within sequences before averaging sequences and tasks. For positive specialist accuracy $S_{b,t}$, task retention is the selected accuracy divided by $S_{b,t}$. Mean retention averages these task ratios within a bank; worst retention takes their task minimum before averaging banks. It is not a ratio of heterogeneous pooled scores. The native oracle maximizes mean task retention over the twenty-four candidates using endpoint labels and is reported only as a retrospective library ceiling.

Panel A's base and specialist accuracies are $31.86\,[31.58, 32.16]\%$ and $50.96\,[50.59, 51.35]\%$. The marginalized selector reaches accuracy $35.21\,[34.90, 35.52]\%$, mean retention $68.96\,[68.39, 69.53]\%$, and worst retention $60.70\,[59.42, 61.98]\%$. Direct-KL selection reaches mean retention $69.47\,[68.79, 70.18]\%$. These modest retentions motivate the explicit library ceiling in panel B.

Table 16: Panel B native outcomes with paired bank intervals. The last row chooses on endpoint labels and bounds attainable mean retention in the rank-matched library; it is not an evaluated selection procedure. Other policies choose without those labels. Each bank contains five tasks and ninety-six endpoint sequences per task.

Panel B policy	Accuracy (%)	Mean retention (%)	Worst retention (%)
Common base	30.88 [30.67, 31.10]	—	—
Specialists	51.93 [51.54, 52.32]	—	—
Uniform output drift	31.89 [31.45, 32.32]	61.37 [60.55, 62.20]	50.85 [49.19, 52.48]
Marginalized Fisher	33.61 [33.24, 33.96]	64.69 [64.03, 65.40]	51.31 [49.90, 52.67]
Weighted output drift	33.58 [33.21, 33.92]	64.64 [64.01, 65.27]	50.46 [49.02, 51.93]
Selection-split KL	33.77 [33.39, 34.11]	65.00 [64.31, 65.70]	52.36 [50.98, 53.81]
Product Fisher, fixed	33.93 [33.58, 34.32]	65.34 [64.64, 66.11]	53.01 [51.47, 54.57]
KnOTS–TIES–SVD16	28.53 [27.89, 29.14]	54.94 [53.76, 56.12]	43.22 [41.04, 45.45]
KnOTS–TIES, native	29.08 [28.52, 29.65]	55.97 [54.90, 57.12]	45.30 [43.50, 47.13]
Native-retention oracle	34.25 [33.94, 34.59]	65.95 [65.28, 66.68]	—

Task identities and teacher functions are fixed within a panel, so repeated banks do not constitute new task populations. Likewise, the Pythia code pools contain only forty-five MBPP and fifty HumanEval problems: one additional success changes their raw scores by 2.22 and 2.00 percentage points. Those small native pools and near-floor specialists do not support precise language-model retention comparisons.

E.5 Bank-level decision records for panel B

Table 17 reports all twenty replication decisions and regrets. Candidate codes are $f = 0.5F$, $p = 0.5P$, $F = F$, $P = P$, and $H =$ factor Fisher. The norm policy chooses P throughout; the weighted policy chooses f throughout. Small positive regrets are retained at six decimal places rather than replaced by an exact zero.

F Covariance-Interpolation Sensitivity

The endpoint comparison can be refined without changing any candidate. Let $G_U = C/V$ be the uniform categorical covariance and interpolate the coordinate metric as

$$q^{\mathrm{mix}}_{t,\beta} = \tfrac{1}{2}\mathbb{E}_{Q_t}\delta^\top\{(1-\beta)G_U + \beta G(p_t)\}\delta, \qquad \beta \in \{0, .25, .5, .75, 1\}. \tag{33}$$

At zero, this is half the uniform score and has the same selector; at one, it is specialist-weighted drift. Intermediate values preserve both positive semidefiniteness and common-offset invariance. The same frozen logits therefore support an explanatory metric sweep, with no additional model training or candidate tuning.

The selected scale changes progressively rather than requiring the endpoint $\beta = 1$ (Table 18). At $\beta = .25$, eighteen banks still choose P and two choose F. At .5, nine choose P, nine choose $.5P$, and two choose F. At .75, every bank chooses $.5P$, while full specialist weighting selects $.5F$ throughout. This connects the uniform-score failure to the metric's preference between ordinary and half-scale updates. The lowest mean regret along this diagnostic sweep occurs at .75, but its coefficient is not promoted to a tuned selector: the sweep is a mechanism analysis on the fixed replication panel.

The residual error after this change has a different source from the original uniform preference. The endpoint oracle alternates between $.5F$, $.5P$, and factor Fisher across banks, whereas a fixed metric can favor one ray systematically. Correcting the output metric is therefore neither sufficient for exact top-1 selection nor evidence that parameter curvature becomes irrelevant. The direct-KL reference

Table 17: Panel B bank-level decisions. Regret uses the maximum task KL on the independent endpoint split and the full twenty-four-candidate oracle. Complete bank tuples, not individual task–candidate cells, define the bootstrap unit.

	Uniform output		Marginalized Fisher		Weighted output		Oracle
Bank	Choice	Regret	Choice	Regret	Choice	Regret	Choice
1	P	0.375338	f	0.049016	f	0.049016	p
2	P	0.433930	f	0.026394	f	0.026394	p
3	P	0.236551	f	0.031033	f	0.031033	p
4	P	0.681308	f	0.000000	f	0.000000	f
5	P	0.243438	f	0.000000	f	0.000000	f
6	P	0.383131	f	0.010604	f	0.010604	p
7	F	0.036866	p	0.000000	f	0.001313	p
8	P	0.573686	p	0.000000	f	0.038034	p
9	P	0.550115	f	0.000000	f	0.000000	f
10	P	0.303822	f	0.000000	f	0.000000	f
11	P	0.505612	f	0.000000	f	0.000000	f
12	P	0.389834	f	0.062540	f	0.062540	p
13	P	0.370629	f	0.024604	f	0.024604	p
14	P	0.352638	H	0.000000	f	0.116556	H
15	P	0.454211	H	0.076421	f	0.000000	f
16	P	0.181384	f	0.063997	f	0.063997	p
17	P	0.398397	H	0.094375	f	0.015395	p
18	P	0.675982	f	0.000000	f	0.000000	f
19	P	0.426768	f	0.035682	f	0.035682	p
20	P	0.387946	f	0.000000	f	0.000000	f

Table 18: Panel B covariance interpolation. Half-scale selections count $.5F$ or $.5P$. Every coefficient uses the same selection and endpoint streams. The grid is a sensitivity analysis, not an endpoint-selected hyperparameter for the main comparison.

Specialist weight β	Half-scale selections	Endpoint hits	Regret [95% paired-bank CI]
0	0/20	0/20	.3981 [.3307, .4643]
.25	0/20	0/20	.3869 [.3187, .4544]
.5	9/20	5/20	.2212 [.1362, .3149]
.75	20/20	11/20	.0186 [.0067, .0330]
1	20/20	8/20	.0238 [.0120, .0377]

and paired bank records separate this remaining sampling and candidate-ordering uncertainty from the large ordinary-scale error.

G Causal Attention: Architecture, Training, and Paired Records

G.1 A second controlled architecture

Panel C is the causal sequence classifier in Equation (18), not a pretrained language model. Its four matrices have shape 32×32, and the output has 32 classes. The mask is zero at the current and preceding positions and negative infinity at later positions; softmax is rowwise. The residual carries each position's observed input directly to the attention output. Only query and readout matrices receive rank-16 LoRA updates with scale two. Keys and values remain fixed, so changing the query update changes which preceding positions are used without changing the candidate support or evaluation inputs.

A NumPy PCG64 generator with primary seed $M = 924671$ draws the two adaptable base matrices, then the key and value matrices, with independent $\mathcal{N}(0, 1/32)$ entries. Next it draws the full teacher right-factor array of shape $5 \times 2 \times 16 \times 32$, with the same variance, and the teacher left-factor array of shape $5 \times 2 \times 32 \times 16$, with variance $.30^2/16$. Each of five teachers uses the common frozen keys and values and query/readout weights $W_{0\ell} + 2B^*_{t\ell}A^*_{t\ell}$. The next draws are task means $\mu_t \sim \mathcal{N}(0, .60^2 I)$

and coordinate scales $d_{tj} = \exp(a_{tj})$, $a_{tj} \sim \mathrm{Unif}[-.5, .5]$. These base weights, teachers, and task distributions remain fixed over all twenty banks.

Each source sequence draws a length uniformly from $\{4, 8, 16\}$, a common latent $z \sim \mathcal{N}(0, I)$, and sixteen independent vectors $\epsilon_\tau \sim \mathcal{N}(0, I)$. Its inputs are $x_\tau = \mu_t + \mathrm{diag}(d_t)(.5z + \sqrt{.75}\epsilon_\tau)$. Positions beyond the sampled length are zero padded and excluded from all losses and scores. Valid causal positions cannot attend to this trailing padding. Native labels are the corresponding teacher's maximizing logits. The teacher's input is the complete valid prefix, so its labels need not be functions of the current vector alone.

For zero-indexed bank b and task t, training sequences use seed $M + 200000 + 100b + t$. Each specialist receives 512 sequences. Initialization uses seed $M + 300000 + b$ to draw the query and readout right factors, with variance $1/32$; these factors are shared across specialists within the bank, and left factors start at zero. A minibatch generator with seed $M + 400000$ draws a 100×8 index array at each of 160 steps, sampling training sequences with replacement. Cross-entropy is averaged over valid positions within a sequence, then sequences and specialist blocks. Blocks share no trainable parameters or optimizer moments. AdamW uses learning rate $.004$, weight decay $.001$, betas $(.9, .999)$, and epsilon 10^{-8}, without scheduling, dropout, or clipping. Panel C's training unit is a sequence, whereas panels A and B train pointwise examples; cross-architecture scores are not presented as compute-matched comparisons.

Geometry, selection, and endpoint streams use seeds $M+100000+1000b+100t+j$, with $j = 1, 2, 3$ and counts $16, 48, 96$. All candidates share these inputs. The implementation uses PyTorch $2.10.0$ on CPU with four threads. Gaussian model arrays and training inputs are first represented in float32; training is float32, and trained weights and evaluation inputs are promoted to float64 for SVDs, categorical reductions, and analytic derivatives. The two matrix derivatives are computed at the actual specialist, in ambient query/readout coordinates. An automatic-directional-derivative check on the first bank agrees at the tested inputs; an independent dense score-Jacobian check at one fixed prefix has maximum absolute errors 1.38×10^{-14} for the ambient Fisher diagonal and 4.18×10^{-14} for the factor diagonals.

G.2 Candidate construction and exact score derivatives

The same twenty-four operator constructions in Table 7 are applied to the two adapted matrices. Each factor/product TIES control retains $\lceil \rho d \rceil$ largest absolute entries in a tensor of size d, including cutoff ties, elects retained-sign frequency, and averages agreeing nonzero values. DARE draws its full right-factor mask, full left-factor mask, and product mask in that order using PCG64 seed $M + 500000 + b$; survival probability is .3. Coordinate-specific model-Fisher diagonals integrate the full categorical distribution on geometry sequences, with uniform sequence and within-sequence weights. Output scaling, final rank-16 truncation, duplicate exclusion, and the eight aligned KnOTS settings follow Appendix C.2. The native-rank KnOTS control omits only the final truncation.

For a prefix position, let $k_i = W_K x_i$, $v_i = W_V x_i$, attention weights a_i, and $\bar{v} = \sum_i a_i v_i$. The derivative of the attended value with respect to the query vector is

$$H = \frac{\sum_i a_i v_i k_i^\top - \bar{v}(\sum_i a_i k_i)^\top}{\sqrt{32}}. \tag{34}$$

For an ambient direction (h_Q, h_O) and residual representation $z = x + \bar{v}$, the directional logit derivative is $j_h = 2W_O H h_Q x + 2h_O z$. The score coefficient for the query matrix is the categorical-centered row of $2W_O H$, multiplied by $x^\top$; the readout coefficient is $2(e_y - p)z^\top$. Their categorical second moments give the ambient diagonal. Chain-rule multiplication by the actual specialist factors gives the factor diagonals. This evaluates the same model Fisher as the marginalized directional score, rather than using a gold-label gradient covariance.

Every endpoint record contains the sequence-balanced marginalized quadratic, full-label KL, uniform and weighted output scores, native accuracy, and specialist accuracy. Selection and endpoint labels are distinct. Ties in the implemented native-selector score are resolved by the first candidate in the fixed library ordering. All reported behavioral minima are checked using the same ordering. Native accuracy, retention, and base-to-specialist gain recovery are recomputed within each bank before aggregation; no ratio of displayed table averages defines the reported gain-recovery interval.

H CAUSAL-ATTENTION RESULTS AND RECONCILIATION

H.1 ESTIMATION ON ATTENTION-MODEL SCORES

The estimator experiment uses four fixed geometry sequences per task and bank, giving 400 sequence–direction units. Each direction is factor averaging minus the relevant specialist. A unit uses 256 independent probe replicates, seed $M + 700000 + 1000b + 100t + i$. Uniform position sampling is with replacement, a fresh categorical label is drawn for every occurrence, and the shorter probe arms use prefixes of the 128-probe stream. The per-token and summed-score arms share both positions and labels; marginalization shares positions. The reference averages the exact categorical variance over all valid positions of that sequence. The three empirical variances in Table 19 are normalized by its squared half-quadratic reference before averaging units within banks.

Table 19: Panel C normalized estimator variance from the retained probe replicates. Intervals resample complete bank summaries with all estimator arms paired. This experiment estimates the local Fisher, not behavioral KL or native accuracy.

Probes	Per-token variance [95% CI]	Corrected-sum variance [95% CI]	Marginalized variance [95% CI]
8	1.2851 [.8838, 1.9590]	3.0301 [2.6086, 3.7095]	.1162 [.1081, .1243]
32	.3024 [.2200, .4387]	2.2902 [2.1752, 2.4342]	.0289 [.0269, .0308]
128	.0822 [.0547, .1299]	2.0766 [1.9916, 2.1942]	.0073 [.0068, .0078]

At 128 probes, relative errors of the finite-replicate mean are .00227 [.00068, .00415] for per-token scores, −.00174 [−.00992, .00692] for corrected summation, and .00036 [−.00008, .00079] for marginalization. They are reported separately from the analytically unbiased expectations. Exact summation of the same retained categorical score moments gives normalized variances .0720, 2.0564, and .00725, respectively. The variance identity applies to those distributional moments; finite-replicate variance and mean estimates fluctuate around them. Exact conditional score means have maximum absolute discrepancy below 4×10^{-15} in this calculation.

H.2 PAIRED SEQUENCE UNCERTAINTY

Table 20 supplies an independent reconciliation of the sequence-level errors for panel C, bank 1, task 3. Every row uses $n = 96$ endpoint sequences. The centered sums have the definitions of Table 11. For example, factor averaging has $q = .5982$, $D = .7062$, marginal SEs .0409 and .0362, and paired SE .0280. These quantities are computed from the sequence contributions, not supplied by a shared correlation model.

Table 20: Panel C means and centered second moments. Marginal SEs are $\sqrt{S_{QQ}/9120}$ and $\sqrt{S_{DD}/9120}$; the paired calibration SE is $\sqrt{(S_{QQ} + S_{DD} - 2S_{QD})/9120}$. Rounding is applied after computing each quantity.

Candidate	$\bar{Q}$	$\bar{D}$	S_{QQ}	S_{DD}	S_{QD}
Factor average	.598207	.706195	15.224444	11.958071	10.019821
Factor TIES (.5)	1.506329	1.410417	120.220082	70.197693	71.433879
Factor DARE	.875864	1.056836	25.329816	21.744051	12.842337
Factor Fisher	.531363	.682893	11.085918	11.465888	8.212150
Product average	.596286	.680345	14.554765	12.124215	9.713614
Product TIES (.5)	1.373930	1.355821	76.602268	49.842180	40.233674
Product Fisher	.542306	.678946	11.146837	11.192845	7.745927

H.3 RANK AND SCALE CONTROLS

The attention-model stress controls give mean-task KL 21.6683 [21.1529, 22.1791] for $25F$ and 2.4719 [2.4110, 2.5357] for corrected arithmetic $5F$. The exact sum $5\bar{u}$ gives 3.0838 [3.0217, 3.1477], compared with 3.0967 [3.0345, 3.1617] for $5P$. The paired truncated-minus-exact difference is

.0129 $[-.0063, .0324]$: unlike panel A, this experiment does not resolve a positive mean KL penalty from rank truncation.

Selection among the eight SVD16 KnOTS configurations gives regret .2338 $[.2083, .2625]$ against the rank-matched oracle; the native-rank selector gives .2292 $[.2016, .2585]$. Holding the SVD16-selected density and scale fixed, native-minus-truncated worst-task KL is $-.0046$ $[-.0172, .0089]$. Native-rank outputs achieve 50.62% accuracy and 66.52% mean retention. These contrasts again do not attribute the baseline's behavior to rank truncation, and they make no claim about the method's published pretrained-model benchmarks.

H.4 Co-primary utility and selection records

For selected mean task accuracy $\bar{S}_b$, base accuracy $\bar{S}_b^0$, and specialist accuracy $\bar{S}_b^{\mathrm{sp}}$, gain recovery is

$$G_b = 100\frac{\bar{S}_b - \bar{S}_b^0}{\bar{S}_b^{\mathrm{sp}} - \bar{S}_b^0}, \qquad \bar{S}_b^{\mathrm{sp}} > \bar{S}_b^0. \tag{35}$$

All panel-C banks have positive denominators. Native selection gives mean recovery $20.43\,[18.70, 22.21]\%$, KL selection $17.80\,[15.75, 19.85]\%$, and the native-label oracle $21.60\,[20.24, 23.07]\%$. Mean worst-task retention is $65.54\,[64.40, 66.60]\%$ for native selection, $64.81\,[63.76, 65.71]\%$ for KL selection, and $65.53\,[64.44, 66.58]\%$ for fixed product Fisher. These intervals quantify the individual estimands; paired method differences are computed before resampling rather than inferred from interval overlap.

The native selector chooses product Fisher in eleven banks, factor Fisher in seven, and P in two. The endpoint-native oracle chooses those three candidates in thirteen, five, and two banks. The KL oracle instead chooses factor Fisher in eleven banks, F in three, product Fisher and $0.5P$ in two each, and $0.5F$ and P in one each. These candidate identities separate an objective shift from a universal preference for a particular scale. Under a retrospective worst-retention floor of .8, no candidate is selection-feasible in any bank, further illustrating the library's native-performance limit.

I Pythia Independent-Bank Transfer Protocol

I.1 Checkpoint, independent banks, and fixed candidate support

The transfer protocol specifies 10 independent Pythia-1.4B banks, with one specialist for each of the five corpora in Appendix B. The design comprises 50 specialist fits and 1,200 bank–task–candidate endpoint cells, but only 10 independent bank-level selection decisions. The pretrained model and tokenizer are `EleutherAI/pythia-1.4b`; their resolved revisions, dataset revisions, and source-ID allocations are recorded before training. The common checkpoint and the fixed task set define the scope of inference. Shared endpoint examples across banks do not create new independently sampled tasks or benchmark problems.

Each bank uses an independent adapter-initialization stream, training-order stream, and merger-mask stream. Within a bank, specialists share their initial right factors, as in the Pythia deployment study; different banks do not share trained adapters. The recipe retains rank 16, LoRA scale two, zero dropout, the `query_key_value` and attention `dense` modules, one full-sequence-loss epoch, at most 500 training examples per task, maximum length 256, and AdamW learning rate 2×10^{-4} and weight decay 10^{-3}. Batch size is two with four-step gradient accumulation. Initialization, optimizer state, and minibatch order are reset independently for every bank. The final partial accumulation group follows the explicitly recorded training rule rather than being silently treated differently between banks.

The primary library is exactly the 24 constructions in Table 7, applied layerwise to the pretrained model. The output-scale grid, TIES densities, DARE survival probability, Fisher coordinates, KnOTS alignment convention, and final rank remain unchanged. A scalar-equivalent corrected update is not added a second time. Native-rank KnOTS and the exact product sum remain separate rank controls. Consequently, improved selection cannot be attributed to a larger library or an extra scale available to only one diagnostic.

I.2 Source separation and the paired record

The original training and endpoint windows are retained. For GSM8K, Alpaca, and HH-RLHF, the geometry pool takes the first 128 eligible source sequences in its existing geometry window; selection and endpoint retain 100 each. MBPP uses 128 test sequences for geometry and the two disjoint 45-problem validation halves for selection and endpoint. HumanEval is reassigned as in Table 21; its geometry no longer reuses training examples. These role assignments are fixed before fitting any bank, and each record retains the original source ID. Exact prompt–response duplicates and truncated-token duplicates are checked across roles; any resulting exclusions and final cardinalities belong to the run manifest.

Table 21: Source-allocation specification for the Pythia replication, before duplicate and eligibility checks. For HumanEval, shuffled source windows are training $[0, 64)$, geometry $[114, 139)$, selection $[139, 164)$, and endpoint $[64, 114)$. Other role windows follow Table 5. Counts denote source examples, not tokens or Monte Carlo labels.

Task	Training	Geometry	Selection	Endpoint
GSM8K	500	128	100	100
HumanEval	64	25	25	50
MBPP	$\min(500, N_{\mathrm{tr}})$	128	45	45
Alpaca	500	128	100	100
HH-RLHF	500	128	100	100

For every bank, task, candidate, and source sequence, retain its eligible-position count and sequence means of Q, D, q^{out}, and q^{wlog}, together with the task's native outcome. All four diagnostics use the same teacher-forced prefixes and masks. Positions are averaged within a source sequence before averaging sequences. Candidate and specialist dropout are disabled, logits have temperature one, categorical probabilities use the full vocabulary, and scalar reductions are accumulated in float64. Streaming vocabulary reductions avoids treating a full persistent logit cache as a prerequisite for pairing.

The marginalized quadratic uses a directional logit derivative at the specialist in the realized adapted-weight direction $v - u_t$, not a derivative in raw factor coordinates or a finite difference mislabeled as an exact derivative. A layerwise directional-derivative check against automatic differentiation is part of the implementation validation. Diagonal Fisher weights used to construct the two Fisher mergers are estimated separately in their stated coordinate systems using 128 sequence-balanced model-label probes per task; the sampled-label diagonal is not substituted for the label-marginalized diagnostic. These probes and their randomness are held fixed across candidate comparisons within a bank.

I.3 Transfer targets and co-primary native outcomes

The preservation contrast is $R_b(q^{\mathrm{wlog}}) - R_b(q^{\mathrm{out}})$, using Equation (19) on the independent endpoint sample. The native contrast compares the two selected candidates on the same endpoint examples. Both are computed within a bank before aggregation. Marginalized Fisher and selection-split KL provide the parameter-local and directly matched-target references. Fixed full-scale product Fisher is evaluated without candidate selection. Its native ordering is tested on Pythia rather than inferred from panels B and C.

Language-model native metrics remain task specific. GSM8K uses parsed numerical exact match, and HumanEval and MBPP use one greedy completion and functional test execution with a 512-token generation limit. Timeouts, compilation failures, and failed tests count as failures. Alpaca reports geometric reference-response token likelihood; HH-RLHF reports length-normalized chosen-versus-rejected response likelihood ordering, with half credit for exact ties. Neither is described as code or mathematical accuracy. Each endpoint report includes the common base, specialists, selected candidates, and fixed product Fisher. A specialist score of zero leaves retention undefined rather than invoking a stabilizing denominator. The small code pools are reported through success counts as well as percentages.

The estimator check uses the same factor-average direction, probe counts $8, 32, 128$, and 256 paired position/label replicates as the controlled experiments. All-label marginalization over every retained

position supplies the finite-sequence reference. Sampled-position and label replicates quantify numerical estimation noise; source-sequence resampling quantifies context variation; the independently trained bank remains the unit for method comparisons. Per-sequence $Q - D$ standard errors follow Equation (27). Recording the actual centered cross-product sum preserves covariance without imposing a common correlation across tasks or candidates.

I.4 Precision and reporting rules

For correlation and regret, the analysis resamples complete bank tuples $10{,}000$ times with the same diagnostics, candidate identities, and endpoints paired. Top-1 uses the Wilson construction in Equation (28). Each native task has its own paired contrast and denominator. Family-removal and scale-removal analyses recompute both the selector and oracle on the restricted library. The protocol records the selected family and output scale in every bank, so a change toward attenuation is distinguishable from improved selection within a scale.

Table 22 reports conditional precision for the bank-level paired KL contrast. Here s_Δ is an assumed bank-level SD and $h = t_{.975,n-1} s_\Delta / \sqrt{n}$ is the corresponding two-sided t-interval half-width under independent, approximately Gaussian differences. Increasing the bank count narrows uncertainty for this checkpoint and task collection, not across architectures.

Table 22: Conditional precision in nats per position for the paired bank-level regret difference. The first column gives assumed SD values; interval half-widths follow the displayed t-interval formula.

Paired bank SD input s_Δ	Half-width, $n = 10$	Half-width, $n = 20$
0.010	0.00715	0.00468
0.020	0.01431	0.00936
0.040	0.02861	0.01872