WHICH TASK LOSES? UNDERSTANDING ASYMMETRIC DEGRADATION IN MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

How many tasks can a merged model retain, and which task is most likely to suffer? We study merging as task-vector superposition and derive two-sided capacity bounds. Under a curvature assumption along the merge path, a stationary expert's loss increase lies between curvature multiples of its interferers' energy. The upper bound limits capacity in the task count K; the lower bound quantifies unavoidable damage. For two tasks, the smaller-vector task loses more whenever the squared norm ratio exceeds the curvature ratio. All 11 Llama pairs satisfying this condition under sampled, uncertified curvature ranges follow the predicted ordering. More broadly, the smaller-norm task's loss rises more in 91% of 58 full-data task-arithmetic pairs. Across 486 directional evaluations spanning four $\sim$7B families and three methods, within-merge contrasts account for 64.1% of degradation variance, which symmetric mergeability scores cannot explain; the smaller-norm task loses more in 83.5% of comparisons. Retraining links this asymmetry to exposure: norms grow with exposure, while matching exposure across tasks (three seeds) compresses their ratios and removes the ordering signal (51.8%). On four pairs, shrinking the larger vector protects the smaller task better than uniform scaling. Equalizing norms at fixed total norm instead stretches the smaller vector and harms 13 of 16 targets.

1 INTRODUCTION

Model merging combines independently fine-tuned models without jointly retraining them. Task Arithmetic (TA) adds their displacements from a shared base model, while TIES and DARE modify these task vectors before combining them (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024). A merge can preserve one task while severely degrading another. Predicting which capability will suffer is therefore as useful as predicting the merge's average performance.

Most work on merge prediction asks whether two experts are *mergeable*, a property shared by both tasks (Zhou et al., 2026b), or how many tasks a model can retain. We connect the capacity question to a directional one: *which* task will a merge sacrifice? Our starting point is the displacement from each expert to the merged model. For any K, two-sided bounds place a stationary expert's loss increase between curvature multiples of the energy of its interferers (§3). The upper bound limits capacity in K; the lower bound requires an explicit curvature assumption along the expert-to-merge path. For pairwise and small-K merges with the small coherences we measure, interferer scale contributes far more to these bounds than coherence (whole-vector cosine). In a pair, this yields a simple prediction: the smaller-vector task incurs the larger loss increase once the squared norm ratio exceeds the curvature ratio.

Our experiments follow this prediction from observation to intervention. Nearly two thirds of merge-damage variance comes from within-merge contrasts that symmetric scores cannot represent, and the smaller-norm task usually loses more. Training exposure sets the norm ratio. When exposure is matched, the ratio falls below the spread of the residual curvature term and no longer orders the loser. Changing the ratio after training also changes the outcome: shrinking the larger vector protects the smaller task, whereas equalizing norms at fixed total norm stretches the smaller vector and hurts retention.

Contributions. Figure 1 connects four results. (1) *Two-sided capacity bounds (§3).* Under a two-sided curvature assumption along the merge path, each task's loss increase lies between $\alpha_i/2$ and $\beta_i/2$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

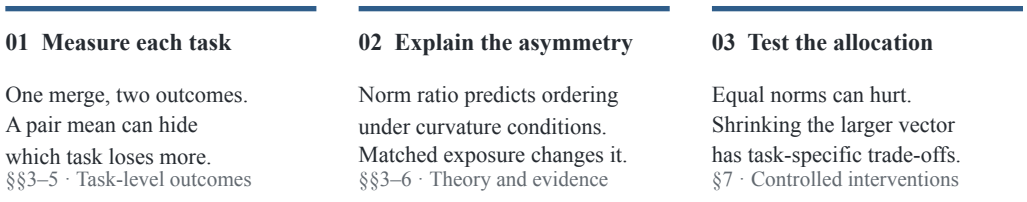

Figure 1: The analysis moves from directional degradation to exposure and then to allocation. The ordering prediction depends on curvature conditions (Corollary 1). Matching exposure changes both the spread of norms and the observed ordering (Section 6). The allocation experiments then test how each target responds to equalization and shrinkage, without assuming that equal norms are desirable (Section 7).

times its interferer energy, with a gradient term and coherence bracket $1 \pm (K-2)\mu_{\max}$ (Theorem 1). The upper bound limits capacity in K, the lower bound remains positive below $K_{\text{bal}} = 2 + 1/\mu_{\max}$, and the pairwise corollary predicts which task loses (Corollary 1). (2) *Scale predicts directional damage (§5).* In task arithmetic, each expert is displaced by its partner's vector. The smaller-norm task has the larger loss increase in 91.4% of 58 pairs, including all 11 Llama pairs satisfying the corollary under sampled, uncertified curvature ranges. Within-merge contrasts account for 64.1% of raw variance across 486 evaluations from four families and three methods (46–72% under metric-homogeneous rescalings). Symmetric predictors, including coherence and pair-level mergeability features, can thus explain at most 35.9% and cannot identify the loser. The smaller-norm rule orders 83.5% of pair–method comparisons correctly, matching the training-set-size rule. Across up to eight tasks (1,016 Llama TA evaluations), coherence burden adds at most 0.0032 to R^2. (3) *Exposure sets the ratio (§6).* Increasing one task's training set raises its norm and its partners' mean degradation in all four dose series. Retraining all eight tasks of Llama and Pythia on 2,490 examples with equal steps (three seeds) reduces norm spread from 28.6× and 75.3× to at most 1.5×. The effective-curvature term then accounts for 98% of loss-contrast variance, and the smaller-norm rule orders 51.8% of metric comparisons correctly. (4) *Shrink the larger vector (§7).* Equalizing norms at fixed total update norm stretches the smaller vector 4.7–50.3-fold and lowers the primary metric on 13 of 16 targets. Shrinking only the larger vector instead protects the smaller-norm target better than scaling both vectors equally.

2 Related Work

Predicting mergeability and task vulnerability. Existing predictors address several distinct outcomes. Demystifying Mergeability predicts pair-average retention from weight, subspace, activation, and gradient features. These include a symmetric min/max norm ratio, whose antisymmetric form we use (Zhou et al., 2026b). Will it Merge? relates retention to base-model knowledge (Rahamim et al., 2026), while task-level collapse has been studied using target-specific hidden-state distances (Cao et al., 2026). SimMerge learns operator and merge-plan selection (Bolton et al., 2026). MergeProbe proposes directional retention prediction from geometry and training metadata, supported by simulator-based pilot evidence (Tang et al., 2026). We measure directional outcomes on real merges, bound what pair-level features can explain, and compare directional predictors with an exposure baseline under identical held-group splits (Appendix L).

Norm, training exposure, and optimization. DisTaC shows that the larger task vector dominates when norms differ, traces this mismatch to training steps and learning rate, and finds that naive rescaling can hurt (Yoshida et al., 2026). Self Positioning relates data imbalance to task-vector magnitude (Li et al., 2024); a safety case study tests data balancing under asymmetric collapse (Choudhary et al., 2026). Viewing task arithmetic as one-shot federated averaging, Tao et al. (2025) correct training heterogeneity with a FedNova-style normalization by each task's summed learning-rate schedule. Our step-ratio baseline is the predictive counterpart of that normalization. A task vector approximates a learning-rate-scaled accumulated gradient (Zhou et al., 2026a), so its norm tracks exposure when learning rate and epoch budget are shared, as they are here. Training duration and the

optimizer's effective noise scale also affect mergeability (Horoi et al., 2026; Kozodoi et al., 2026; Zhang et al., 2026). Within each family, our experts share the epoch budget, learning rate, and optimizer settings. Their training duration varies through dataset size, which spans two orders of magnitude, and best-checkpoint selection, until we control it directly (§6). Building on DisTaC, we derive when the norm ratio determines the loser and test how training and rescaling change that ratio.

Calibration, geometry, and evaluation regime. Magnitude-aware merging includes the coefficients used by DO-Merging, NAN, and FroM (Zheng et al., 2025; Si et al., 2025; Li et al., 2025). AdaMerging learns coefficients from unlabeled test inputs (Yang et al., 2024), and LiNeS scales updates by layer depth (Wang et al., 2025). Task Singular Vectors, isotropic merging, SVC, and Task Alignment operate on subspaces, singular components, or representation discrepancies (Gargiulo et al., 2025; Marczak et al., 2025; Li et al., 2026; de Jorge et al., 2026). Activation-level factorial interventions further separate nonadditivity from recovery (Zhu, 2026). Our experiments are limited to fixed, data-free, whole-vector rules. Benchmarks, large-scale and in-the-wild studies, and scaling laws show that merge outcomes depend on the expert collection and evaluation goal (He et al., 2025; Yadav et al., 2025; Hitit et al., 2026; Wang et al., 2026). Gradient-matching and tangent-space analyses provide complementary functional explanations (Daheim et al., 2024; Ortiz-Jimenez et al., 2023).

3 Interferer Energy Bounds Each Task's Loss

Let $\boldsymbol{\theta}_0$ denote the shared base model, $\boldsymbol{\theta}_i$ an expert, and $\boldsymbol{\tau}_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}_0$ its task vector. Task arithmetic combines K experts as $\boldsymbol{\theta}_M = \boldsymbol{\theta}_0 + \sum_{k=1}^{K} \boldsymbol{\tau}_k$. Relative to expert i, the merge adds the other task vectors: $\boldsymbol{r}_i = \boldsymbol{\theta}_M - \boldsymbol{\theta}_i = \sum_{j\neq i} \boldsymbol{\tau}_j$. Its effect on that expert's training or validation loss L_i is therefore $\Delta L_i = L_i(\boldsymbol{\theta}_i + \boldsymbol{r}_i) - L_i(\boldsymbol{\theta}_i)$. To bound this change, write $\boldsymbol{g}_i = \nabla L_i(\boldsymbol{\theta}_i)$ and define the interferer energy $A_i = \sum_{j\neq i} \|\boldsymbol{\tau}_j\|^2$. Let $\mu_{\max} = \max_{j\neq k;\, j,k\neq i} |\cos(\boldsymbol{\tau}_j, \boldsymbol{\tau}_k)|$ be the largest absolute coherence among interferers ($\mu_{\max} = 0$ when $K = 2$), with $\sigma_{\min} = \min_{j\neq i} \|\boldsymbol{\tau}_j\|$ and $\sigma_{\max} = \max_{j\neq i} \|\boldsymbol{\tau}_j\|$.

Assumption 1 (Curvature along the merge path)**.** *L_i is twice continuously differentiable on a neighbourhood of the segment $\{\boldsymbol{\theta}_i + t\boldsymbol{r}_i : t \in [0,1]\}$, and there are $0 < \alpha_i \leq \beta_i$ such that $\alpha_i\|\boldsymbol{r}_i\|^2 \leq \boldsymbol{r}_i^\top \nabla^2 L_i(\boldsymbol{\theta}_i + t\boldsymbol{r}_i)\boldsymbol{r}_i \leq \beta_i\|\boldsymbol{r}_i\|^2$ for all $t \in [0,1]$.*

The upper constant imposes the usual smoothness condition along this segment. The positive lower constant is a separate assumption: it fails wherever the loss is flat or concave along $\boldsymbol{r}_i$. An upper curvature bound alone cannot supply it.

Theorem 1 (Two-sided merge bound)**.** *Under Assumption 1,*

$$\boldsymbol{g}_i^\top \boldsymbol{r}_i + \tfrac{\alpha_i}{2} A_i\big(1 - (K-2)\mu_{\max}\big) \;\leq\; \Delta L_i \;\leq\; \boldsymbol{g}_i^\top \boldsymbol{r}_i + \tfrac{\beta_i}{2} A_i\big(1 + (K-2)\mu_{\max}\big). \tag{1}$$

At a stationary expert ($\boldsymbol{g}_i = \boldsymbol{0}$), replacing every interferer norm by $\sigma_{\max}$ gives the capacity bound $\varepsilon^\star(K) = \frac{\beta}{2}\sigma_{\max}^2(K-1)(1 + (K-2)\mu)$, where β is restricted to the merge segment and μ to interferer pairs. Under the assumption, sufficiently small K keeps this bound below a loss tolerance ε. The corresponding lower bound is $\Delta L_i \geq \frac{\alpha_i}{2}(K-1)\sigma_{\min}^2(1 - (K-2)\mu_{\max})$. It is positive for $K < K_{\mathrm{bal}} = 2 + 1/\mu_{\max}$ and, for fixed α_i and $\mu_{\max}$, grows with K while $2(K-1)\mu_{\max} < 1$. Beyond K_{bal}, signed cancellation can make $\|\boldsymbol{r}_i\|$ small, leaving the lower bound vacuous. The constants α_i, β_i and $\mu_{\max}$ depend on the selected tasks. At stationarity and fixed curvature, task-vector dependence enters only through A_i and $\mu_{\max}$. For pairs, the bracket $1 \pm (K-2)\mu_{\max}$ equals one. For larger merges, each additional task changes it by at most $\pm 3\%$ for Llama ($\mu_{\max} = 0.027$) and $\pm 17\%$ for Pythia (0.168). Interferer scale therefore dominates both bounds in the pairwise and small-K settings studied here.

Corollary 1 (Which task loses)**.** *For a two-task merge $\boldsymbol{\theta}_0 + \boldsymbol{\tau}_i + \boldsymbol{\tau}_j$ ($K = 2$, unit coefficients) with nonzero task vectors, at stationary experts and under Assumption 1 for both tasks, $\Delta L_i = \frac{q_i}{2}\|\boldsymbol{\tau}_j\|^2$ and $\Delta L_j = \frac{q_j}{2}\|\boldsymbol{\tau}_i\|^2$ for realized curvatures $q_i \in [\alpha_i, \beta_i]$ and $q_j \in [\alpha_j, \beta_j]$, so*

$$\log \frac{\Delta L_i}{\Delta L_j} = 2\log \frac{\|\boldsymbol{\tau}_j\|}{\|\boldsymbol{\tau}_i\|} + \log \frac{q_i}{q_j}. \tag{2}$$

Task i incurs a larger loss increase than task j whenever $\|\boldsymbol{\tau}_j\|^2/\|\boldsymbol{\tau}_i\|^2 > \beta_j/\alpha_i$.

To examine this relation empirically, define the effective curvature $\hat{q}_i = 2\Delta L_i/\|\boldsymbol{\tau}_j\|^2$ for pairwise task arithmetic. Substituting $\hat{q}$ for q makes Equation 2 an identity whenever $\Delta L_i, \Delta L_j > 0$. The loss contrast then separates into twice the log norm ratio and an effective-curvature term. Away from stationarity, however, $\hat{q}_i = q_i + 2\boldsymbol{g}_i^\top \boldsymbol{\tau}_j/\|\boldsymbol{\tau}_j\|^2$ can fall outside $[\alpha_i, \beta_i]$; the guarantee concerns q. The corollary thus predicts the loser when the norm ratio is large enough to dominate curvature variation, and makes no prediction otherwise. Section 6 tests whether training exposure determines this ratio.

4 Setup and a Cap on Symmetric Predictors

We merge backbone parameters and retain the target task's trained head. For a primary metric M_i in which higher is better, we define degradation as $D_i = M_i(\boldsymbol{\theta}_i) - M_i(\boldsymbol{\theta}_M)$. Our pairwise predictors are relative scale $s_i = \log_{10}(\|\boldsymbol{\tau}_j\|/\|\boldsymbol{\tau}_i\|)$ and absolute coherence $\mu_{ij} = |\langle \boldsymbol{\tau}_i, \boldsymbol{\tau}_j \rangle|/(\|\boldsymbol{\tau}_i\|\|\boldsymbol{\tau}_j\|)$.

Experts and merges. We fine-tune Llama 3.1-8B, Mistral 7B, Qwen 2.5-7B, and Pythia 6.9B on eight GLUE classification/regression tasks using LoRA (Hu et al., 2022) (rank 16, alpha 32). Training runs for three epochs at learning rate 2×10^{-4}, and we retain the best validation checkpoint. Task vectors exclude classification heads; each evaluation restores the target expert's head. Quality control (QC) retains 8, 5, 6, and 8 viable experts, respectively (Appendix B). We evaluate both tasks of every valid pair under TA, TIES (retention 0.2), and DARE (pruning rate 0.9). This yields 486 directional evaluations from 81 family–pair groups, which share trained experts.

Decomposition. For a fixed pair and merging method, order its tasks alphabetically and write

$$m = \frac{D_i + D_j}{2}, \qquad a = \frac{D_i - D_j}{2}, \qquad (D_i, D_j) = (m + a, m - a). \tag{3}$$

The mean m measures average damage, while the contrast a identifies which task loses more. Swapping the targets preserves m and μ_{ij} but reverses the signs of a and s_i. Across P pair–method combinations, the total centered sum of squares decomposes exactly as

$$\sum_{p=1}^{P} \left[(D_{pi} - \bar{D})^2 + (D_{pj} - \bar{D})^2 \right] = 2 \sum_{p=1}^{P} (m_p - \bar{D})^2 + 2 \sum_{p=1}^{P} a_p^2. \tag{4}$$

A predictor that assigns the same value to both targets, such as μ_{ij} or a pair-level mergeability score, cannot capture variation within a pair–method combination. Its in-sample R^2 is therefore bounded by the pair-mean share of Equation 4, and it cannot identify which target loses. An antisymmetric predictor such as s_i captures only the contrast.

Inference. We quantify uncertainty using pair-cluster and shared-expert (dyadic, t_{26}) standard errors, along with pair-group and expert-resampling bootstraps (Appendix Table 7). These summaries are conditional on the observed experts; they do not provide sampling guarantees for new tasks or families.

5 Scale, Not Coherence, Predicts How Much and Which Task Loses

We first test the pairwise prediction on validation loss, the quantity bounded by Corollary 1, and then examine task metrics across four families and three methods (§5.1–5.3). We also extend the analysis to Llama TA merges of up to eight tasks, where interferer energy tracks damage and coherence adds little (§5.4).

5.1 The smaller-norm task has the larger loss increase

A TA pair displaces expert i by exactly $\boldsymbol{\tau}_j$, allowing us to test Corollary 1 on validation loss. The smaller-norm task has the larger loss increase in 91.4% of 58 pairs (pair-bootstrap 95% CI $[82.8, 98.3]$; Appendix Table 3). We exclude STS-B because its MSE loss is not comparable. Using sampled,

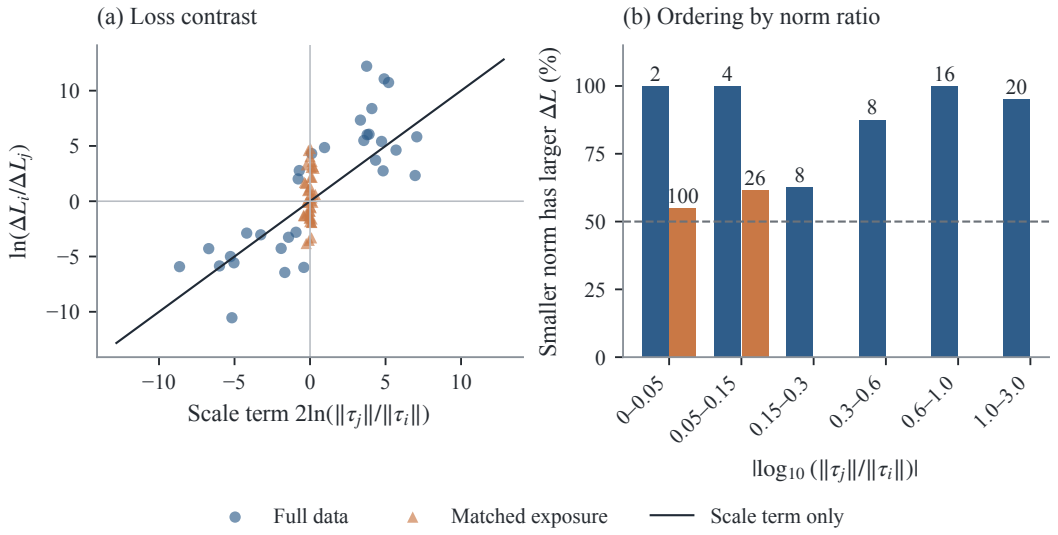

Figure 2: Larger norm ratios more often identify the task with the larger loss increase (Corollary 1). Results use task-arithmetic pairs, exclude STS-B, and include matched-exposure pairs from Section 6. (a) Log loss contrast versus the scale term for pairs with two positive loss changes: 31 full-data pairs (circles) and 41 matched-exposure pairs (triangles). The line shows the scale term alone, with slope one and no fitted parameter; vertical deviations are the effective-curvature term. (b) Fraction of pairs in which the smaller-norm task has the larger loss increase, grouped by absolute log norm ratio and source. Bar labels give pair counts. Low-ratio bins contain almost only matched pairs; high-ratio bins contain only full-data pairs. The dashed line marks chance.

uncertified curvature ranges for Llama (Appendix J), 11 of 21 pairs satisfy the corollary's condition, and all 11 follow its prediction. Loss does not rise for 29 of 116 targets, including 25 larger-norm tasks. These non-stationary LoRA experts therefore violate the stationary premise of the lower bound on validation loss. The ordering test uses all pairs; the decomposition uses only the 31 with two positive changes. For those pairs, Equation 2 decomposes $\ln(\Delta L_i/\Delta L_j)$ exactly: scale accounts for 54% of the variance, effective curvature for 32%, and their covariance for 14%. The scale term alone gives $R^2 = 0.66$ with slope one and no fitted parameter (Figure 2a). The median absolute curvature term is 2.36 in natural-log units, a ratio of about 10 estimated from these same pairs. This suggests that scale dominates once $|\log_{10}(\|\boldsymbol{\tau}_j\|/\|\boldsymbol{\tau}_i\|)| \gtrsim 0.5$. Ordering accuracy rises with the ratio (Figure 2b), although low-ratio bins contain almost only experts retrained at equal steps (Section 6). Among full-data pairs alone, the rule is correct in 11/14 below 0.3 and 42/44 above.

5.2 Scale, not coherence, explains the within-merge contrast

Most variation across the 486 evaluations occurs between the two tasks within a merge. These contrasts account for 64.1% of total squared variation in the 486 outcomes (Equation 4), limiting any symmetric predictor to 35.9%. This helps explain the pooled comparison of $R^2 = 0.284$ for relative scale and 0.003 for absolute coherence (Appendix Table 6): the two predictors address different components, with $0.284 = 0.641 \times 0.443$ and $0.003 = 0.359 \times 0.0083$. Within their respective components, s explains 44.3% of contrast energy (Appendix Figure 6a), while μ explains 0.8% of pair-mean variance (also 0.8% within family and method). The pair's mean log norm, another symmetric feature, explains 6.3% (12.1%), so symmetry alone does not account for coherence's weak performance. The pooled scale slope is 0.175 (pair-cluster SE 0.022; dyadic SE 0.031, $p < 0.001$), rising to 0.252 with family, method, and task fixed effects (dyadic SE 0.036, $p < 0.001$). Pooled coherence is not significant (-0.808, dyadic SE 0.959, $p = 0.41$).

Outside Pythia, coherence spans only 0.001–0.027. Its negative fixed-effects coefficient (-1.575) has an expert-bootstrap interval that includes zero. When outcomes are made metric-homogeneous, the smaller-norm rule still correctly orders 74.9–88.1% of contrasts (Appendix B).

5.3 Norm and training-set size order the loser equally well

The simple rule that the smaller-norm expert degrades more is correct in 203/243 pair–method comparisons (83.5%; expert bootstrap $[73.5, 92.0]$; Appendix Figure 6b). Accuracy is 78.6% for Llama, 76.7% for Mistral, 91.1% for Qwen, and 86.9% for Pythia. It is 42/51 (82.4%; expert bootstrap $[70.4, 97.0]$) for pairs sharing a primary metric. By comparison, the gap between the experts' own scores orders held-out contrasts near or below chance: 50.2% across held families and 44.9% across held tasks. Scale also reduces held-out contrast RMSE, although expert-bootstrap intervals for this improvement include zero (Appendix B.1).

Training exposure as the baseline. A simpler predictor requires no model weights. Let $d = \log_{10}(N_j/N_i)$ describe relative training-set size and $\omega = \log_{10}(T_j/T_i)$ the relative optimizer steps of the selected checkpoints; metadata cover all 27 experts. Predicting greater degradation for the task with fewer examples correctly orders 83.5% of comparisons ($[77.4, 89.3]$). Using fewer steps gives 84.8% ($[79.0, 90.1]$), and the scale and data log ratios correlate at 0.979. Because N_i is fixed by task across families, the data rule is also a fixed ordering: RTE<MRPC<STS-B<CoLA<SST-2<QNLI<QQP<MNLI. The pairwise study alone therefore cannot separate exposure from task identity, difficulty, or metric. Section 6 breaks this link by retraining every task at equal exposure. In held-family and held-task prediction, the scale-minus-data contrast RMSE differences are $+0.003$ ($[-0.011, +0.019]$) and $+0.001$ ($[-0.008, +0.009]$), respectively. Neither is resolved, and adding scale to data does not help (Appendices B.2 and B.3, Appendix Table 10). Thus, norm and exposure are statistically indistinguishable ordering predictors in this study, consistent with norm tracking exposure under a shared learning rate and epoch budget (Zhou et al., 2026a). Directional predictors should be compared with this weight-free baseline.

5.4 Up to eight tasks, capacity is an interferer-energy budget

Under Assumption 1, the bounds apply to any K. In the Llama TA sweep, coherence adds little beyond interferer scale through eight tasks (Appendix Figure 7). For K-task TA, we use $s_i^{(K)} = \log_{10}(A_i/\|\boldsymbol{\tau}_i\|^2)$, with interferer energy $A_i = \sum_{j\neq i}\|\boldsymbol{\tau}_j\|^2$, and define the *coherence burden* as $b_i = B_i/A_i$. Here $B_i = 2\sum_{j<k;\, j,k\neq i}\mu_{jk}\|\boldsymbol{\tau}_j\|\|\boldsymbol{\tau}_k\|$ bounds cross-term energy among interferers. Both predictors follow from Theorem 1: A_i is the energy term, and $b_i \leq (K-2)\mu_{\max}$ bounds the relative cross-term controlled by the coherence bracket (Appendix A). Applying unit-coefficient TA to every subset of eight Llama tasks gives 1,016 target evaluations on overlapping subsets ($K = 2, \ldots, 8$). Scale-only R^2 using $s_i^{(K)}$ ranges from 0.130–0.370; coherence burden adds at most 0.0032. In a separate $K = 3$ DARE check, it adds 0.087 ($p = 0.130$; Appendix C). Many scores are at or below the best constant predictor: 29.8% in the sweep and 30.7% in pairwise evaluations. At $K \geq 5$, scale therefore mainly identifies which tasks collapse. Excluding these floor rows leaves a pairwise scale slope of 0.135 (Appendix C). For Llama, the bracket in Theorem 1 remains within 17% of one at $K = 8$, making interferer energy the main capacity constraint in this setting. Fixed-norm perturbations of Llama interferer directions can move outcomes in either direction (Appendix I).

6 Training Exposure Sets the Norm Ratio

The pairwise results leave open whether norms reflect exposure or intrinsic differences between tasks (Section 5.3). We address this by retraining all eight tasks in Llama 3.1-8B and Pythia 6.9B with the same recipe and equal exposure. Each task uses a fixed random subset of 2,490 examples, the size of RTE's full training set, and retains its final checkpoint. Every expert within a family thus takes the same number of optimizer steps: 234 for Llama and 468 for Pythia (Appendix Table 4). Three seeds (42, 1, 2) vary both the subset and training seed. We merge all 28 pairs with TA, TIES, and DARE and score both targets on the full validation sets. The pairwise-study evaluator scores both the merge and the corresponding individual experts (Appendix D). Matching exposure compresses the norm

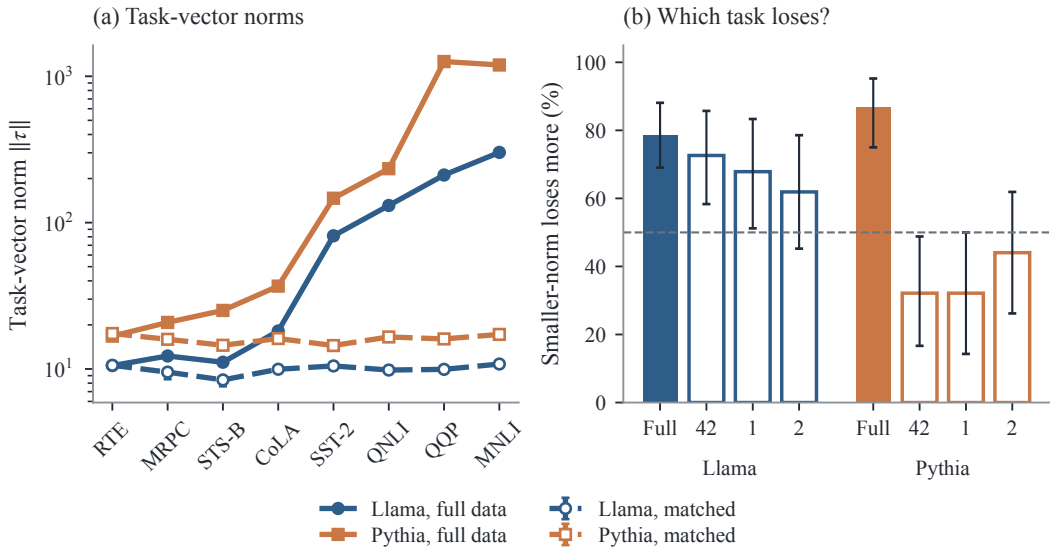

Figure 3: Matching exposure compresses norm spread, but ordering accuracy differs across families. (a) Task-vector norms on a log scale for full-data experts (solid) and experts trained on 2,490 examples with identical optimizer steps (dashed; median and range over three seeds). Tasks follow full training-set size. (b) Accuracy of the smaller-norm ordering rule for full data (filled) and matched seeds 42, 1, and 2 (open). Pair-bootstrap 95% intervals use 5,000 resamples of the 28 pairs and do not account for the eight-expert structure; Section 6 gives exact expert-level permutation tests. The dashed line marks chance. Figure 11 reports the separate shrink experiment.

spread and removes the pooled ordering signal; a separate dose experiment tests what happens when one task receives more data.

Matched exposure removes the norm signal. The max/min norm ratio across eight tasks falls from 28.6 to 1.25–1.50 for Llama and from 75.3 to 1.22–1.25 for Pythia (Figure 3a). In loss-test pairs, the median ratio is 1.07 (median $|\log_{10}|$ ratio 0.028, versus 0.81 for full-data pairs). These ratios are far below the curvature spread, where Corollary 1 makes no prediction. Of 126 TA pairs, only 41 have two positive loss changes; 104 of 252 targets have $\Delta L \leq 0$. In those 41 pairs, Equation 2, with $\hat{q}$ defined as the residual, assigns 98% of loss-contrast variance to effective curvature and 0.7% to scale. Across all 126 pairs, the smaller-norm task has the larger loss increase in 56.3% ($[46.0, 66.7]$). On task metrics, the rule correctly orders 51.8% of 504 comparisons pooled across families and seeds ($[42.3, 61.3]$). Per-seed accuracy is 72.6, 67.9, and 61.9% for Llama and 32.1, 32.1, and 44.0% for Pythia (Figure 3b); within-pair scale explains at most 2.4% of contrast variance in any seed. Exact expert-level permutation tests use all $8!$ norm assignments per seed. Only Llama seed 42 exceeds chance at one-sided $p < 0.05$ ($p = 0.041$; seeds 1 and 2: 0.090, 0.208), and no Pythia seed falls below chance (lower-tail $p \geq 0.087$). Norm therefore does not reliably order the loser in either family at matched exposure. This is consistent with the corollary's no-prediction regime. The residual scale term is about 30 times smaller than for full-data experts, so curvature variation would obscure even an effect of full-data strength.

Task-intrinsic differences remain. Within-merge contrasts still account for 31–54% of the reduced variance (Appendix Table 17), but the tasks that lose are different. With full-data experts, a task's share of lost comparisons decreases with training-set size (Spearman -0.85 for Llama and -0.86 for Pythia). After matching exposure, these shares no longer track their full-data values: correlations are -0.04 and 0.14 for seed 42, and 0.00 and 0.13 for the seed mean (Appendix Table 18 and Figure 8). The effective-curvature residual absorbs these remaining task differences.

Dose-response. We next vary exposure for one task at a time. MNLI and QQP are retrained on nested subsets of 2,490, 10,000, and 40,000 examples using one seed. The full-data expert, trained with best-epoch selection, provides a fourth level. Each expert is merged with the same seed-42 matched RTE, MRPC, and CoLA experts. Norm grows monotonically with training-set size (Llama MNLI $11 \to 27 \to 69 \to 302$; Pythia MNLI $17 \to 49 \to 130 \to 1195$), as does partners' mean degradation in all four series (Llama MNLI $0.168 \to 0.233 \to 0.325 \to 0.365$; Llama QQP $0.008 \to 0.149 \to 0.381 \to 0.406$; Pythia MNLI $0.126 \to 0.245 \to 0.313 \to 0.346$; Pythia QQP $0.051 \to 0.265 \to 0.348 \to 0.565$). The ordering changes less consistently. Llama QQP loses more in 7 of 9 comparisons at 2,490 examples and in none at 10,000 or 40,000. The other three series lose at most 3 of 9 at every level (Appendix Figure 9 and Table 19). More exposure thus increases relative scale and partners' mean degradation, without a uniform change in loser identity.

Damage is lower at matched exposure. Mean pairwise degradation falls from 0.270 to 0.052–0.097 for Llama and from 0.244 to 0.082–0.101 for Pythia. Matching also reduces every task's exposure and norm, however; Llama MNLI falls from 302 to 11. The improvement is consistent with lower interferer energy (Theorem 1), but cannot be attributed to equalization alone because we have no matched-exposure arm at a larger common size. Retraining establishes how exposure changes the norm ratio. We next hold training fixed and test how changing coefficients affects the loser (Section 7).

7 Shrink the Larger Vector; Do Not Stretch the Smaller

Having changed the norm ratio through retraining (Section 6), we now change it directly through merge coefficients. The experiments distinguish shrinking the larger vector from stretching the smaller one. We use the same trained Llama 3.1-8B and Pythia 6.9B experts in four fixed pairs covering all eight tasks: MNLI+RTE, QNLI+STS-B, MRPC+SST-2, and CoLA+QQP. For alphabetically ordered tasks i, j, let $\hat{\boldsymbol{\tau}}_i = \boldsymbol{\tau}_i / \|\boldsymbol{\tau}_i\|$, $n_i = \|\boldsymbol{\tau}_i\|$, $\eta = \langle \hat{\boldsymbol{\tau}}_i, \hat{\boldsymbol{\tau}}_j \rangle$, and $\Lambda = \|\boldsymbol{\tau}_i + \boldsymbol{\tau}_j\|$. Define

$$\boldsymbol{\tau}_M(\rho) = \Lambda \frac{\hat{\boldsymbol{\tau}}_i + \rho\, \hat{\boldsymbol{\tau}}_j}{\|\hat{\boldsymbol{\tau}}_i + \rho\, \hat{\boldsymbol{\tau}}_j\|}, \qquad \boldsymbol{\theta}(\rho) = \boldsymbol{\theta}_0 + \boldsymbol{\tau}_M(\rho). \tag{5}$$

This construction sets the component norm ratio to ρ while preserving both directions, their signed cosine, and the total update norm Λ. We evaluate $\rho \in \{1/4, 1/2, 1, 2, 4\}$ and natural TA at $\rho_0 = n_j/n_i$ on full validation sets using a controlled BF16 evaluator. The 96 merge evaluations have a maximum post-BF16 norm deviation of 0.142%; Appendix Figure 10 shows all 96.

Equalizing at fixed total norm stretches the smaller vector and harms retention. We prespecify the comparison between $\rho = 1$ and natural TA. At fixed Λ, equalization gives each component norm $\Lambda/\sqrt{2 + 2\eta}$, close to $\max(n_i, n_j)/\sqrt{2}$ for these nearly orthogonal, unequal pairs. The larger vector is therefore scaled by about 0.7, while the smaller vector grows 4.7–50.3-fold (coefficients $c_i(1), c_j(1)$ in Appendix Table 22). Equalization lowers the primary metric on 13 of 16 targets: 7/8 for Llama and 6/8 for Pythia. Both improvements occur on MRPC, whose F1 remains below the all-positive reference of 0.812; Pythia CoLA stays at zero. Llama QQP F1 falls from 0.865 to 0.142, and Pythia MNLI accuracy from 0.778 to 0.318. We do not select a ratio after evaluation (Appendix E).

Matched single-component controls. Evaluating each rescaled component alone separates the target's loss change exactly into isolated and interaction terms. Isolated rescaling raises loss for 15 of 16 targets, while the interaction change is negative for 13 of 16. The observed harm therefore does not require an increase in interaction (Appendix Table 24).

Shrinking the larger vector instead. On the same four pairs of full-data experts, we keep the smaller vector's coefficient at 1 and scale the larger by $c \in \{1, 0.75, 0.5, 0.25\}$, with $c = 1$ recovering TA. Both targets are scored with the pairwise-study evaluator (Appendix Figure 11; Appendix Table 21). At $c = 0.75$, close to the 0.696–0.715 coefficients that equalization assigns to the larger vector (Appendix Table 22), 13 of 16 targets improve over TA in this evaluator. Equalization, by comparison, harmed 13 of 16 in the controlled evaluator. The smaller-norm target improves in 4/4 Llama and 3/4 Pythia pairs; Pythia CoLA remains at zero MCC. Mean degradation of the larger-norm target changes by -0.004 for Llama and -0.003 for Pythia, with one clear loss: Pythia QNLI ($+0.057$).

To isolate the role of preserving the smaller vector, we also scale both vectors by c. The larger vector receives the same coefficient in both treatments. Shrinking only the larger vector better protects the smaller task: mean D is 0.245 versus 0.281 for Llama and 0.213 versus 0.259 for Pythia at $c = 0.75$, and 0.222 versus 0.294 and 0.180 versus 0.267 at $c = 0.5$. Uniform scaling also moves the smaller vector away from its trained endpoint. The larger-norm task, in contrast, does as well or better under uniform scaling (Pythia 0.034 versus 0.100 at $c = 0.5$; Appendix Table 20). Appendix Figure 11 compares stretching, uniform scaling, and shrinking. More aggressive shrinking reveals a trade-off. At $c = 0.25$, Llama's smaller-norm targets improve from mean $D = 0.301$ to 0.073, at a cost of $+0.019$ to the larger targets. Pythia's larger targets instead lose 0.236 on average (MNLI $+0.461$, QNLI $+0.351$), and only 2 of its 4 smaller targets improve. These controls locate equalization's harm in stretching the smaller vector; the appropriate amount of shrinkage still depends on the task and family.

Equalizing down instead of up. Mean-one inverse-norm coefficients equalize component norms at their harmonic mean. In a pair, they can at most double the smaller vector and primarily shrink the larger one (Equation 13, Appendix K). In a separate calibration experiment, this rule lowers the larger-norm task's metric in 49 of 53 pairs. The smaller-norm task loses more in only 20/53 pairs, compared with 47/53 under uniform TA (Appendix K). Changing the coefficient ratio alone can therefore change which task loses.

8 Discussion and Limitations

Take-home. Interferer energy connects merge capacity to task-level damage. Under the curvature assumption, a stationary expert's loss increase is bounded above and below by multiples of that energy. In pairwise merges and the Llama K-task sweep ($\mu_{\max} \approx 0.027$), the coherence bracket stays within 17% of one, making energy the main capacity constraint in K. This is an empirical scale relation, not a certified $K_{\max}$. The directional prediction agrees with all 11 Llama pairs satisfying the condition under sampled curvature ranges, 91.4% of TA loss comparisons, and 83.5% of metric comparisons. Symmetric scores cannot explain the 64.1% within-merge contrast share, and coherence burden adds at most 0.0032 to R^2 through eight tasks in Llama TA. Exposure sets the norm ratio; matching it removes reliable norm-based ordering. For practice, these results favor reporting each task's outcome, comparing directional predictors with the weight-free exposure baseline, and rebalancing by shrinking the larger vector rather than equalizing at fixed total norm.

Limitations. The curvature constants in Assumption 1 are not certified for real networks (Appendix J). We therefore report conditional bounds and empirical scale dependence, not a certified maximum K. The bounds concern loss rather than the metrics defining D. Our loss test covers only task arithmetic and excludes STS-B, which uses MSE. Loss does not rise for 29 of 116 full-data and 104 of 252 matched targets, so the stationary premise fails in those cases. The main evidence comes from GLUE tasks with task-specific heads, LoRA, and QC-selected experts. Exposure is confounded with task identity, difficulty, and metric in the pairwise study. Matching exposure addresses this confounding for two families at RTE's size, but cannot separate equalization from lower total exposure; the dose experiment uses one seed and three small partners.

Full-data checkpoints are selected on the evaluation sets. Shared experts create dependence that our robust intervals only approximate, and the same pairwise study informed the hypotheses tested here. Rescaling covers four pairs in two families with one training seed, using different evaluators for stretching and shrinking. Shrinkage can protect the smaller task at the larger task's expense, with the trade-off varying by task and family. We do not study model-size trends or full fine-tuning. The K-task, mask, generative-LoRA, direction-perturbation, and curvature analyses use Llama (Appendices C, F, H, I, and J). The ordering rules do not transfer to the public generative MergeBench experts, which are instruction-tuned and have approximate training counts (Appendix G).

Reproducibility Statement

The bundle provides training and merge code, repository-relative CPU analyses, tables, figures, input hashes, and all evaluation records, including per-example losses for the exposure-matched, dose-response, and rescaling experiments. It reconstructs summaries and held-group predictions

without checkpoints, and includes split membership, train-only preprocessing, fitted coefficients, and every held-out prediction. Stored logits, labels, and aligned example IDs support recomputation of controlled metrics and loss decompositions. Before computing dependence-robust intervals, the supplementary code recomputes every reused estimate; all 85 checks agree to floating-point precision. The external MergeBench package (Appendix G) includes model revisions, geometry, all 112 task scores, per-example records, and ordering and screening computations. Appendices document QC, metrics, evaluation settings, and scoring rules. Reproducing GPU training and evaluation requires weights and datasets, which are not included.

AI USE STATEMENT

In this work, the authors used generative AI tools to aid in polishing the writing and improving the readability of the manuscript, and to assist with writing and debugging the experimental code. The authors carefully reviewed, verified, and edited all AI-generated content and code, and take full responsibility for the content and correctness of this final work.

REFERENCES

Peter M. Aronow, Cyrus Samii, and Valentina A. Assenova. Cluster-robust variance estimation for dyadic data. *Political Analysis*, 23(4):564–577, 2015. doi: 10.1093/pan/mpv018.

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/2601.09473v2.

Nicolò Bossi. DMS: LoRA multi-task merging via per-module magnitude reallocation. Master's thesis, Politecnico di Milano, 2026. URL https://hdl.handle.net/10589/260744.

Yuan Cao, Dezhi Ran, Yuzhe Guo, Mengzhou Wu, Simin Chen, Linyi Li, Wei Yang, and Tao Xie. An empirical study and theoretical explanation on task-level model-merging collapse. *arXiv preprint arXiv:2603.09463*, 2026. URL https://arxiv.org/abs/2603.09463v1.

Aarnav Choudhary, Matheus Fonseca Rocha, Jiwon Seo, Vasu Sharma, and Maheep Chaudhary. Asymmetric collapse in model merging: When refusal overwrites recognition. *arXiv preprint arXiv:2607.27240*, 2026. URL https://arxiv.org/abs/2607.27240.

Nico Daheim, Thomas Möllenhoff, Edoardo Maria Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model merging by uncertainty-based gradient matching. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.12808v2.

Pau de Jorge, César Roberto de Souza, Björn Michele, Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Florent Perronnin, Diane Larlus, and Yannis Kalantidis. Task alignment: A simple proxy for practical model merging across diverse vision tasks. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2604.12935v3.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *CVPR*, 2025.

Yifei He, Siqi Zeng, Yuzheng Hu, Rui Yang, Tong Zhang, and Han Zhao. MergeBench: A benchmark for merging domain-specialized LLMs. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2025. URL https://openreview.net/forum?id=rw50iUoyLu.

Oğuz Kağan Hitit, Leander Girrbach, and Zeynep Akata. A systematic study of in-the-wild model merging for large language models. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2511.21437v2.

Stefan Horoi, Guy Wolf, Eugene Belilovsky, and Gintare Karolina Dziugaite. From memorization to parameter interference: How overtraining experts harms model merging. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2506.14126v2.

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.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Nikita Kozodoi, Zainab Afolabi, and Jack Butler. Are we merging the right models? Impact of expert training duration on model merging for LLMs. In *ICML 2026 Workshop on Weight-Space Symmetries: from Foundations to Practical Applications*, 2026. URL https://openreview.net/forum?id=KVAdINcwe5.

Mingxin Li, Zhijie Nie, Yanzhao Zhang, Dingkun Long, Richong Zhang, and Pengjun Xie. Improving general text embedding model: Tackling task conflict and data imbalance through model merging. *arXiv preprint arXiv:2410.15035*, 2024. URL https://arxiv.org/abs/2410.15035v1.

Yayuan Li, Ze Peng, Jian Zhang, Jintao Guo, Yue Duan, and Yinghuan Shi. When shared knowledge hurts: Spectral over-accumulation in model merging. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.05536v2.

Zijian Li, Xiaocheng Feng, Huixin Liu, Yichong Huang, Ting Liu, and Bing Qin. FroM: Frobenius norm-based data-free adaptive model merging. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4674–4687, 2025. URL https://aclanthology.org/2025.findings-emnlp.251/.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=RBZpAa27ls.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.12827v3.

Adir Rahamim, Asaf Yehudai, Boaz Carmeli, Leshem Choshen, Yosi Mass, and Yonatan Belinkov. Will it merge? On the causes of model mergeability. In *Findings of the Association for Computational Linguistics: ACL*, pp. 26551–26570, 2026. doi: 10.18653/v1/2026.findings-acl.1322. URL https://aclanthology.org/2026.findings-acl.1322/.

Chongjie Si, Kangtao Lv, Jingjing Jiang, Yadao Wang, Yongwei Wang, Xiaokang Yang, Wenbo Su, Bo Zheng, and Wei Shen. NAN: A training-free solution to coefficient estimation in model merging. *arXiv preprint arXiv:2505.16148*, 2025. URL https://arxiv.org/abs/2505.16148v1.

Lin Tang, Wei Zhang, Jing Li, Hongyu Chen, Ming Zhao, and Yuxuan Wang. Predicting mergeability of parameter-efficient fine-tuning updates. *arXiv preprint arXiv:2606.19549*, 2026. URL https://arxiv.org/abs/2606.19549v1.

Zhixu Tao, Ian Mason, Sanjeev Kulkarni, and Xavier Boix. Task arithmetic through the lens of one-shot federated learning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Cgyo7S7Oy0.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. LiNeS: Post-training layer scaling prevents forgetting and enhances model merging. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=J5sUOvlLbQ.

Yuanyi Wang, Yanggan Gu, Yiming Zhang, Qi Zhou, Zhaoyi Yan, Congkai Xie, Xinyao Wang, Jianbo Yuan, and Hongxia Yang. Model merging scaling laws in large language models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2509.24244v4.

Yongxian Wei, Anke Tang, Li Shen, Zixuan Hu, Chun Yuan, and Xiaochun Cao. Modeling multi-task model merging as adaptive projective gradient descent. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2501.01230v3.

Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, Lu Hou, Sinan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. OptMerge: Unifying multimodal LLM capabilities and modalities via model merging. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.19892v3.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2023.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. What matters for model merging at scale? *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=9sbetmvNpW.

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=nZP6NgD3QY.

Kotaro Yoshida, Yuji Naraki, Takafumi Horie, Ryotaro Shimizu, Ioannis Mitliagkas, and Hiroki Naganuma. DisTaC: Conditioning task vectors via distillation for robust model merging. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=W70w5JCzdq.

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 *International Conference on Machine Learning*, 2024.

Chenxiang Zhang, Alexander Theus, Damien Teney, Antonio Orvieto, Jun Pang, and Sjouke Mauw. How does the optimizer implicitly bias the model merging loss landscape? In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RU76KTF1Da.

Shenghe Zheng, Hongzhi Wang, Chenyu Huang, Xiaohui Wang, Tao Chen, Jiayuan Fan, Shuyue Hu, and Peng Ye. Decouple and orthogonalize: A data-free framework for LoRA merging. *arXiv preprint arXiv:2505.15875*, 2025.

Luca Zhou, Daniele Solombrino, Donato Crisostomi, Maria Sofia Bucarelli, Giuseppe Alessio D'Inverno, Fabrizio Silvestri, and Emanuele Rodolà. On task vectors and gradients. In *Proceedings of UniReps: the Third Edition of the Workshop on Unifying Representations in Neural Models*, volume 322 of *Proceedings of Machine Learning Research*, pp. 398–417. PMLR, 2026a. URL https://proceedings.mlr.press/v322/zhou26a.html.

Luca Zhou, Bo Zhao, Rose Yu, and Emanuele Rodolà. Demystifying mergeability: Interpretable properties to predict model merging success. In *International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=uFQwehcepH.

Chencheng Zhu. Reading the gate, not the interference: Output-side interference measurement does not track merge collapse. *arXiv preprint arXiv:2608.11797*, 2026. URL https://arxiv.org/abs/2608.11797v3.

Appendix overview. Table 1 summarizes the notation, and Table 2 links the main claims to their supporting evidence.

A Two-Sided Bounds: Proofs and Scope

This appendix develops the theory and its empirical test. We prove Theorem 1 and Corollary 1, derive the displacement identities underlying $s_i^{(K)}$ and b_i, and detail the loss-level test in Section 5.1. We also spell out the scope of the bounds. Figure 4 connects the expert-to-merge displacement, curvature along that path, and the resulting capacity bounds.

Symbol	Meaning	Symbol	Meaning
$\boldsymbol{\theta}_0, \boldsymbol{\theta}_i, \boldsymbol{\tau}_i$	base model, expert i, and its task vector $\boldsymbol{\tau}_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}_0$	$\varepsilon^\star(K)$	capacity bound $\beta\sigma_{\max}^2(K-1)\big(1+(K-2)\mu\big)/2$ at a stationary expert; $\sigma_{\max} = \max_{j\neq i}\|\boldsymbol{\tau}_j\|$
$\boldsymbol{\theta}_M$	merged model; for TA, $\boldsymbol{\theta}_0 + \sum_{k=1}^{K}\boldsymbol{\tau}_k$	K_{bal}	$2 + 1/\mu_{\max}$; the lower bound is positive for $K < K_{\text{bal}}$
K	number of merged tasks	D_i	degradation $M_i(\boldsymbol{\theta}_i) - M_i(\boldsymbol{\theta}_M)$, expert minus merged score on primary metric M_i (higher is better)
$\boldsymbol{r}_i$	displacement of expert i by the merge, $\boldsymbol{\theta}_M - \boldsymbol{\theta}_i = \sum_{j\neq i}\boldsymbol{\tau}_j$	m, a	pair mean $(D_i + D_j)/2$ and within-merge contrast $(D_i - D_j)/2$; tasks in alphabetical order
ΔL_i	loss increase of task i, $L_i(\boldsymbol{\theta}_i + \boldsymbol{r}_i) - L_i(\boldsymbol{\theta}_i)$	s_i	relative scale $\log_{10}(\|\boldsymbol{\tau}_j\|/\|\boldsymbol{\tau}_i\|)$; antisymmetric $(s_j = -s_i)$
$\boldsymbol{g}_i$	expert gradient $\nabla L_i(\boldsymbol{\theta}_i)$; $\mathbf{0}$ at a stationary expert	μ_{ij}	pair coherence $\lvert\cos(\boldsymbol{\tau}_i, \boldsymbol{\tau}_j)\rvert$; symmetric
A_i	interferer energy $\sum_{j\neq i}\|\boldsymbol{\tau}_j\|^2$	$s_i^{(K)}, b_i$	K-task scale $\log_{10}(A_i/\|\boldsymbol{\tau}_i\|^2)$; coherence burden (interferer cross-term bound over A_i)
$\mu_{\max}$	largest $\lvert\cos\rvert$ between two interferers of task i (0 when $K = 2$)	d, ω	$\log_{10}$ ratios N_j/N_i of training-set sizes and T_j/T_i of selected-checkpoint optimizer steps
α_i, β_i	lower and upper curvature of L_i along $\boldsymbol{r}_i$ on the segment from $\boldsymbol{\theta}_i$ to $\boldsymbol{\theta}_M$ (Assumption 1)	ρ, ρ_0	component norm ratio (task j to i) at fixed total norm $\|\boldsymbol{\tau}_i + \boldsymbol{\tau}_j\|$; natural TA ratio $\rho_0 = \|\boldsymbol{\tau}_j\|/\|\boldsymbol{\tau}_i\|$
$q_i, \hat{q}_i$	realized curvature, $q_i \in [\alpha_i, \beta_i]$; effective curvature $\hat{q}_i = 2\Delta L_i/\|\boldsymbol{\tau}_j\|^2$ of a TA pair	c	coefficient on the larger vector (shrink-only; smaller kept at 1) or on both (uniform scaling); $c_i(\rho), c_j(\rho)$: fixed-total-norm coefficients

Table 1: Notation used throughout the paper. Pair quantities refer to target task i and partner task j. Norms and cosines use backbone task-vector parameters, excluding classification heads.

Claim	Evidence (statistic, n)	Type	Where
Under a curvature assumption, a stationary expert's loss rise lies between curvature multiples of its interferer energy; the upper side is a capacity bound in K.	Theorem 1 and proof; assumption constants are not certified for real networks.	theorem	§3; App. A
In a pair, the smaller-vector task loses more once the squared norm ratio exceeds the curvature ratio.	Corollary 1: $\|\boldsymbol{\tau}_j\|^2/\|\boldsymbol{\tau}_i\|^2 > \beta_j/\alpha_i$ ($K = 2$, stationary experts).	theorem	§3; App. A.1
On validation loss, the smaller-norm task has the larger loss increase.	91.4% of 58 full-data TA pairs $[82.8, 98.3]$; 11 of 21 Llama pairs meet the condition under sampled, uncertified curvature ranges, and 11/11 agree.	test	§5.1; Fig. 2; Tab. 3; App. A.3
Most merge-damage variance is a within-merge contrast that symmetric (pair-level) scores cannot explain.	Contrast share 64.1% (46–72% under metric-homogeneous rescalings); 486 evaluations, 243 pair–methods, 81 family–pairs, 4 families, 3 methods.	descriptive	§4, §5.2; Fig. 5; Tab. 6
The smaller-norm task loses more on task metrics; training-set size and step count order equally well.	203/243 comparisons (83.5%, $[77.4, 88.9]$); data 83.5% $[77.4, 89.3]$, steps 84.8% $[79.0, 90.1]$.	descriptive	§5.3; Fig. 6; Tabs. 8, 9
Coherence adds little beyond interferer scale up to eight tasks.	Llama TA, $K = 2, \ldots, 8$, 1,016 evaluations: scale R^2 0.130–0.370, burden adds $\leq$0.0032.	descriptive	§5.4; Fig. 7; App. C
Exposure sets the norm ratio: at equal exposure the spread collapses and norm no longer orders the loser.	Max/min norm falls from 28.6× (Llama) and 75.3× (Pythia) to $\leq$1.50×; ordering 51.8% of 504 $[42.3, 61.3]$; 2,490 examples, 3 seeds × 2 families.	intervention	§6; Fig. 3a,b; Tabs. 16, 17, 18
More training data raises a task's norm and its partners' degradation.	Both rise monotonically over $N \in \{2{,}490, 10^4, 4\times10^4, \text{full}\}$ in 4/4 series (MNLI, QQP; Llama, Pythia); one seed, 9 comparisons per level.	intervention	§6; Tab. 19; Fig. 9
Shrink the larger vector, do not stretch the smaller: equalizing at fixed total norm hurts.	Stretch (×4.7–50.3 on the smaller vector) lowers 13/16 targets; shrink-only at $c = 0.75$ beats uniform scaling for the smaller target (mean D 0.245 vs 0.281 Llama, 0.213 vs 0.259 Pythia); 4 pairs × 2 families.	intervention	§7; Figs. 3c, 10; Tabs. 20, 21, 22, 23
The ordering rules do not transfer to public generative experts.	MergeBench: norm rule 55.6% (Llama) and 50.0% (Gemma) of 18 each, exact $p = 0.45, 0.66$.	descriptive	App. G

Table 2: Evidence for the main claims. *Theorem* rows give guarantees under Assumption 1, whose constants remain uncertified for real networks. The *test* row evaluates the corollary on validation loss, the quantity it bounds. *Descriptive* rows summarize observed merges; *intervention* rows change exposure or merge coefficients. Brackets give pair-bootstrap 95% intervals from 5,000 draws. Comparisons and evaluations reuse experts across pairs and methods, so the counts do not represent independent trials.

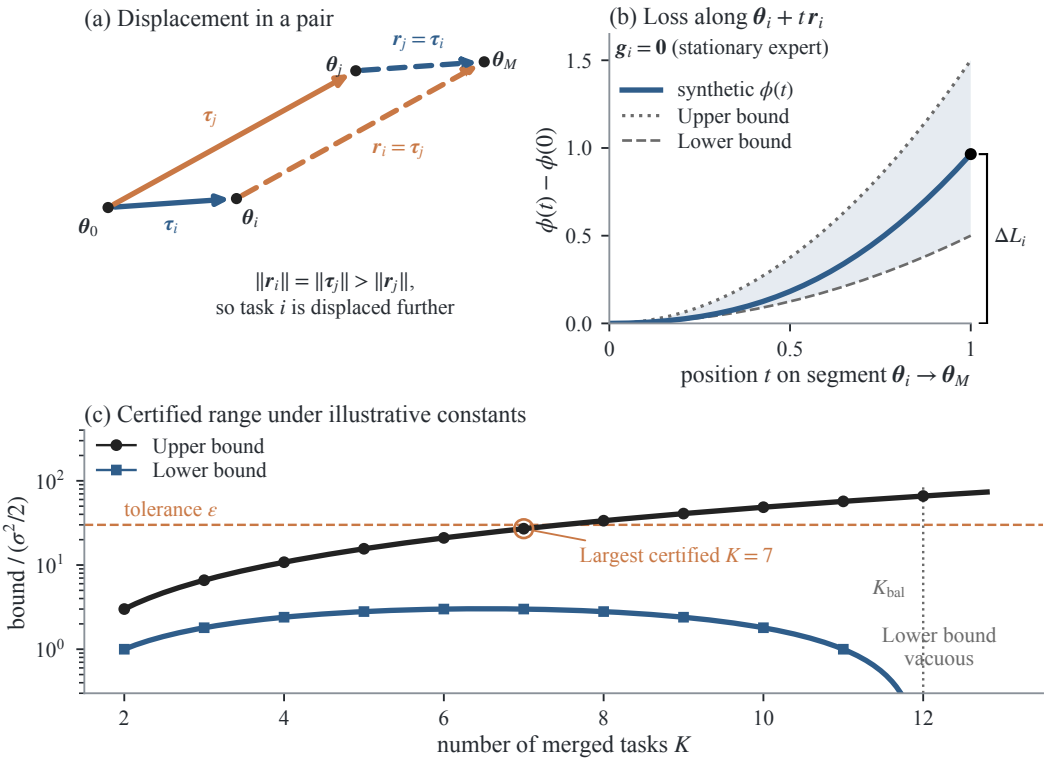

Figure 4: Geometry of Theorem 1. (a) In task arithmetic, expert i is displaced by $\boldsymbol{r}_i = \boldsymbol{\theta}_M - \boldsymbol{\theta}_i = \sum_{j \neq i} \boldsymbol{\tau}_j$. For a pair, $\boldsymbol{r}_i = \boldsymbol{\tau}_j$ and $\boldsymbol{r}_j = \boldsymbol{\tau}_i$, so the task with the smaller vector receives the larger displacement. (b) Let $\phi(t) = L_i(\boldsymbol{\theta}_i + t\boldsymbol{r}_i)$. Under Assumption 1 and at a stationary expert ($\boldsymbol{g}_i = \boldsymbol{0}$), $\phi(t) - \phi(0)$ lies between $\frac{\alpha_i}{2}\|\boldsymbol{r}_i\|^2 t^2$ and $\frac{\beta_i}{2}\|\boldsymbol{r}_i\|^2 t^2$ along the segment, so $\Delta L_i = \phi(1) - \phi(0)$ lies between $\frac{\alpha_i}{2}\|\boldsymbol{r}_i\|^2$ and $\frac{\beta_i}{2}\|\boldsymbol{r}_i\|^2$. The blue curve illustrates a synthetic loss with directional curvature in $[\alpha_i, \beta_i]\,\|\boldsymbol{r}_i\|^2$ ($\alpha_i = 1$, $\beta_i = 3$, $\|\boldsymbol{r}_i\|^2 = 1$). (c) Stationary bounds in units of $\sigma^2/2$, with all interferer norms equal to σ and illustrative constants $\alpha = 1$, $\beta = 3$, $\mu_{\max} = 0.1$. The upper bound $\beta(K-1)\big(1 + (K-2)\mu_{\max}\big)$ gives a sufficient task-count limit for a tolerance ε. Here $\varepsilon = 30\,\sigma^2/2$ certifies $K \leq 7$ under these assumptions, although larger merges may still be feasible. The lower bound $\alpha(K-1)\big(1 - (K-2)\mu_{\max}\big)$ is positive for $K < K_{\text{bal}} = 2 + 1/\mu_{\max} = 12$ and becomes vacuous beyond that point. Panels (b) and (c) are schematic, not fitted.

A.1 Proof of Theorem 1 and Corollary 1

Taylor step. Let $\phi(t) = L_i(\boldsymbol{\theta}_i + t\boldsymbol{r}_i)$, which inherits twice continuous differentiability on $[0, 1]$ from L_i. By Taylor's theorem with integral remainder, $\phi(1) - \phi(0) = \phi'(0) + \int_0^1 (1-t)\phi''(t)\,dt$, with $\phi'(0) = \boldsymbol{g}_i^\top \boldsymbol{r}_i$ and $\phi''(t) = \boldsymbol{r}_i^\top \nabla^2 L_i(\boldsymbol{\theta}_i + t\boldsymbol{r}_i)\boldsymbol{r}_i$. Assumption 1 gives $\alpha_i\|\boldsymbol{r}_i\|^2 \leq \phi''(t) \leq \beta_i\|\boldsymbol{r}_i\|^2$, and $\int_0^1 (1-t)\,dt = \frac{1}{2}$, so $\boldsymbol{g}_i^\top \boldsymbol{r}_i + \frac{\alpha_i}{2}\|\boldsymbol{r}_i\|^2 \leq \Delta L_i \leq \boldsymbol{g}_i^\top \boldsymbol{r}_i + \frac{\beta_i}{2}\|\boldsymbol{r}_i\|^2$.

Energy step. Expanding the squared displacement gives $\|\boldsymbol{r}_i\|^2 = A_i + C_i$ with cross-term energy $C_i = 2\sum_{j<k;\, j,k\neq i} \langle \boldsymbol{\tau}_j, \boldsymbol{\tau}_k \rangle$. Writing $a_j = \|\boldsymbol{\tau}_j\|$ and using $|\langle \boldsymbol{\tau}_j, \boldsymbol{\tau}_k \rangle| \leq \mu_{\max} a_j a_k$,

$$|C_i| \leq \mu_{\max}\Big[\Big(\sum_{j\neq i} a_j\Big)^2 - \sum_{j\neq i} a_j^2\Big] \leq \mu_{\max}(K-2)\sum_{j\neq i} a_j^2 = \mu_{\max}(K-2)A_i,$$

where the second inequality follows from Cauchy–Schwarz, $(\sum_{j\neq i} a_j)^2 \leq (K-1)\sum_{j\neq i} a_j^2$. Hence $A_i(1 - (K-2)\mu_{\max}) \leq \|\boldsymbol{r}_i\|^2 \leq A_i(1 + (K-2)\mu_{\max})$. Since $\alpha_i, \beta_i > 0$, substituting these bounds, valid for all K, into the Taylor bounds gives Equation 1. The lower bound is informative only when $(K-2)\mu_{\max} < 1$. The stationary capacity forms also require $1 - (K-2)\mu_{\max} \geq 0$ before the

lower multiplier is applied to $A_i \geq (K-1)\sigma_{\min}^2$. (For $K = 2$, $C_i = 0$ and $\|\boldsymbol{r}_i\|^2 = \|\boldsymbol{\tau}_j\|^2$ exactly.) The stated capacity forms follow from $(K-1)\sigma_{\min}^2 \leq A_i \leq (K-1)\sigma_{\max}^2$.

Direction. For $K = 2$, $\boldsymbol{r}_i = \boldsymbol{\tau}_j$ and $\boldsymbol{r}_j = \boldsymbol{\tau}_i$. At a stationary expert, the mean-value form of the Taylor step gives $\Delta L_i = \frac{1}{2}\int_0^1 2(1-t)\phi''(t)\,dt = \frac{q_i}{2}\|\boldsymbol{\tau}_j\|^2$, where $q_i = \int_0^1 2(1-t)\phi''(t)\,dt/\|\boldsymbol{\tau}_j\|^2 \in [\alpha_i, \beta_i]$; likewise for j. Dividing gives Equation 2. If $\|\boldsymbol{\tau}_j\|^2/\|\boldsymbol{\tau}_i\|^2 > \beta_j/\alpha_i$, then $\Delta L_i \geq \frac{\alpha_i}{2}\|\boldsymbol{\tau}_j\|^2 > \frac{\beta_j}{2}\|\boldsymbol{\tau}_i\|^2 \geq \Delta L_j$. □

Scope. These bounds concern the loss L_i. They do not directly bound task metrics, and their curvature constants are not certified for real networks. Appendix J reports sampled finite-difference curvature only as a diagnostic. Away from stationarity, the gradient term $\boldsymbol{g}_i^\top \boldsymbol{r}_i$ enters both sides, and $|\boldsymbol{g}_i^\top \boldsymbol{r}_i| \leq \|\boldsymbol{g}_i\|\sqrt{A_i(1+(K-2)\mu_{\max})}$ bounds it. For TIES and DARE, $\boldsymbol{r}_i$ is the edited displacement, not $\sum_{j\neq i}\boldsymbol{\tau}_j$, so applying the bounds requires the edited vectors. Our empirical tests of Equation 2 use task arithmetic, for which $\boldsymbol{r}_i = \boldsymbol{\tau}_j$ holds exactly.

A.2 Displacement identities, task-count scaling, and certificates

Displacement identities. For a task subset S of size K merged by unit-coefficient TA, the exact identity $\|\boldsymbol{r}_i\|^2 = A_i + C_i$ and $|C_i| \leq B_i$ hold with the burden numerator $B_i = 2\sum_{j<k;\, j,k\neq i}\mu_{jk}\|\boldsymbol{\tau}_j\|\|\boldsymbol{\tau}_k\|$ of Section 5.4, because each inner product lies between $\pm\mu_{jk}\|\boldsymbol{\tau}_j\|\|\boldsymbol{\tau}_k\|$. Hence

$$\max\{0, A_i - B_i\} \leq \|\boldsymbol{r}_i\|^2 \leq A_i + B_i, \qquad b_i = B_i/A_i \leq (K-2)\mu_{\max}, \tag{6}$$

so b_i provides a task-specific bound at least as tight as the bracket $(K-2)\mu_{\max}$ in Theorem 1. The quantity B_i bounds the cross-term magnitude, not the realized signed cross term, and $b_i = 0$ at $K = 2$. Our multi-task predictors $s_i^{(K)} = \log_{10}(A_i/\|\boldsymbol{\tau}_i\|^2)$ and b_i therefore use the theorem's scale and coherence quantities. Normalizing by the target norm is an empirical design choice that makes $s_i^{(K)}$ reduce to twice the pairwise s_i at $K = 2$.

Task-count scaling. If $\|\boldsymbol{\tau}_j\| \leq \sigma$ and all coherences are at most $\mu_{\max}$, then $A_i + B_i \leq \sigma^2(K-1)[1+(K-2)\mu_{\max}]$: there are $K-1$ squared-norm terms and $\binom{K-1}{2}$ distinct cross terms. The linear and quadratic terms of this envelope balance at $K_{\text{bal}} = 2 + 1/\mu_{\max}$ (infinite when $\mu_{\max} = 0$), which is also where the lower side of Theorem 1 becomes vacuous. The balance point need not correspond to a sharp empirical transition; heterogeneous norms can also make the envelope loose. Under Assumption 1 the bounds bracket each task's loss increase for any K. Certifying the constants for real networks remains open, so we report the bounds and their empirical scale dependence, not a certified $K_{\max}$. The natural maximum pairwise coherence is about 0.027 for Llama and 0.168 for Pythia 6.9B, giving K_{bal} near 38.5 and 7.95. Our K-task sweep uses Llama and shows no onset of a coherence-dominated regime (Appendix C). The lower Pythia balance point remains untested.

Coefficient changes and edited vectors. For a weighted merge of edited vectors $\widetilde{\boldsymbol{\tau}}_j$, the actual displacement is $\boldsymbol{r}_i = \sum_{j\in S}\lambda_j\widetilde{\boldsymbol{\tau}}_j - \boldsymbol{\tau}_i$. This displacement includes any shrinkage or editing of the target's own vector, as well as interference from other tasks. The Taylor step still applies to this displacement when Assumption 1 holds along it, but the energy step and Equation 6, as written, describe unedited unit-coefficient TA. Regressions using pre-edit geometry for TIES or DARE are therefore descriptive diagnostics, not direct tests of the bound.

Loss versus metrics. Experiments use $D_i = M_i(\boldsymbol{\theta}_i) - M_i(\boldsymbol{\theta}_M)$ with higher-is-better task metrics M_i. Accuracy, F1, and correlation are not the differentiable loss L_i, and we do not claim an inequality that converts Equation 1 into a bound on these metrics. The metric-level results of Sections 5.2 and 5.3 test the corollary's directional prediction, not the bound itself; Section 5.1 tests it on the loss.

Conditional success and failure certificates. If the upper side of Equation 1 is at most a tolerance ε_i, then $\Delta L_i \leq \varepsilon_i$; if the lower side exceeds it, the tolerance is violated. Neither implication holds in reverse. A usable certificate requires verified curvature and gradient bounds for the same loss, expert, and path. Symmetric validation-set probes do not establish those constants, so we report bounds and their empirical scale dependence rather than a certified $K_{\max}$.

An upper curvature constant cannot yield a lower bound. A constant loss function is β-smooth for any positive upper bound β, yet its degradation is zero for arbitrarily many task vectors. A positive lower bound proportional only to β, vector norms, and coherence would therefore be invalid. Accordingly, Theorem 1 uses a separate lower constant α_i. Even when lower curvature is available, a negative gradient contribution can offset quadratic growth. The gradient term must therefore be retained for an expert trained for a finite budget or under constraints.

A.3 Loss-level test details

We test every task-arithmetic pair from the pairwise study and the three matched-exposure seeds, excluding pairs containing STS-B because it uses MSE rather than cross-entropy. This leaves 21 Llama, 6 Mistral, 10 Qwen, and 21 Pythia full-data pairs, plus 21 pairs per matched seed and family. For task i, ΔL_i is the merged model's mean validation loss minus the expert's own loss. Expert losses come from individual evaluations for Llama, Mistral, and Qwen, and from the merge evaluator for Pythia and all matched-exposure experts. A pair is correctly ordered when its smaller-norm task has the larger ΔL. Intervals resample family–pair clusters using 5,000 draws and seed 2027.

The log contrast requires both $\Delta L > 0$, leaving 31 of 58 full-data pairs and 41 of 126 matched pairs for the variance decomposition. Scale, curvature, and covariance shares sum to one. In the remaining pairs, at least one target's loss does not rise: 29 of 116 full-data targets (25 larger-norm and 4 smaller-norm; minimum -0.287) and 104 of 252 matched targets. At a stationary expert with $\alpha_i > 0$, the lower bound in Theorem 1 rules out $\Delta L_i \leq 0$. These observations therefore violate its premises. Here L_i is validation loss, and LoRA experts selected by validation score or trained for a fixed budget need not be stationary. We retain all pairs in the ordering test and condition on positive changes only for the variance decomposition.

In Figure 2b, the two lowest-ratio bins contain 2 and 4 full-data pairs alongside 100 and 26 matched pairs. All 52 pairs above 0.15 use full data. Within full-data pairs alone, ordering is correct in 11/14 below 0.3 and 42/44 above. The curvature ratio of about 10 in Section 5.1 is estimated from these same 31 positive-change pairs; it is a descriptive summary, not an independent prediction.

Corollary condition with sampled curvature ranges. We also check the condition using Llama finite-difference probes from Appendix J. These symmetric second differences use directions that were not fitted to the loss contrasts. We substitute each task's sampled minimum $q_{\min}$ for α_i and sampled maximum $q_{\max}$ for β_i. For each of the 21 full-data Llama TA pairs excluding STS-B, we label the smaller-norm task i and check $\|\boldsymbol{\tau}_j\|^2/\|\boldsymbol{\tau}_i\|^2 > \beta_j/\alpha_i$. Eleven pairs satisfy this condition: squared norm ratios range from 44–820, compared with sampled curvature ratios of 9–40. All 11 follow the predicted ordering. The two Llama ordering errors, MRPC with CoLA and QQP with MNLI, have squared norm ratios of 2.2 and 2.0 and fail the condition. In 4 of the 11 qualifying pairs, the larger-norm task's loss does not rise, so the stationary premise also fails there.

This remains a diagnostic. A sampled minimum does not establish a uniform lower bound (Appendix J); agreement under sampled ranges is not a verified instance of the corollary. Table 3 reports the results by family. Mistral has only two pairs with two positive loss changes, so we omit its variance shares. Pythia's scale share exceeds one because its covariance share is negative.

Experts	Family	Pairs	Loss ordering	Median $\lvert \log_{10} \text{ratio} \rvert$	Scale share	Curvature share
Full data	Llama 8B	21	90.5%	0.82	0.55	0.26
Full data	Mistral 7B	6	100.0%	0.44	–	–
Full data	Qwen 7B	10	100.0%	0.77	0.17	0.39
Full data	Pythia 6.9B	21	85.7%	0.91	1.43	0.32
Matched exposure	Llama 8B	63	68.3%	0.02	0.01	0.91
Matched exposure	Pythia 6.9B	63	44.4%	0.03	0.01	1.01

Table 3: Loss-level test of Corollary 1 on task-arithmetic pairs (STS-B excluded). Loss ordering measures how often the smaller-norm task has the larger loss increase. Scale and curvature shares decompose the variance of $\ln(\Delta L_i/\Delta L_j)$ over pairs with both loss changes positive. The remaining share is covariance. Pooled: full data 91.4% of 58 pairs ($[82.8, 98.3]$; scale 0.54, curvature 0.32); matched exposure 56.3% of 126 pairs ($[46.0, 66.7]$; scale 0.007, curvature 0.98).

B Experimental Protocol and Data Provenance

Primary GLUE settings. The eight tasks are MNLI, QQP, QNLI, SST-2, CoLA, STS-B, MRPC, and RTE. Metrics are accuracy for MNLI/QNLI/SST-2/RTE, F1 for QQP/MRPC, Matthews correlation for CoLA, and Pearson correlation for STS-B. Degradation is the individual expert score minus the merged score. Negative degradation and below-baseline scores are retained. The individual score in D comes from direct checkpoint evaluation (Table 5); for Pythia, the merge evaluator's own individual scores differ from it by at most 0.0072. All evaluations use GLUE validation sets. The analysis script performs no training on task labels.

We use LoRA rank 16, alpha 32, and dropout 0.05. Attention and MLP projections are adapted; Pythia uses the GPT-NeoX query-key-value, dense attention, and two MLP projections. The GLUE configuration specifies AdamW, three epochs, learning rate 2×10^{-4}, warmup ratio 0.06, weight decay 0.01, maximum sequence length 128, and BF16 training. Batch size and accumulation vary with the model. The experiments use a common base checkpoint within each family and exclude classification heads from task-vector geometry and merging. At evaluation, we restore the target's trained head. The testbed therefore evaluates task-specific heads, rather than a single multitask classifier. Table 4 lists training exposure and task-vector norms for the Llama and Pythia experts, at full data and at matched exposure.

		Llama 3.1-8B			Pythia 6.9B		
Task	N_{full}	$\|\boldsymbol{\tau}\|$ full	$\|\boldsymbol{\tau}\|$ matched	Steps	$\|\boldsymbol{\tau}\|$ full	$\|\boldsymbol{\tau}\|$ matched	Steps
MNLI	392,702	301.9	$10.8_{10.6–11.4}$	234	1194.7	$17.2_{17.0–17.6}$	468
QQP	363,846	211.4	$9.9_{9.7–10.2}$	234	1261.1	$16.0_{15.9–16.5}$	468
QNLI	104,743	130.9	$9.8_{9.8–10.5}$	234	233.6	$16.5_{16.3–16.7}$	468
SST-2	67,349	81.3	$10.5_{10.1–10.6}$	234	146.7	$14.5_{14.3–14.6}$	468
CoLA	8,551	18.2	$10.0_{9.9–10.1}$	234	36.8	$16.1_{14.9–17.0}$	468
STS-B	5,749	11.1	$8.4_{7.6–8.7}$	234	25.1	$14.5_{13.8–14.9}$	468
MRPC	3,668	12.3	$9.5_{8.5–9.7}$	234	20.8	$15.9_{15.1–16.1}$	468
RTE	2,490	10.5	$10.6_{10.2–10.9}$	234	16.8	$17.5_{16.8–18.0}$	468
max/min	157.7×	28.6×	$1.26\times_{1.25–1.50}$	1.00×	75.3×	$1.23\times_{1.22–1.25}$	1.00×

Table 4: Expert training exposure and task-vector norms for the two families used in the exposure experiments. N_{full} is the full GLUE training-set size and $\|\boldsymbol{\tau}\|$ the norm of the LoRA task vector (classification heads excluded). All experts train for three epochs at learning rate 2×10^{-4} with batch 32 (Llama) or 16 (Pythia). Full-data experts train on the whole training set and keep the best-epoch checkpoint. Matched experts train on 2,490 examples of every task (RTE's full set) and retain the final checkpoint, giving every matched expert in a family the same number of optimizer steps (Steps). Matched norms are medians over three seeds (42, 1, 2), with subsets redrawn for each seed and ranges shown as subscripts. The last row gives max/min over the eight tasks; for matched experts it is computed within each seed and shown as median and range. Mistral and Qwen experts are trained on full data only; Table 5 lists which of them pass QC.

Expert quality control. We retain an expert only if its score exceeds both an absolute task threshold and a margin above the null baseline. The table below gives the resulting thresholds and individual scores. This selection limits what the study can say about failed or near-null experts. Averages over different valid task sets are not equal-difficulty cross-family comparisons. We did not diagnose the excluded experts. Mistral's MNLI, QQP, and QNLI experts (0.370, 0.542, 0.563) lie within 0.06 of the constant-predictor scores on the same labels (0.354, 0.538, 0.505), and Qwen's MNLI and QNLI experts (0.527, 0.617) fall well below threshold. We have not tested whether family-specific learning rates or longer training would rescue these experts, and our conclusions do not extend to them.

Regression and resampling. We construct the primary table by joining stored individual evaluations, task-vector measurements (norms and coherences), and merge evaluations. For each family, valid pairs are evaluated in both directions and with all three methods. Percentile bootstrap intervals use 5,000 resamples of 81 pair groups, drawn with replacement using analysis seed 2027. It conditions on a fixed collection of experts. All quoted R^2 values are in-sample explanatory summaries; held-out prediction is separately evaluated by RMSE and skill in Appendix B.3. Directional evaluations share

Family	MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE
Threshold	0.750	0.588	0.750	0.800	0.300	0.700	0.862	0.627
Llama 8B	0.902	0.869	0.953	0.966	0.668	0.920	0.927	0.892
Mistral 7B	0.370	0.542	0.563	0.948	0.689	0.918	0.938	0.881
Qwen 7B	0.527	0.686	0.617	0.970	0.690	0.895	0.900	0.845
Pythia 6.9B	0.781	0.670	0.922	0.947	0.598	0.898	0.883	0.809

Table 5: QC thresholds and individual expert primary metrics, from Trainer evaluations before QC exclusions. Llama and Pythia pass all eight tasks; Mistral excludes MNLI/QQP/QNLI, and Qwen excludes MNLI/QNLI.

experts and cannot be counted as independent experiments. We do not combine p-values across analyses into a single test.

Pool	N	R^2_μ	R^2_s	$\hat{b}_s$ (cluster SE)	$\hat{b}_\mu$ (cluster SE)
All	486	0.003	0.284	+0.175 (0.022)	-0.808 (0.584)
DARE	162	0.017	0.104	+0.103 (0.028)	-1.902 (0.965)
TA	162	0.001	0.416	+0.207 (0.024)	-0.361 (0.442)
TIES	162	0.000	0.444	+0.215 (0.024)	-0.162 (0.455)

Table 6: Pairwise results computed from stored individual scores, task geometry, and merged metrics. Coefficients are fitted with $D \sim s + \mu$, and standard errors cluster by family–pair (shared-expert dyadic SEs are in Table 7). All method rows contain the same 81 groups. The two-predictor pooled R^2 is 0.287. R^2_s and R^2_μ refer to different variance components (Section 5.2).

Pooled regressions and sensitivity. Table 6 and Figure 5 give the pooled and per-method fits. Pair-group bootstrap intervals are $[0.135, 0.220]$ for the pooled scale coefficient and $[-6.309, -0.144]$ for coherence. Although the coherence interval excludes zero, the pair-cluster and dyadic tests are not significant ($p = 0.17$ and 0.41) and the expert-bootstrap interval ($[-9.660, 0.530]$) includes zero. With family, method, and task fixed effects, R^2 rises to 0.416. Leaving out each family keeps the scale coefficient between 0.153 and 0.210. Standardizing degradation within each family–target-task cell and refitting with the same fixed effects gives a scale coefficient of 0.867 (cluster $p < 0.001$; dyadic SE 0.122, $p < 0.001$). For a pair evaluated in both directions, s_i changes sign while μ stays fixed, making their pooled covariance zero in the balanced design. The two univariate R^2 values therefore sum almost exactly to the two-predictor value.

Dependence-robust inference. The supplementary code computes these intervals after recomputing all reused point estimates from stored evaluations; all 85 checks agree to floating-point precision. Different pairs can share an expert, a dependence that pair clustering misses. Table 7 therefore adds shared-expert (dyadic) standard errors, an Aronow–Samii–Assenova sandwich (Aronow et al., 2015) over the 27 experts with a small-sample factor and a t_{26} reference, and an expert-resampling (vertex) bootstrap. The method-specific scale slopes, with dyadic uncertainty, are 0.207 (SE 0.030, $p < 0.001$) for TA, 0.215 (0.030, $p < 0.001$) for TIES, and 0.103 (0.048, $p = 0.040$) for DARE. By family they are 0.154 (0.037, $p = 0.004$) for Llama, 0.324 (0.035, $p < 0.001$) for Qwen, and 0.142 (0.038, $p = 0.007$) for Pythia. With only five experts, Mistral has a non-positive dyadic variance estimate, which we report as undefined. The vertex bootstrap gives $[0.104, 0.251]$ for the pooled scale slope. Resampling experts can be conservative for dyadic data, whereas resampling pair groups ignores dependence through shared experts and can be anti-conservative, so we report both.

For held-group RMSE differences, we refit every fold in every resample. Conditioning on the saved predictions instead would make the held-family steps-plus-scale-minus-steps gap appear resolved ($[0.0009, 0.0141]$), which is not robust to refitting ($[-0.0021, 0.0285]$).

Coherence coefficient by family. Within-family pair coherences span 0.0011–0.0274 (Llama), 0.0020–0.0233 (Mistral), 0.0009–0.0251 (Qwen), and 0.0002–0.168 (Pythia). All 53 pairwise signed cosines for Llama, Mistral, and Qwen are positive, so μ equals the signed cosine there. The pooled fixed-effects coefficient is -1.575 (pair-cluster SE 0.448, dyadic SE 0.582, $p = 0.012$). Per-family coefficients from $D \sim \mu + s$ are -11.65 ($p = 0.008$), -11.83 ($p = 0.019$), -11.51 ($p = 0.001$), and $+0.37$ ($p = 0.46$) for Llama, Mistral, Qwen, and Pythia. Over each family's observed μ range,

Regression slopes (SE, p); pair groups G=81, experts G=27			
Quantity	Estimate	Pair-cluster	Shared-expert (dyadic)
Pooled $D \sim s + \mu$: $\hat{b}_s$	+0.175	0.022 (<.001)	0.031 (<.001)
Pooled $D \sim s + \mu$: $\hat{b}_\mu$	-0.808	0.584 (0.171)	0.959 (0.407)
With fixed effects: $\hat{b}_s$	+0.252	0.026 (<.001)	0.036 (<.001)
With fixed effects: $\hat{b}_\mu$	-1.575	0.448 (<.001)	0.582 (0.012)
Ordering accuracy, % (95% CI)			
Rule	Estimate	Pair bootstrap	Expert bootstrap
Smaller norm	83.5	[77.4, 88.9]	[73.5, 92.0]
Fewer examples	83.5	[77.4, 89.3]	[73.7, 92.5]
Fewer steps	84.8	[79.0, 90.1]	[75.6, 93.0]
Steps − norm	+1.2	[-3.7, +7.4]	[-7.4, +10.1]
Contrast RMSE difference $\times 10^3$, refitted per resample (95% CI)			
Comparison	Estimate	Pair bootstrap	Expert bootstrap
Held family: scale − data	+3.3	[-11.2, +19.3]	[-24.2, +32.4]
Held family: scale − zero	-53.5	[-79.3, -25.2]	[-100.1, +1.6]
Held family: data − zero	-56.9	[-83.4, -29.7]	[-102.7, -4.8]
Held task: scale − data	+1.1	[-7.6, +9.1]	[-16.9, +19.5]
Held task: scale − zero	-49.9	[-76.3, -23.7]	[-97.2, +0.6]
Held task: data − zero	-51.0	[-76.9, -25.6]	[-96.1, -4.4]
External screening: rank of the selected 3/6 Llama set among all 20 subsets			
Merge	Gain (pp)	Rank / 20	P(subset $\leq$ selected)
TA	+3.02	9	0.45
TIES	+12.81	6	0.30
DARE	+5.43	9	0.45
Mean of three	+7.09	9	0.45

Table 7: Dependence-robust inference. Pair-cluster SEs group the six evaluations of a family–pair. Shared-expert (dyadic) SEs also allow correlation between any evaluations that share one of the 27 experts (t_{26} reference). Bootstrap intervals use 5,000 draws (seed 2027) that resample family–pair groups or, more conservatively, experts within each family. Each RMSE-difference resample refits every held-out fold. Negative differences favor the first predictor. The external rank is the selected Llama set's position among all $\binom{6}{3} = 20$ candidate sets (1 = best); the last column is the fraction of sets at least as good.

these imply a change of at most 0.31 in D. Adding method and task fixed effects within each family gives $p = 0.10$, 0.57, 0.14, and 0.31.

In the first three families the three most aligned pairs are all among RTE, STS-B, and MRPC, the smallest-data and smallest-norm experts, with low pair-mean damage (e.g., Llama RTE+STS-B: $\mu = 0.027$, $m = 0.035$). Within each family, μ correlates with the pair's mean log norm at -0.67 to -0.93. Adding the pair's mean log norm gives -5.76 ($p = 0.24$), $+3.66$ ($p = 0.51$), -7.44 ($p = 0.20$), and -0.01 ($p = 0.99$).

In the pooled fixed-effects fit, Pythia supplies 84.7% of the residualized μ variance and the single Pythia MNLI+QQP pair ($\mu = 0.168$) supplies 71.3%, but this outlier does not account for the coefficient: dropping the pair gives -3.17 (SE 1.27) and dropping Pythia gives -10.67 (SE 3.42). It survives controls for the pair's mean log norm (-1.84, dyadic $p = 0.032$) and mean log training size (-1.56, dyadic $p = 0.013$), while 91.2% of expert-bootstrap draws are negative and the interval $[-9.6, 3.6]$ includes zero (in 1,623 of 5,000 draws a task or family level is absent and its unidentified dummy is dropped). We therefore interpret the coefficient as an association that depends on the composition of the expert collection.

Metric-homogeneous outcomes. We compute null scores from the stored labels for the full validation sets (MNLI 0.3545, QQP F1 0.5382, QNLI 0.5054, SST-2 0.5092, MRPC F1 0.8122, RTE 0.5271; zero for CoLA and STS-B). After expressing degradation as $D/(M_{\text{expert}} - M_{\text{null}})$, the norm rule orders 84.0% of 243 contrasts ($[77.8, 89.7]$) with contrast share 58.9%. Relative to raw D, 19 orderings change, and the scale slope is 0.778 (dyadic $p = 0.004$). Clipping at the null gives 88.1% of 227 contrasts ($[81.4, 93.9]$; share 71.8%), relative degradation D/M_{expert} gives 84.0% ($[77.8, 89.7]$; share 65.1%), and standardizing within family–task gives 74.9% ($[67.5, 81.9]$; share 45.9%; data and step rules 75.7% and 77.0%).

On accuracy-only pairs (MNLI, QNLI, SST-2, RTE; 42 contrasts in 14 groups), raw ordering is 81.0% ($[72.2, 89.7]$; share 50.5%) and chance-adjusted ordering 85.7% (36/42; $[76.2, 94.4]$; share 61.3%). On same-metric pairs the norm rule orders 82.4% raw, 86.3% chance-adjusted, and 93.3% clipped. The 9 of 486 rows with $D > 1$ all come from STS-B merges with negative Pearson correlation.

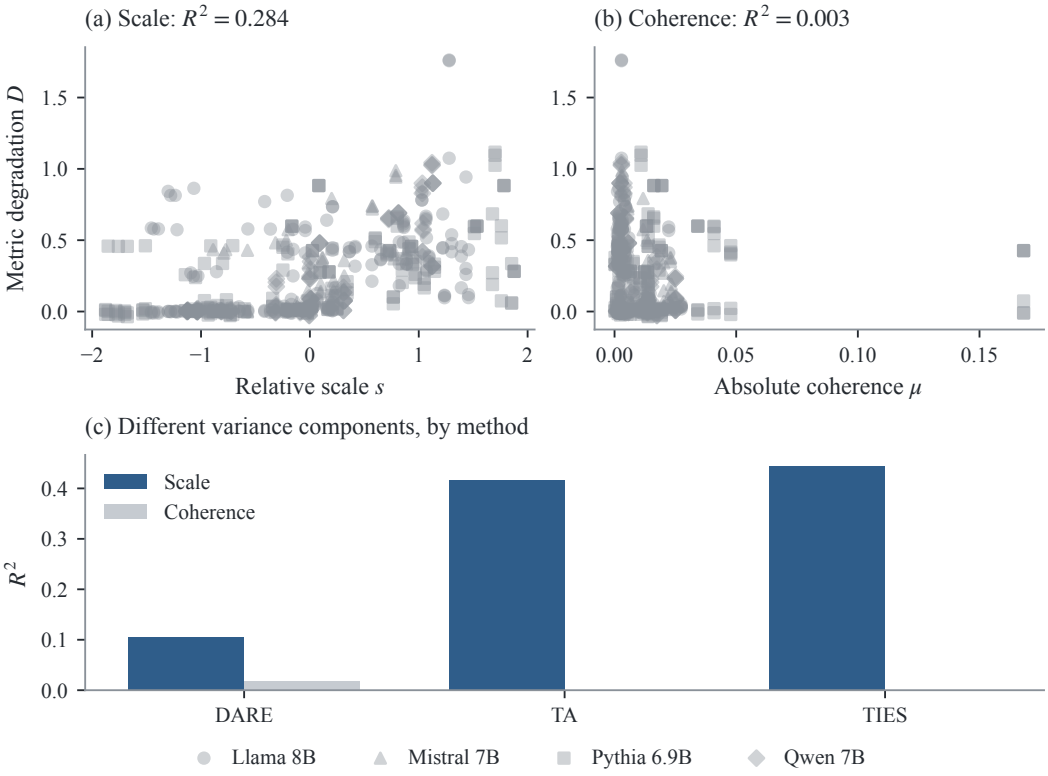

Figure 5: Directional degradation pooled across four families. The univariate fits give $R^2_s = 0.284$ and $R^2_\mu = 0.003$, but address different components of Equation 4. The antisymmetric predictor s explains 44.3% of the within-pair contrast, which accounts for 64.1% of total variance. The symmetric predictor μ explains 0.8% of pair-mean variance, which accounts for the remaining 35.9%.

B.1 Directional contrast prediction

We form 243 pair–method contrasts from the 486 directional observations. Alphabetical task ordering defines a and s in Section 4; the quality feature is $q = (M_i(\boldsymbol{\theta}_i) - M_j(\boldsymbol{\theta}_j))/2$. Four prespecified models predict a as 0, $b_q q$, $b_s s$, or $b_q q + b_s s$. All have zero intercept, so swapping tasks reverses the prediction. Coefficients are fitted by OLS on training folds only, with no tuning, clipping, or outcome normalization. This analysis uses the same pairwise study as the main regressions.

Held-family folds train on the other families and test each pair once. Held-task folds exclude every pair containing the held task from training and test all such pairs. Each pair is therefore tested twice across the eight task folds, once for each member. The resulting 486 fold-level predictions represent 243 unique pair–method combinations, rather than 486 independent pairs. Table 8 reports all four models; the bundle contains their splits, coefficients, and predictions.

The fixed norm rule predicts that the smaller-norm expert loses more. Fitted scale slopes are positive in every fold, so the fitted scale predictor and this untrained rule give the same task ordering. Their equal accuracy reflects this shared ordering and provides no separate validation. Exact outcome ties are excluded from ranking, and tied predictions score 0.5. There are no outcome ties in the pairwise study. The expert-score-gap predictor yields held-family ordering accuracy of 50.2% overall and 45.1% on same-metric pairs. We define the same-metric subset before running the analysis: accuracy

for MNLI/QNLI/SST-2/RTE and F1 for MRPC/QQP; CoLA and STS-B have no same-metric partner. It contains 51 pair–method rows from 17 family–pair groups. We evaluate the existing predictors on this subset without refitting them.

Scope	Predictor	Family RMSE	Task RMSE	Family order	Task order
All	Equal degradation	0.237	0.237	50.0%	50.0%
All	Expert-score gap	0.239	0.275	50.2%	44.9%
All	Relative scale	0.183	0.187	83.5%	83.5%
All	Gap + scale	0.184	0.203	81.5%	77.8%
Same metric	Equal degradation	0.206	0.206	50.0%	50.0%
Same metric	Expert-score gap	0.212	0.225	45.1%	47.1%
Same metric	Relative scale	0.178	0.183	82.4%	82.4%
Same metric	Gap + scale	0.182	0.193	82.4%	82.4%

Table 8: Geometry and expert-quality contrast predictors. RMSE is measured on $a = (D_i - D_j)/2$ in raw metric units. Ordering accuracy is the fraction of correctly ranked task pairs, with predicted ties assigned 0.5. Held-task summaries count the two task folds per pair explicitly as described above.

The fitted zero-intercept scale predictor lowers contrast RMSE from the zero-contrast baseline of 0.2367 to 0.1832 across held families and to 0.1868 across held tasks. Refitted pair-bootstrap intervals resolve both the held-family gain (-0.054, $[-0.079, -0.025]$) and the held-task gain (-0.050, $[-0.076, -0.024]$), but both expert-bootstrap intervals include zero. On same-metric pairs, scale's held-family RMSE (0.1781) is also below zero contrast (0.2059). Relative scale reduces pooled squared error against the quality-gap baseline by 41.1% for held families and 53.8% for held tasks. On same-metric pairs these gains are 29.7% and 34.2%, and scale also has lower RMSE than the equal-degradation baseline in all four comparisons.

Correctly identifying the loser does not ensure an accurate estimate of the gap. Comparisons across different primary metrics remain a descriptive diagnostic, and shared experts prevent interpreting the ordering counts as independent Bernoulli trials. For the norm rule, the same-metric ordering accuracy (82.4%, 17 groups) has pair-group interval $[74.1, 90.5]$ and expert interval $[70.4, 97.0]$; accuracy-only pairs (14 groups) give 81.0% ($[72.2, 89.7]$ and $[66.7, 100.0]$).

B.2 Training-exposure comparison

The cached GLUE metadata give training-set sizes of 392,702 (MNLI), 104,743 (QNLI), 363,846 (QQP), 67,349 (SST-2), 8,551 (CoLA), 5,749 (STS-B), 3,668 (MRPC), and 2,490 (RTE). We obtain selected optimizer steps from each expert's final Trainer state, which identifies the best checkpoint loaded before the expert was saved. Selected epochs range from one to three. The analysis bundle includes all 27 records and hashes of the supporting dataset metadata and Trainer states. These metadata provide observational exposure proxies; training was not randomized in this comparison.

For the contrast sensitivity analysis, we add four zero-intercept models: d, $d + s$, ω, and $\omega + s$. They reuse the exact splits, outcomes, and same-metric subset of Appendix B.1. Table 9 reports all eight models without tuning. The comparison is exploratory. The directional-outcome comparison adds method-conditioned d, $d + s$, ω, and $\omega + s$ to the eight base specifications and leaves every base split and prediction unchanged. Both comparisons are retrospective rather than prospective validation.

The untrained data-size and scale rules have the same aggregate accuracy but disagree on 4/81 family–pair orderings. Equal aggregate accuracy therefore does not imply identical predictions. Selected steps order 206/243 contrasts correctly (84.8%; steps minus norm $+1.2$ points, $[-3.7, +7.4]$). All three rules achieve 42/51 (82.4%) on the same-metric subset. Data-only contrast RMSE is 0.1799 for held families and 0.1857 for held tasks, both below scale's point estimates. However, the refitted intervals for the difference include zero. Because the three methods reuse experts, the ordering counts are not independent Bernoulli trials.

B.3 Held-group prediction protocol and failure cases

This analysis uses the 486-row pairwise table reconstructed from stored evaluations. Within each fold, we fit OLS with an intercept, method indicators, and the specified continuous predictors; the context/quality specifications also use the available context indicators. There are no interactions, regularization, tuned thresholds, outcome normalization, or prediction clipping. Continuous predictors

Predictor	All: family	All: task	Same: family	Same: task
Equal degradation	0.2367	0.2367	0.2059	0.2059
Expert-score gap	0.2387	0.2749	0.2125	0.2255
Relative scale	0.1832	0.1868	0.1781	0.1829
Score gap + scale	0.1839	0.2033	0.1820	0.1928
Training-data ratio	0.1799	0.1857	0.1821	0.1871
Data ratio + scale	0.1829	0.1870	0.1879	0.1882
Selected-step ratio	0.1797	0.1867	0.1770	0.1872
Step ratio + scale	0.1873	0.1889	0.1884	0.1878

Table 9: Contrast RMSE for every exposure sensitivity specification. "All" includes every pair–method contrast; "Same" evaluates pairs sharing a primary metric without refitting. Columns give held-family and held-task prediction. On all pairs, training-data and selected-step ratios have lower RMSE point estimates than relative scale; on same-metric pairs, scale is below the data ratio in both columns and below the step ratio for held tasks. Adding scale increases RMSE for both exposure predictors in every column. Refitted pair-bootstrap intervals for these differences ($\times 10^3$; family, task) include zero: data+scale minus data $+3.1$ $[-4.7, +18.7]$ and $+1.3$ $[-4.4, +8.5]$; steps+scale minus steps $+7.5$ $[-2.1, +28.5]$ and $+2.1$ $[-4.8, +9.8]$; scale minus steps $+3.5$ $[-9.7, +19.0]$ and $+0.1$ $[-9.9, +9.1]$. Table 7 gives the scale-minus-data comparison.

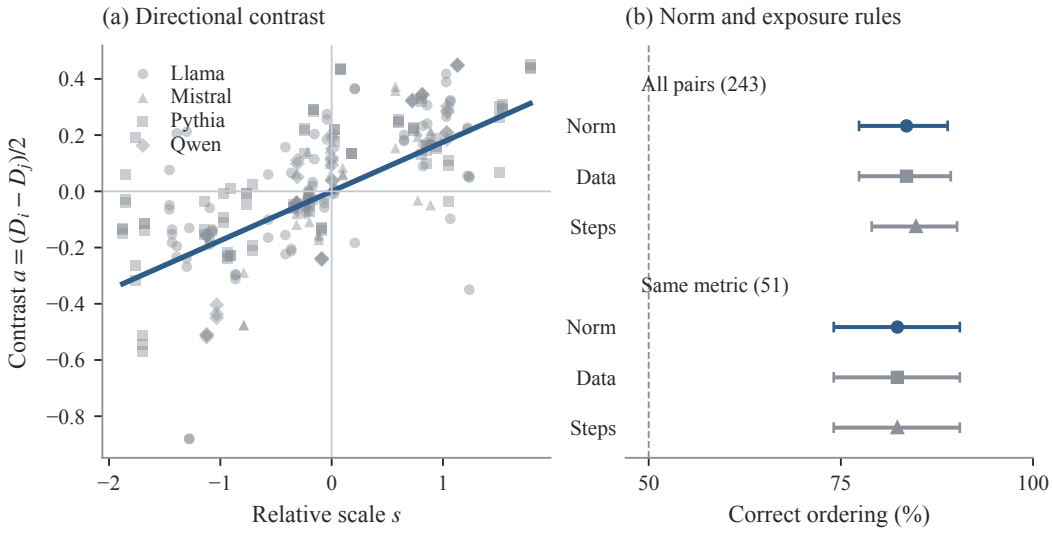

Figure 6: Norm and training exposure predict the loser with similar accuracy. (a) Each point is one pair–method contrast, with a pooled descriptive fit shown by the line. (b) Three untrained rules predict greater degradation for the expert with the smaller norm, training set, or selected checkpoint step count (pair-bootstrap 95% intervals: $[77.4, 88.9]$, $[77.4, 89.3]$, $[79.0, 90.1]$). The all-pair analysis combines different primary metrics; the same-metric subset contains accuracy–accuracy and F1–F1 pairs. Horizontal bars show pair-bootstrap 95% intervals for both subsets; the dashed line scores an equal-degradation prediction as a tie (50%). The counts include repeated evaluations across methods and do not represent independent expert pairs.

are standardized using training means and standard deviations, and categorical encodings are fitted on training data. The richer baseline requires the individual expert's evaluation score and therefore is not label-free. QC selection is inherited from the experimental protocol.

For a held task, training excludes *every pair containing that task*, including observations evaluated on the other task; testing uses only the held target's observations. Each of the 486 rows is predicted once per scheme. The simple baseline predicts the training mean for each merging method; a richer context/quality baseline adds the individual expert's score and indicators for available context (target task in the held-family scheme, family in the held-task scheme).

Held-family train/test row counts are 318/168 (Llama), 426/60 (Mistral), 318/168 (Pythia), and 396/90 (Qwen). In held-task folds, removing both directions of all affected pairs gives $486-2n_t$ training rows and n_t test rows, where n_t is in Table 13. We audit each split to confirm that training and test share no family–pair, that training excludes every held-task expert, and that each scheme and specification produces exactly one prediction per row. Folds still share other experts and evaluation examples, so they are not independent replications. We report descriptive errors without fold-independence significance tests.

Predictor	Family RMSE	Skill$_M$	Task RMSE	Skill$_M$
Method means	0.292	+0.000	0.288	+0.000
Method $+s$	0.250	+0.264	0.256	+0.213
Method $+d$	0.248	+0.279	0.257	+0.207
Method $+d+s$	0.250	+0.265	0.258	+0.198

Table 10: Held-group prediction of directional degradation D. Skill$_M = 1 - \sum(D - \widehat{D})^2 / \sum(D - \widehat{D}_M)^2$ measures improvement over method means estimated on training data; it differs from in-sample R^2. s: relative scale; d: training-data ratio. Refitted pair-bootstrap intervals for RMSE differences from method means: $+s$ $[-0.061, -0.020]$ (family) and $[-0.054, -0.011]$ (task); $+d$ $[-0.064, -0.023]$ and $[-0.052, -0.011]$. Step-ratio, coherence, and context/quality rows are in Tables 11 and 12.

Predictor	Family RMSE	Skill$_M$	Task RMSE	Skill$_M$
Method means	0.292	+0.000	0.288	+0.000
Method $+s$	0.250	+0.264	0.256	+0.213
Method $+d$	0.248	+0.279	0.257	+0.207
Method $+d+s$	0.250	+0.265	0.258	+0.198
Method $+\omega$	0.248	+0.279	0.258	+0.201
Method $+\omega+s$	0.253	+0.247	0.259	+0.195

Table 11: All exposure specifications for held-group prediction of D, extending Table 10 with the selected-step ratio ω.

Predictors	Family RMSE	Skill$_M$	Task RMSE	Skill$_M$
Method means	0.292	+0.000	0.288	+0.000
Method $+s$	0.250	+0.264	0.256	+0.213
Method $+\mu$	0.360	-0.523	0.300	-0.081
Method $+s+\mu$	0.327	-0.259	0.269	+0.128
Context/quality	0.307	-0.105	0.313	-0.182
Context/quality $+s$	0.250	+0.267	0.261	+0.181
Context/quality $+\mu$	0.573	-2.856	0.326	-0.280
Context/quality $+s+\mu$	0.323	-0.224	0.276	+0.081

Table 12: Additional directional-outcome specifications, sharing the method and method-plus-scale baselines with Table 10. Context/quality adds available context and the individual expert score. Together with Tables 10 and 11, this reports all twelve specifications.

Adding relative scale s to method intercepts reduces pooled squared error by 26.4% across families and 21.3% across tasks (refitted pair-bootstrap RMSE-difference intervals exclude zero, as do those for adding d; Table 10). Performance improves for all four families and six of eight tasks, but worsens for MNLI and RTE. Method means plus ω attain RMSE 0.248 and 0.258, and adding s worsens them to 0.253 and 0.259 (Table 11). Scale improves the context/quality baseline, whereas adding coherence to scale worsens pooled RMSE.

For method $+s$, pooled MAE falls from 0.234 to 0.187 in held-family prediction and from 0.230 to 0.192 in held-task prediction. Equal-fold mean RMSE also falls, from 0.289 to 0.244 and from 0.266 to 0.240, respectively. However, equal-fold mean *relative squared-error skill* is $+0.282$ held-family but only $+0.011$ held-task. The distinction matters because pooled skill weights each group's relative improvement by its baseline squared error, while equal-fold skill gives every group equal weight. Thus, improvement on six task folds does not remove the failures on MNLI and RTE.

Held out	N	Method	$+s$	Context/quality	$+s$	$+s+\mu$
Llama	168	0.309	0.280	0.366	0.272	0.271
Mistral	60	0.257	0.219	0.261	0.218	0.216
Pythia	168	0.267	0.230	0.260	0.248	0.428
Qwen	90	0.322	0.247	0.294	0.229	0.227
CoLA	69	0.293	0.212	0.401	0.235	0.231
MNLI	42	0.215	0.287	0.215	0.286	0.426
MRPC	69	0.320	0.284	0.328	0.287	0.286
QNLI	42	0.202	0.141	0.201	0.144	0.145
QQP	57	0.266	0.241	0.313	0.253	0.252
RTE	69	0.150	0.235	0.150	0.234	0.236
SST-2	69	0.216	0.130	0.221	0.138	0.139
STS-B	69	0.465	0.391	0.468	0.391	0.390

Table 13: Prediction RMSE for each held-out group. The first $+s$ column extends method means; the final two extend context/quality. All rows and all twelve specifications, including the exposure specifications, are available in the analysis bundle.

The Pythia fold requires extrapolation: 48/168 test observations have s outside the training range and 36/168 have coherence outside it. Its maximum coherence is 0.168, versus 0.027 in the training families, and adding μ to context/quality$+s$ raises its RMSE from 0.248 to 0.428 (Table 13). This range check describes the fold but does not establish the cause of its error. We retain all observations and do not refit models in response. The pairwise study contains only four families and eight task definitions, and the hypotheses were informed by that study. The study therefore does not establish transfer to a fresh benchmark, accurate predictions for individual tasks, or better selection of merging methods.

B.4 Evaluation collections and matching

The multi-task analysis uses the stored Llama $K = 2$–8 evaluation records. We recompute geometry from the stored task-vector norms and coherences, then verify for every target that its individual baseline equals merged score plus degradation. There are 1,016 unique subset–target rows, with no missing members of the complete $K = 2$–8 enumeration.

The task-count sweep constructs each merged backbone from the fixed base weights and the selected task vectors using float32 accumulation, then casts to the evaluation model dtype. It restores the trained head for the target task. Geometry is held fixed across task counts. The $K = 3$ method comparison selects the TA rows for exactly the same 15 subsets evaluated by TIES and DARE. Calibration comparisons match the task set and evaluation configuration within that experiment. We analyze each evaluation collection separately so that method comparisons do not mix scores from different experimental settings.

Analysis checks. The analysis script checks row uniqueness, sample counts, metric-baseline consistency for every experiment, and finite-difference recomputation. Tables and plots are generated directly from result inputs, whose SHA-256 hashes are recorded in a manifest. The bundle reconstructs the numerical summaries from stored evaluations; model checkpoints and datasets are not included.

C Multi-task Results

We evaluate unit-coefficient TA on every subset of the eight Llama tasks, with $K = 2, \ldots, 8$. There are $28, 56, 70, 56, 28, 8, 1$ subsets at these sizes and 1,016 task evaluations. All subsets reuse the same eight experts, so they are not independent task collections. We use the same evaluation implementation at every task count, retaining the fixed base model and each target expert's classification head (Appendix B.4).

Mean degradation and mean worst-task degradation increase with K (Figure 7; Table 14 lists the full sweep). At $K = 5$, mean worst-task degradation is 1.009; at $K = 8$, it is 1.630. Values above one are possible because Pearson and Matthews correlations can become negative. We fit scale and coherence burden separately at each K. A pooled rise in degradation with K would not establish

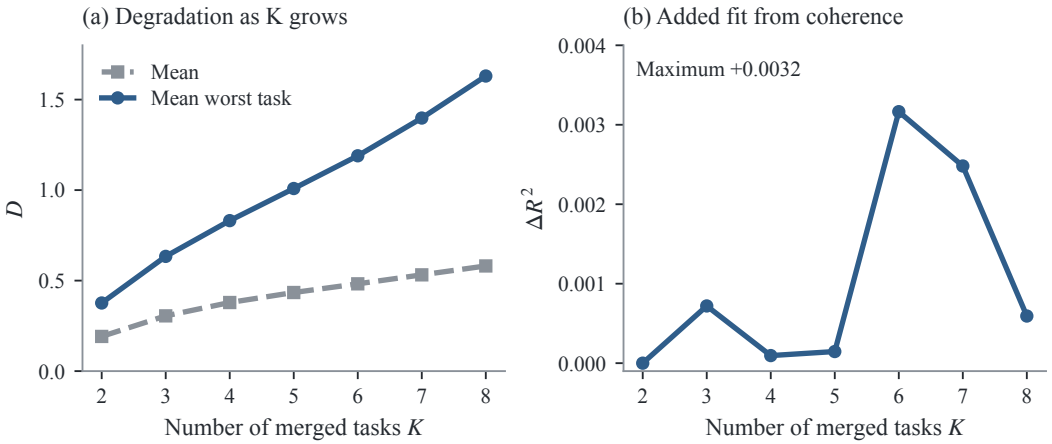

Figure 7: Exhaustive $K = 2$–8 Llama TA sweep. For each subset, we take the largest degradation across its targets, then average these maxima across subsets. Added R^2 is from adding the coherence burden $b_i = B_i/A_i$ to $D_i \sim s_i^{(K)}$ separately for each K. The curves are descriptive: subsets overlap, $K = 8$ has only one subset, and many large-K values are collapse-floor values (see text).

a causal effect of coherence. Scale-only R^2 ranges from 0.130 to 0.370. Adding coherence burden contributes at most 0.0032. We find no evidence here for the onset of a coherence-dominated regime.

Family	K	Subsets	Mean D	Mean worst D	R_s^2	ΔR_b^2
Llama 8B	2	28	0.192	0.376	0.370	0.000
Llama 8B	3	56	0.305	0.634	0.285	0.001
Llama 8B	4	70	0.379	0.831	0.220	0.000
Llama 8B	5	56	0.434	1.009	0.175	0.000
Llama 8B	6	28	0.482	1.190	0.145	0.003
Llama 8B	7	8	0.532	1.398	0.133	0.002
Llama 8B	8	1	0.581	1.630	0.130	0.001

Table 14: Complete Llama TA sweep. Mean worst D averages the maximum degradation in each subset. The additional R^2 from burden is descriptive. With only one subset at $K = 8$, a cluster test is unavailable. At $K = 2$, $b_i = 0$ identically.

Collapse-floor censoring. We classify a merged score as a floor value if it is at or below the best constant-predictor score on the same labels (majority accuracy; all-positive F1 for QQP and MRPC; zero for MCC and Pearson). In the pairwise pool, 149 of 486 evaluations (30.7%) are floor values (Llama DARE 51.8%, Pythia DARE 46.4%, Pythia TA 30.4%, Llama TA 12.5%). In the Llama sweep, 303 of 1,016 (29.8%) are floor values. By $K = 2, \ldots, 8$ the fractions are 12.5, 22.0, 27.1, 32.5, 37.5, 44.6, and 50.0% (none exactly a constant prediction). Floor values account for 78.6% of worst-task outcomes at $K = 5$ and 100% at $K = 8$. Mean worst-task degradation at large K therefore mainly measures collapse below chance.

Restricting the analysis to non-floor rows changes the scale-only R^2 at $K = 2, \ldots, 7$ from 0.370, 0.285, 0.220, 0.175, 0.145, and 0.133 to 0.481, 0.361, 0.213, 0.077, 0.003, and 0.072 (at $K = 8$, 4 rows remain). At $K \geq 5$ the scale signal therefore comes mostly from which tasks collapse. We did not refit the burden increment on this restricted set. In the pairwise pool, the scale slope on the 337 non-floor rows is 0.135 (dyadic SE 0.023, $p < 0.001$), and on the 110 contrasts with no floor row the norm, data, and step rules order 87%, 87%, and 85%.

Other merging methods at $K = 3$. To compare merging methods, we select 15 Llama subsets spanning low, middle, and high full-subset coherence burden. Evaluating TIES and DARE on all three targets gives 90 evaluations. We compare to the 45 TA rows for exactly these subsets from the

exhaustive sweep. For this comparison, the 56 possible Llama subsets are ranked by the full-subset coherence burden, and five each are selected from the low, middle, and high strata. This selection defines a stratified geometric stress test rather than a random task sample. Scale-only R^2 is 0.297 for TA, 0.349 for TIES, and 0.033 for DARE; burden adds 0.0001, 0.0008, and 0.0871, respectively. The DARE burden coefficient is not significant under subset clustering ($p = 0.130$), but the larger descriptive increment limits how far the TA result can be generalized across methods (Table 15).

Method	Mean D	$\hat{b}_s$	p_s	R^2_s	ΔR^2_b
DARE	0.404	0.028	0.260	0.033	0.087
TA	0.236	0.078	$< .001$	0.297	0.000
TIES	0.241	0.087	$< .001$	0.349	0.001

Table 15: Exactly matched 15-subset Llama $K = 3$ comparison. The scale-only fit supplies the scale coefficient and its cluster p; the additional R^2 measures the contribution of burden. Subset clusters still share experts.

D Exposure-Matched Experiments

Training and metadata. Retraining retains each family's full-data configuration: LoRA rank 16, alpha 32, dropout 0.05 on all attention and MLP projections; AdamW for three epochs at learning rate 2×10^{-4}, warmup ratio 0.06, weight decay 0.01, maximum length 128, BF16; batch 32 (Llama) or 16 (Pythia) without accumulation.

We change two aspects of training. First, we shuffle the GLUE training split with a subset seed and retain its first N examples. The 2,490-, 10,000-, and 40,000-example subsets are therefore nested within a seed. Second, we omit intermediate evaluation and checkpoint selection, saving the final checkpoint instead.

Saved metadata for all 56 retrained experts confirm three epochs, final-checkpoint selection, the requested number of examples, and 234 (Llama) or 468 (Pythia) optimizer steps for every one of the 48 equal-exposure experts. The dose experts take 939 and 3,750 steps (Llama, 10,000 and 40,000 examples) and 1,875 and 7,500 (Pythia). Seed 42 uses subset seed 42 and the configuration's training seed 42; seeds 1 and 2 set both the subset and the training seed. All 48 equal-exposure experts pass the QC thresholds (Table 5) when scored by the merge evaluator; the smallest margin is Pythia MRPC seed 1 (F1 0.878 versus 0.862).

Merging and evaluation. We reuse the pairwise study's task-vector extraction, merge operators (unit-coefficient TA; TIES retaining the top 20%; DARE at $p = 0.9$ with seed 42), merged-state construction, and evaluator. Two experts are loaded at a time. We score both targets on the full GLUE validation sets (sizes in Appendix E) using their own trained heads. Degradation D is measured against each expert's individual score from the same evaluator; repeated individual evaluations agree exactly. We recompute norms and coherences from the loaded vectors. Each complete seed has 28 pairs $\times$ 3 methods $\times$ 2 targets $= 168$ merge evaluations (84 contrasts). All three seeds are complete for both families. The dose experiment adds 108 Llama and 108 Pythia evaluations; its 2,490-example level reuses the seed-42 merges. The shrink-only experiment adds 32 evaluations per family (4 pairs $\times$ 4 values of $c \times$ 2 targets).

Per-seed results. Table 16 compares norm spread, ordering, and within-pair scale R^2 between the full-data and matched experts. Table 17 gives each seed's mean degradation, contrast share, ordering by method, a one-sided cluster sign-flip test of above-chance ordering (20,000 random sign flips of each pair's scale), and an exact expert-level permutation test that reassigns the eight norms to tasks in all $8! = 40{,}320$ ways. The sign-flip test treats the 28 pairs as independent although their norms come from eight experts, so it is anti-conservative. We therefore base the conclusions on the permutation test.

Pooling both families and all seeds gives 51.8% ($[42.3, 61.3]$) of 504 comparisons. Point estimates exceed 50% for every Llama seed and fall below 50% for every Pythia seed. Under the permutation test, however, only Llama seed 42 is significant at one-sided $p < 0.05$ ($p = 0.041$), and no Pythia seed is significantly below chance (lower-tail $p = 0.087, 0.133, 0.353$). The fitted within-pair scale

Family	Expert set	Norm spread	Ordering acc.	95% CI	Contrast R^2
Llama 3.1-8B	Full data	28.6×	78.6%	[69.0, 88.1]	0.34
	Equal exposure, seed 42	1.5×	72.6%	[58.3, 85.7]	0.00
	Equal exposure, seed 1	1.2×	67.9%	[51.2, 83.3]	0.00
	Equal exposure, seed 2	1.3×	61.9%	[45.2, 78.6]	0.02
Pythia 6.9B	Full data	75.3×	86.9%	[75.0, 95.2]	0.44
	Equal exposure, seed 42	1.3×	32.1%	[16.7, 48.8]	0.02
	Equal exposure, seed 1	1.2×	32.1%	[14.3, 50.0]	0.02
	Equal exposure, seed 2	1.2×	44.0%	[26.2, 61.9]	0.00

Table 16: Full-data versus exposure-matched experts. Norm spread is the max/min ratio of the eight task-vector norms. Ordering accuracy is the share of pair–method comparisons in which the smaller-norm expert loses more, with pair-bootstrap 95% intervals (5,000 resamples of pairs, seed 2027). Contrast R^2 is the share of contrast energy explained by a zero-intercept fit on relative scale within the same set. Norm spread is shown to one decimal. At two decimals, the matched spreads are 1.50, 1.25, and 1.26 (Llama) and 1.25, 1.23, and 1.22 (Pythia). Pair-bootstrap intervals do not account for norms being attached to just eight experts. Table 17 adds exact expert-level permutation tests.

slopes change sign across seeds in both families ($+0.006$, $+0.009$, -0.163 for Llama; $+0.201$, -0.292, -0.067 for Pythia).

At matched exposure, task and head type largely determine the remaining norm ordering (STS-B, a one-output regression head, has the smallest or near-smallest norm in every seed), so the effective sample is about eight tasks per family. Table 18 and Figure 8 give each task's share of lost comparisons. For example, Llama RTE loses 86% of its full-data comparisons but 19% at matched exposure, whereas Llama QQP rises from 29% to 76% and Pythia QNLI from 29% to 100%. The matched shares no longer track their full-data values. Spearman correlations between matched seeds are 0.73–0.88 for Llama and 0.19–0.81 for Pythia.

Family	Expert set	Contrasts	Mean D	Contrast share	TA	TIES	DARE	p_{flip}	p_{perm}
Llama 3.1-8B	Full data	84	0.270	52.2%					
	Matched, seed 42	84	0.068	44.2%	75.0%	71.4%	71.4%	0.004	0.041
	Matched, seed 1	84	0.097	53.0%	67.9%	71.4%	64.3%	0.022	0.090
	Matched, seed 2	84	0.052	30.9%	67.9%	60.7%	57.1%	0.097	0.208
Pythia 6.9B	Full data	84	0.244	75.1%					
	Matched, seed 42	84	0.082	44.0%	28.6%	35.7%	32.1%	0.985	0.925
	Matched, seed 1	84	0.084	53.4%	32.1%	32.1%	32.1%	0.983	0.915
	Matched, seed 2	84	0.101	53.7%	46.4%	39.3%	46.4%	0.751	0.677

Table 17: Per-seed details. Mean D averages both targets across merges. Contrast share measures the within-merge contribution to degradation variance (Equation 4). TA/TIES/DARE report smaller-norm ordering accuracy for each method. Both p values are one-sided for ordering above 50%: p_{flip} from the cluster sign-flip test over pairs (anti-conservative), p_{perm} from the exact permutation of the eight expert norms across tasks.

Dose-response details. Each dosed MNLI or QQP expert is merged with the seed-42 equal-exposure RTE, MRPC, and CoLA experts under TA, TIES, and DARE (nine contrasts per level). At 2,490 examples, we reuse the seed-42 equal-exposure expert. The 10,000- and 40,000-example experts extend the same shuffled subset, with optimizer steps proportional to N. For the full-data level, we rescore the pairwise-study expert with this evaluator. That expert was trained on the entire split with best-epoch selection, so its selected steps are not proportional to N. Norm and the partners' mean degradation increase with N in every series. The dosed task's own degradation stays at or below 0.069 at every subset level but rises to 0.148–0.289 with three of the four full-data experts (Pythia QQP: 0.000) (Table 19; Figure 9). With one training seed and three small partner tasks, these loser shares (nine contrasts per level) remain descriptive.

Shrink-only details. This experiment uses the full-data experts and four fixed pairs from Section 7, scored with the pairwise evaluator. Thus $c = 1$ reproduces TA under that evaluator; it does not reproduce the separate controlled evaluation in Appendix E. Mean worst-target degradation over the

Task (full N)	Llama 3.1-8B Full	s42	s1	s2	Pythia 6.9B Full	s42	s1	s2
RTE (2,490)	0.86	0.19	0.24	0.29	0.95	0.81	0.71	0.29
MRPC (3,668)	0.67	0.67	0.81	0.81	0.62	0.24	0.43	0.62
STS-B (5,749)	0.76	0.62	0.48	0.19	0.57	0.48	0.14	0.05
CoLA (8,551)	0.62	0.86	0.67	0.62	0.86	0.57	1.00	1.00
SST-2 (67,349)	0.29	0.14	0.00	0.14	0.38	0.10	0.00	0.48
QNLI (104,743)	0.24	0.67	0.57	0.76	0.29	1.00	0.86	0.86
QQP (363,846)	0.29	0.76	0.90	0.95	0.00	0.29	0.43	0.57
MNLI (392,702)	0.29	0.10	0.33	0.24	0.33	0.52	0.43	0.14
Spearman with full data		−0.04	−0.12	−0.17		0.14	0.24	0.00
Spearman with $\log N$	−0.85				−0.86			

Table 18: Share of each task's pair–method comparisons (seven partners × three methods) in which it loses more than its partner. With full data, this share decreases as training-set size grows. After exposure matching, it no longer follows the full-data pattern.

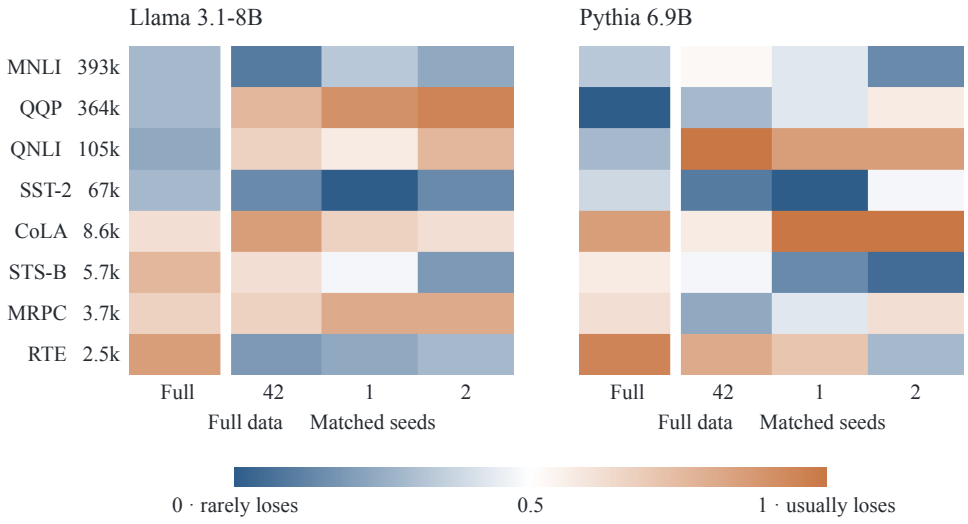

Figure 8: Share of each task's 21 pair–method comparisons (seven partners × three methods) in which it degrades more than its partner, with ties counting one half. Orange indicates frequent losses; blue indicates infrequent losses. Rows are ordered by full training-set size N_{full} (grey). With full data (first column), the loss share is at most 0.38 for each of the four largest tasks and at least 0.57 for each of the four smallest, in both families (Spearman with $\log N_{\text{full}}$: -0.85 for Llama, -0.86 for Pythia). With exposure matched at 2,490 examples (three seeds), the shares no longer track training-set size. Llama QQP rises from 0.29 to 0.76–0.95 and Llama RTE falls from 0.86 to 0.19–0.29. The Spearman correlation between the full-data shares and the seed-mean matched shares is 0.00 for Llama and 0.13 for Pythia. Matched shares are more consistent across seeds for Llama (pairwise Spearman 0.73–0.88) but less so for Pythia (0.19–0.81).

four pairs is 0.301, 0.245, 0.222, and 0.079 for Llama at $c = 1, 0.75, 0.5, 0.25$, and 0.280, 0.222, 0.249, and 0.377 for Pythia, making 0.25 the best evaluated coefficient for the worst target in Llama and 0.75 in Pythia. At $c = 0.25$, Pythia RTE, the smaller-norm target of its pair, also degrades ($+0.278$), and Pythia CoLA stays at zero MCC for every c.

Uniform-scaling control. We also merge the same four pairs after multiplying both vectors by $c \in \{0.75, 0.5\}$ (16 evaluations per family), scored with the same evaluator. Table 20 compares each target's mean D under uniform scaling and shrinkage of only the larger vector, using the same c.

Family	Task	Train examples	Norm	D (task)	D (partners)	Task loses more
Llama-3.1-8B	MNLI	2,490	11.44	0.017	0.168	11%
Llama-3.1-8B	MNLI	10,000	26.59	0.006	0.233	22%
Llama-3.1-8B	MNLI	40,000	69.31	0.014	0.325	0%
Llama-3.1-8B	MNLI	392,702	301.89	0.189	0.365	22%
Llama-3.1-8B	QQP	2,490	9.94	0.053	0.008	78%
Llama-3.1-8B	QQP	10,000	22.69	0.004	0.149	0%
Llama-3.1-8B	QQP	40,000	61.47	0.010	0.381	0%
Llama-3.1-8B	QQP	363,846	211.44	0.289	0.406	33%
Pythia-6.9B	MNLI	2,490	17.23	0.069	0.126	22%
Pythia-6.9B	MNLI	10,000	48.82	0.025	0.245	0%
Pythia-6.9B	MNLI	40,000	130.13	0.030	0.313	11%
Pythia-6.9B	MNLI	392,702	1194.73	0.148	0.346	22%
Pythia-6.9B	QQP	2,490	16.04	0.027	0.051	33%
Pythia-6.9B	QQP	10,000	45.89	0.007	0.265	0%
Pythia-6.9B	QQP	40,000	112.71	0.016	0.348	11%
Pythia-6.9B	QQP	363,846	1261.09	-0.000	0.565	0%

Table 19: Dose-response. Each row averages nine contrasts (three partners $\times$ three methods). Norm refers to the dosed task vector, and D (partners) averages degradation on RTE, MRPC, and CoLA. The final column gives the fraction of contrasts in which the dosed task loses more. The largest N in each series is the full-data expert (best-epoch selection).

Shrinking protects the smaller-norm target more in both families and at both values of c, while the larger-norm target does as well or better under uniform scaling.

Family	c	D smaller-norm		D larger-norm	
		shrink	uniform	shrink	uniform
Llama-3.1-8B	0.50	0.222	0.294	0.004	0.001
Llama-3.1-8B	0.75	0.245	0.281	-0.002	-0.003
Pythia-6.9B	0.50	0.180	0.267	0.100	0.034
Pythia-6.9B	0.75	0.213	0.259	0.020	-0.005

Table 20: Uniform-scaling control: mean D over the four pairs under scaling by c of either the larger vector alone (shrink) or both vectors (uniform). Lower is better.

Family	c	D smaller-norm	improved	D larger-norm	improved
Llama-3.1-8B	0.25	0.073	4/4	0.021	0/4
Llama-3.1-8B	0.50	0.222	3/4	0.004	2/4
Llama-3.1-8B	0.75	0.245	4/4	-0.002	3/4
Llama-3.1-8B	1.00	0.301	0/4	0.002	0/4
Pythia-6.9B	0.25	0.235	2/4	0.259	0/4
Pythia-6.9B	0.50	0.180	2/4	0.100	1/4
Pythia-6.9B	0.75	0.213	3/4	0.020	3/4
Pythia-6.9B	1.00	0.280	0/4	0.023	0/4

Table 21: Shrink-only allocation: the smaller vector keeps coefficient 1 and the larger receives c ($c = 1$ is TA). D columns average over the four pairs; "improved" counts targets whose degradation falls relative to TA in the same evaluator.

E Controlled Norm-Allocation Experiment

Design and construction. We use the trained BF16 expert and base checkpoints. We form task-vector differences in FP32, construct each candidate backbone from the base in FP32, and then cast to BF16. Classification/regression heads are excluded from task-vector geometry and restored from the target expert. Evaluation uses right padding, maximum sequence length 128, and batch size 16. Full validation sizes are MNLI 9,815; RTE 277; QNLI 5,463; STS-B 1,500; MRPC 408; SST-2 872;

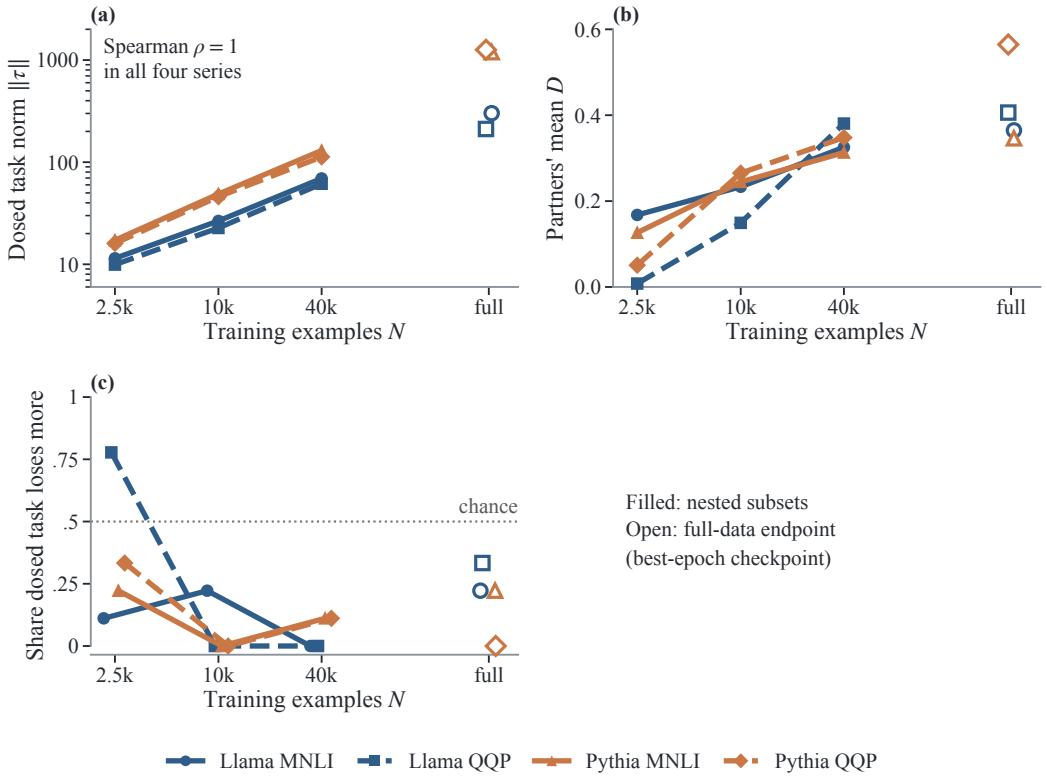

Figure 9: Dose-response for MNLI and QQP. (a) The dosed expert's task-vector norm rises monotonically with its number of training examples N in all four series (Spearman $\rho = 1$). (b) The mean degradation D of its partners (RTE, MRPC, CoLA) also rises monotonically with N, and hence with its norm, in all four series. (c) The fraction of contrasts lost by the dosed task does not change consistently with N. It is at most 3 of 9 at every level except Llama QQP at 2,490 examples (7 of 9). The dotted line marks 0.5. Each point averages nine contrasts (three partners $\times$ TA, TIES, DARE). Experts at 2,490–40,000 examples are trained on nested subsets with one seed. The largest N ("full": 392,702 for MNLI, 363,846 for QQP) uses best-epoch selection and appears as an open marker, separated from the nested-subset series. Horizontal offsets in (c) separate tied points. Exact values are in Table 19.

CoLA 1,043; and QQP 40,430. The four fixed pairs include every task once per family. Each family contributes one trained expert collection; this experiment does not replicate training seeds.

Expert scores in Table 23 are matched re-evaluations of the same checkpoints with this evaluator (a plain BF16 forward pass without autocast, batch 16), whereas Table 5 reports the Trainer evaluation used for QC (BF16 weights under mixed-precision autocast; batch 32 for Llama). Validation rows, padding side, and maximum length are identical. The floating-point execution paths nevertheless change a few borderline predictions (e.g., 2 of 277 for Pythia RTE, 0.809 versus 0.801; +0.015 MCC for Pythia CoLA). We therefore measure every intervention effect against its matched baseline in Table 23.

With $n_i = \|\boldsymbol{\tau}_i\|$, $n_j = \|\boldsymbol{\tau}_j\|$, and $\eta = \langle \hat{\boldsymbol{\tau}}_i, \hat{\boldsymbol{\tau}}_j \rangle$, the coefficients in Equation 5 are

$$c_i(\rho) = \frac{\Lambda}{n_i \sqrt{1 + \rho^2 + 2\rho\eta}}, \qquad c_j(\rho) = \frac{\Lambda\rho}{n_j \sqrt{1 + \rho^2 + 2\rho\eta}}. \tag{7}$$

At $\rho_0 = n_j/n_i$, both coefficients are one. At $\rho = 1$, both components have norm $\Lambda/\sqrt{2 + 2\eta}$. This fixed-total-norm constraint differs from the mean-one coefficient normalization in Table 31, which equalizes at the harmonic mean of the norms (Equation 13). The experiments therefore test different

calibration rules. Under fixed-total-norm equalization, the large coefficients assigned to smaller vectors can move them far from their trained endpoints (Table 22).

Family	Pair (i, j)	$\rho_0 = n_j/n_i$	$c_i(1)$	$c_j(1)$
Llama 8B	MNLI + RTE	0.03493	0.706	20.228
Llama 8B	QNLI + STS-B	0.0849	0.708	8.339
Llama 8B	MRPC + SST-2	6.626	4.736	0.715
Llama 8B	CoLA + QQP	11.64	8.256	0.709
Pythia 6.9B	MNLI + RTE	0.01403	0.705	50.294
Pythia 6.9B	QNLI + STS-B	0.1075	0.709	6.596
Pythia 6.9B	MRPC + SST-2	7.055	5.035	0.714
Pythia 6.9B	CoLA + QQP	34.25	23.850	0.696

Table 22: Natural component ratios and equalization coefficients. Every intervention preserves the unit-coefficient TA update norm before quantization. The smaller vector's coefficient ranges from 4.736 to 50.294.

Coverage and numerical checks. The evaluation covers 96 unique merges, 16 exact individual experts, and 16 unique base conditions. Repeating base and individual conditions across execution shards gives 160 evaluation records for 128 unique cells. Repeated baselines have identical primary metrics and per-example losses. The largest relative BF16 update-norm deviation is 0.142% for Llama and 0.032% for Pythia. Stored predictions, labels, example IDs, and per-example losses permit independent recomputation of every primary metric and loss. Hash manifests verify the downloaded artifacts. We use cross-entropy for classification and MSE for STS-B, keeping these loss units separate. The evaluator records zero Pearson correlation for constant predictions.

Family	Target	Expert	TA	Equal	ΔM
Llama 8B	MNLI	0.902	0.903	0.354	-0.549
Llama 8B	RTE	0.892	0.794	0.527	-0.267
Llama 8B	QNLI	0.954	0.950	0.495	-0.456
Llama 8B	STS-B	0.920	0.638	0.105	-0.533
Llama 8B	MRPC	0.927	0.746	0.797	+0.052
Llama 8B	SST-2	0.966	0.962	0.740	-0.222
Llama 8B	CoLA	0.669	0.022	-0.012	-0.035
Llama 8B	QQP	0.869	0.865	0.142	-0.723
Pythia 6.9B	MNLI	0.781	0.778	0.318	-0.460
Pythia 6.9B	RTE	0.801	0.751	0.473	-0.278
Pythia 6.9B	QNLI	0.922	0.844	0.495	-0.349
Pythia 6.9B	STS-B	0.899	0.637	0.109	-0.528
Pythia 6.9B	MRPC	0.883	0.685	0.802	+0.117
Pythia 6.9B	SST-2	0.947	0.943	0.517	-0.425
Pythia 6.9B	CoLA	0.614	0.000	0.000	+0.000
Pythia 6.9B	QQP	0.669	0.662	0.548	-0.113

Table 23: Prespecified equalization comparisons: $\Delta M = M(\boldsymbol{\theta}(1)) - M(\boldsymbol{\theta}(\rho_0))$ (positive means better retention). Both improvements occur on MRPC, whose F1 scores remain below the all-positive reference (0.812). Pythia CoLA is tied at zero. Expert scores use matched controlled re-evaluations; Table 5 uses the Trainer evaluator.

Figure 11 compares three ways to change the norm ratio: stretching the smaller vector (this appendix), scaling both vectors, and shrinking only the larger one (Appendix D).

Single-component controls. For each target loss L_t, we also evaluate the base and four single-component backbones: $\boldsymbol{\theta}_0 + \boldsymbol{\tau}_i$, $\boldsymbol{\theta}_0 + \boldsymbol{\tau}_j$, $\boldsymbol{\theta}_0 + c_i(1)\boldsymbol{\tau}_i$, and $\boldsymbol{\theta}_0 + c_j(1)\boldsymbol{\tau}_j$. Each uses the target t's trained head, including when the component comes from the other task. These 64 single-component evaluations and 16 base checks use the same examples, numerical construction, tokenizer, batch size, and precision as the primary sweep. This exploratory diagnostic includes every fixed pair and both targets.

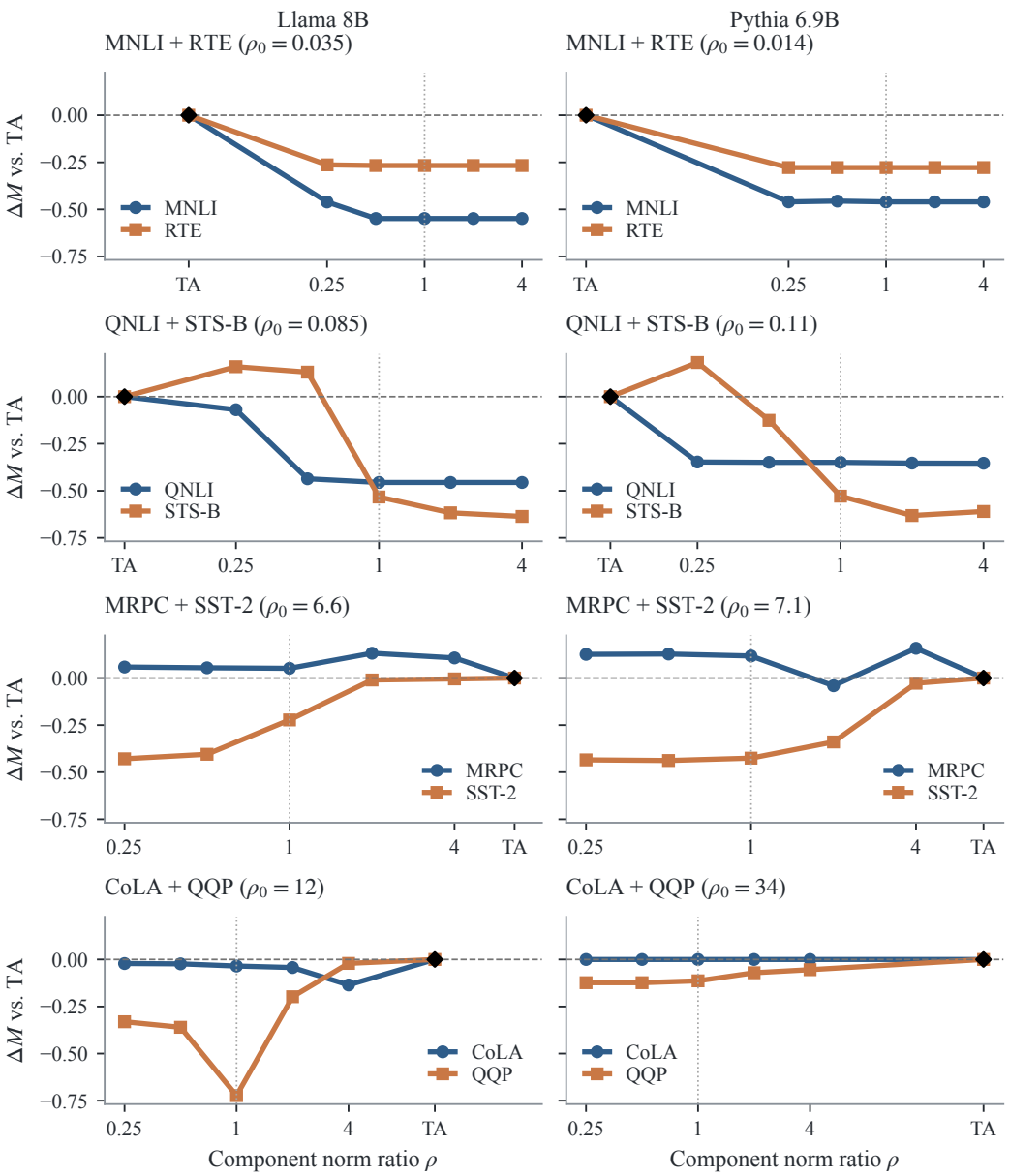

Figure 10: All 96 controlled merge evaluations. Each line shows one target's change in its own primary metric relative to natural TA. Black diamonds mark natural TA at its actual logarithmic coordinate (ρ_0 in panel titles). The dotted vertical line marks equal component norms. Connected segments guide the eye between evaluated ratios. Targets within a panel may use different metrics, so each curve should be read as that target's change. No ratio is selected to define a reported method.

Define the base-anchored finite-difference interaction

$$I_t(c_i, c_j) = L_t(\boldsymbol{\theta}_0 + c_i\boldsymbol{\tau}_i + c_j\boldsymbol{\tau}_j) - L_t(\boldsymbol{\theta}_0 + c_i\boldsymbol{\tau}_i) - L_t(\boldsymbol{\theta}_0 + c_j\boldsymbol{\tau}_j) + L_t(\boldsymbol{\theta}_0). \tag{8}$$

Subtracting natural TA from equal allocation gives the exact identity

$$\Delta L_t = \Delta L_{t,\text{iso}} + \Delta I_t, \tag{9}$$

where $\Delta L_{t,\text{iso}}$ is the sum of the two single-component loss changes and $\Delta I_t = I_t(c_i(1), c_j(1)) - I_t(1, 1)$. This decomposition is conditional on the chosen base and component parameterization. It

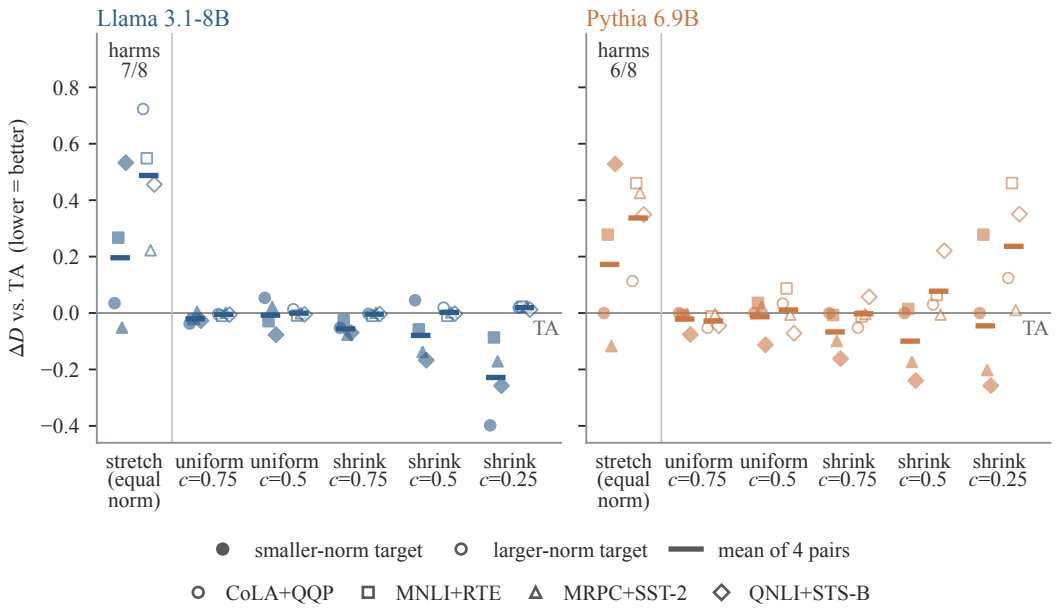

Figure 11: Allocation trade-offs on four fixed pairs (Llama left, Pythia right). Each target's change in degradation is relative to TA in the same evaluator, $\Delta D = D_{\text{intervention}} - D_{\text{TA}}$. Values below zero indicate better retention. Shape identifies the pair; filled/open markers denote smaller/larger-norm targets. Horizontal bars average four pairs. Stretch harms 13/16 targets (7/8 Llama, 6/8 Pythia). On average, shrinking protects the smaller targets more than uniform scaling. Stronger shrinkage can harm the larger targets, especially for Pythia. Pythia CoLA stays at zero MCC. Stretch uses the controlled evaluator ($\Delta D = -\Delta M$; Table 23); uniform and shrink use the pairwise evaluator (Tables 20 and 21).

Family	Target	ΔL	ΔL_{iso}	ΔI	95% interval
Llama 8B	MNLI	+1.998	+3.280	-1.282	[-1.379, -1.190]
Llama 8B	RTE	+0.843	+5.562	-4.719	[-5.349, -4.097]
Llama 8B	QNLI	+1.331	+0.800	+0.531	[+0.507, +0.555]
Llama 8B	STS-B (MSE)	-7.287	+2.981	-10.268	[-10.582, -9.936]
Llama 8B	MRPC	+0.165	+1.481	-1.316	[-1.543, -1.084]
Llama 8B	SST-2	+0.417	+1.317	-0.900	[-1.090, -0.709]
Llama 8B	CoLA	+0.005	+1.004	-0.999	[-1.074, -0.923]
Llama 8B	QQP	+0.466	-1.718	+2.184	[+2.166, +2.201]
Pythia 6.9B	MNLI	+0.875	+41.706	-40.831	[-41.517, -40.202]
Pythia 6.9B	RTE	+0.623	+52.216	-51.593	[-57.588, -46.102]
Pythia 6.9B	QNLI	+0.718	+0.455	+0.262	[+0.241, +0.283]
Pythia 6.9B	STS-B (MSE)	-3.508	+1.963	-5.471	[-5.808, -5.145]
Pythia 6.9B	MRPC	+0.788	+3.053	-2.265	[-2.594, -1.955]
Pythia 6.9B	SST-2	+1.853	+2.072	-0.219	[-0.559, +0.104]
Pythia 6.9B	CoLA	-1.391	+10.818	-12.210	[-13.360, -11.126]
Pythia 6.9B	QQP	+0.555	+18.759	-18.204	[-18.344, -18.061]

Table 24: Loss decomposition for equal allocation minus natural TA. Positive ΔL denotes higher joint loss. The isolated term sums both single-component changes on the same target loss. The residual gives the interaction change; intervals use a paired-example bootstrap. All rows use CE except the explicitly marked STS-B MSE rows; these losses are not pooled.

describes the evaluated changes without estimating a Hessian constant or establishing a universal mechanism. In particular, joint loss can increase even when the interaction term decreases, if isolated rescaling incurs a larger increase.

Results. The single-component evaluations give the exact base-anchored decomposition $\Delta L_t = \Delta L_{t,\mathrm{iso}} + \Delta I_t$ of each target loss change (Table 24). Loss rises for 13 targets, which are not exactly the same 13 targets whose metrics fall (both MRPC targets gain F1 while CE rises; both STS-B targets lose Pearson correlation while MSE falls). Isolated rescaling raises loss for 15 of 16 targets (all but Llama QQP) and, in 12 of the 13 targets whose loss rises, by at least half the joint increase. The interaction change is negative for 13 of 16 (12 with paired-example intervals excluding zero; Pythia SST-2's $[-0.559, +0.104]$ crosses zero) and falls for 10 of the 13 targets whose joint loss rises. Joint loss can rise with either sign of interaction change (Llama MNLI: $+1.998 = +3.280 - 1.282$; Llama QQP: $+0.466 = -1.718 + 2.184$). A negative interaction change is an accounting result and does not establish reduced interference: in the pairs stretched 24–50-fold it mostly offsets a catastrophic single-component loss (Pythia MNLI: $+41.706$ isolated, -40.831 interaction).

We align conditions by stored example IDs and labels and verify Equation 9 for every example before averaging. Table 24 reports all targets with 1,000 paired-example bootstrap resamples, seed 2027, and percentile 95% intervals for ΔI_t. The intervals reflect resampling of evaluation examples only. They do not capture training-seed uncertainty or make the tasks independent replications. CE and STS-B MSE retain their own units.

F DARE Mask Sensitivity

We assess mask sensitivity in a separate study of five Llama pairs at DARE pruning rates $0.5/0.9/0.99$, totaling 200 evaluations (Table 25). At $p = 0.9$, median within-cell degradation SD is 0.013, while worst-task degradation spans 0.569–0.792 across masks. At $p = 0.99$, the maximum cell SD reaches 0.243. The mask experiment uses five selected pairs, two target directions, and 20 total masks per pair spread across three rates. At a given rate, within-cell SD is computed across mask seeds for one pair–target cell. We also report the maximum SD across the ten cells to show variation that the median can hide. The worst-task range takes the maximum degradation over those cells for each seed and reports its range over seeds.

Family	p	Mean D	Median SD	Max SD	Worst D range
Llama 8B	0.5	0.117	0.003	0.019	0.468–0.505
Llama 8B	0.9	0.290	0.013	0.119	0.569–0.792
Llama 8B	0.99	0.450	0.037	0.243	0.575–0.871

Table 25: Mask variability in 200 directional Llama evaluations.

These five selected pairs cannot establish mask stability for every pairwise or full-merge configuration. At high pruning rates, substantial tail variation remains even when the typical cell is stable.

G External Generative Experts

We apply the two untrained rules of Section 5.3 to public MergeBench experts (He et al., 2025): coding, instruction following, math, and multilingual reasoning on Llama 3.1-8B-Instruct and Gemma 2-2B-it, merged with TA, TIES, and DARE. Both rules are applied without fitting coefficients.

Fixed expert collections and construction. We use the public MergeBench coding, instruction, math, and multilingual checkpoints for `meta-llama/Llama-3.1-8B-Instruct` and `google/gemma-2-2b-it`. The analysis bundle includes an asset manifest fixing all ten base/expert revisions and weight-file hashes; the inspected MergeBench code revision is `dedf1b1b76d69018c058e2ee45406ba5d0021238`. We retain every expert and pair regardless of its evaluation outcome. Each family has six pairs and two targets per pair. The three primary methods yield 72 directional scores. Arithmetic weight averaging adds 24 scores as a sensitivity check. Two bases evaluated on four domains and eight own-domain experts add 16 baseline scores, for 112 task scores from 58 model jobs. Shared experts and tasks are not independent replications.

Task vectors are FP32 differences of the released BF16 weights. We accumulate merges in FP32, store BF16, verify tokenizer and tensor compatibility, and count tied parameters once in geometry. TA uses unit coefficients. TIES retains the largest 20% of entries per tensor, elects the positive sign on ties, and averages sign-consistent contributions. This is the pairwise-study implementation; MergeBench instead uses a global-threshold/sum variant. DARE-TA uses independent pruning at $p = 0.9$ with seed 2027. Weight averaging is TA with both coefficients 0.5. We do not search over coefficients.

Full-split scoring. GSM8K uses all 1,319 test examples, the eight fixed CoT demonstrations from `lm-eval` 0.4.8, and exact match after flexible numeric extraction. IFEval uses all 541 prompts and strict prompt-level accuracy. Coding uses all 164 native HumanEval+ tasks and deterministic single-sample pass@1, requiring both base and extended tests to pass. Multilingual reasoning uses translated ARC-Challenge with French/German/Russian sizes 1,169 each and Spanish size 1,170; we report the macro-average of length-normalized multiple-choice accuracies. Generation is greedy, with caps of 1,024 tokens for math/code and 1,280 for IFEval. Each family uses its unchanged base tokenizer and chat template for generation; ARC uses the raw multiple-choice prompt. Our matched checkpoint evaluation uses different decoding settings and a narrower task suite than MergeBench. These scores therefore do not reproduce its leaderboard.

Table 26 lists all base and expert scores. The runtime uses PyTorch 2.6.0, Transformers 4.51.3, vLLM 0.8.5.post1, `lm-eval` 0.4.8, and EvalPlus 0.3.1. Code runs in a network-free filesystem sandbox with CPU, memory, process and wall-time limits. The native EvalPlus polynomial-root oracle is taken from upstream EvalPlus at revision `26d6d00bb1fd0fa37f39c99d5290da67891d1c5e`. Qualification confirms that 163 canonical solutions pass. The native Newton reference for HumanEval/32 violates the official residual tolerance on 7/788 extended cases. We retain that task, its tolerance, and all tests. Independently correct and incorrect root controls validate the oracle.

Family / domain	Base	Expert	Gain
Llama coding	0.6037	0.6159	+0.0122
Llama instruction	0.7431	0.7357	-0.0074
Llama math	0.8309	0.8317	+0.0008
Llama multilingual	0.4558	0.4631	+0.0073
Gemma coding	0.3902	0.4085	+0.0183
Gemma instruction	0.5231	0.6359	+0.1128
Gemma math	0.5171	0.6209	+0.1039
Gemma multilingual	0.4082	0.4094	+0.0013

Table 26: All external base and own-domain expert scores under the fixed evaluation protocol. Gains are measured on the same task. All experts remain in the evaluation, including the Llama instruction expert whose gain is negative.

Untrained rules. The data feature uses reported SFT counts: coding 110,000; instruction 29,900; math 591,000; multilingual 5,940. These rounded published counts do not reconstruct optimizer steps; the Llama math expert also received GRPO training. The ordering rule predicts that the task with the smaller task-vector norm (or smaller training set) loses more. For screening, we keep the three of six candidate pairs with the smallest absolute log norm ratio (or training-set-size ratio). We compare their mean worst-target degradation $w_p = \max(D_{pi}, D_{pj})$ against expected random selection (the mean over all six pairs) and against all $\binom{6}{3} = 20$ possible 3-pair sets.

Results. On Llama, the ordering rule is correct in 55.6% of pair–method comparisons (10/18: 2/6, 5/6, and 3/6 for TA, TIES, and DARE) for both norm and data. On Gemma, accuracy is 50.0% for norm and 33.3% for data. An exact test over the $2^6 = 64$ orientations of the six pairs finds that Llama's 55.6% is matched or exceeded by 45.3% of orientations ($p = 0.453$), Gemma norm's 50% by $p = 0.656$, and Gemma data's 33.3% by $p = 0.891$. Per pair, Llama ordering is 3/3 for coding+instruction and math+multilingual but 0/3 for instruction+math. Excluding every Llama-math pair leaves nine comparisons, and each rule correctly orders five.

On Llama, norm and data screening select the same pairs (coding+instruction, coding+math, instruction+multilingual) and lower mean worst-target degradation relative to random selection by 3.02,

12.81, and 5.43 percentage points for TA, TIES, and DARE (7.09 on their mean, from 0.2585 to 0.1876; the oracle is 0.0741). Among candidate sets, however, this selection ranks 9th, 6th, and 9th of the 20 possible sets (8th under weight averaging; 9th on the mean of the three methods, where 45% of sets do at least as well; Table 7), leaving the Llama gain statistically indistinguishable from chance. On Gemma, screening is worse than random: -0.88 points for norm and -0.46 for data. Weight averaging does not reverse either family-level result.

Llama experts improve over their base by at most 0.012, leaving little signal for this test. Candidate-pair screening also compares different task combinations. It does not choose coefficients for a fixed set of required capabilities.

Scoring rules. We apply two evaluator compatibility rules uniformly to every model, without consulting reference answers. First, the standard flexible GSM8K regex can select a trailing punctuation-only string such as `$.` after a boxed answer. We retain the same regex and normalization, discard candidates with no digit, and select the last remaining candidate. Second, the Gemma instruction checkpoint's generation config omits the base chat turn-ending token, so we truncate generated text at the first canonical base end token before scoring (534 of its 541 outputs; the remaining seven hit the token cap) and score IFEval with the official strict evaluator and language-detection seed 2027. Generating the Gemma coding baseline with explicit base end-token IDs instead gives identical sanitized programs and pass/fail results.

We independently reconstruct full-split means and all 96 directional outcomes from the numerical package. Across each family's 14 evaluations per generative domain, token-cap counts are respectively 149/2,200 for Llama/Gemma math, 338/1,000 for instruction, and 44/24 for code. The results are conditional on these fixed generation budgets and do not establish quality under unrestricted generation.

H Generative LoRA Experts

Generative setup. For the generative LoRA test, we train Llama 3.1-8B experts with rank 16, alpha 32, learning rate 2×10^{-4}, effective batch size 32, maximum sequence length 1,024, and up to 8,000 training examples per task. GSM8K trains for three epochs; code (CodeAlpaca training data) and SAMSum train for two. Greedy generation evaluates 200 GSM8K examples by exact match, all 164 HumanEval problems by pass@1, and 200 SAMSum examples by ROUGE-L F1. The summarization evaluator uses its pure-Python LCS implementation rather than the optional stemmed `rouge_score` backend. The resulting scores are specific to this ROUGE implementation.

Individual primary scores are 0.665 (GSM8K), 0.067 (HumanEval), and 0.460 (SAMSum). Task-vector norms are 25.737, 17.130, and 19.684. We compute the target-relative log ratio from these norms in each evaluation direction. Each fixed target and method has exactly two possible interferers. Table 27 subtracts degradation with the smaller interferer from degradation with the larger interferer.

Target	TA ΔD	TIES ΔD	DARE ΔD
GSM8K	0.020	0.020	0.030
Code	0.061	0.012	0.073
SAMSum	0.006	0.003	0.022

Table 27: Generative two-point comparisons at fixed target and method. All nine differences are positive. Because experts, task examples, and methods are shared, these are not nine independent Bernoulli trials.

This experiment extends the analysis to generative tasks, but uses one family, one training seed, and a weak code expert. They do not establish general LLM capability retention, broad generative scaling, or a universal relative importance of scale and coherence.

I Fixed-Norm Direction Stress Test

For a target i and interferer j, we blend the normalized unperturbed vector with a common normalized random direction $\boldsymbol{z}$ orthogonalized against $\boldsymbol{\tau}_i$, then restore the original norm. The blend grid is

$\gamma \in \{0, .2, .35, .5, .65, .8\}$, with random seeds 42 and 123. Four interferers give $K = 5$, and the target itself is not edited. The two random directions and three Llama targets (MNLI, SST-2, RTE) give 36 task evaluations. The total interferer norm energy A_i is fixed, while mean interferer coherence rises from approximately 0.003 to 0.94. At the strongest perturbation, degradation increases by 0.087–0.111 for MNLI and 0.092–0.119 for SST-2, while RTE improves by 0.072–0.087. Changing directions can therefore substantially alter outcomes even at fixed norm energy. The intervention also lowers target–interferer coherence and changes alignment with the loss landscape. Its measured effect combines these directional changes. Recorded A_i is constant across doses within each target cell. Orthogonality of $\boldsymbol{z}$ to the target does not keep the edited target–interferer cosine constant: normalization and blending reduce it.

Family	Target	ΔD_{42}	ΔD_{123}	Target μ start	Target μ end
Llama 8B	MNLI	0.111	0.087	0.003	0.001
Llama 8B	RTE	-0.072	-0.087	0.003	0.001
Llama 8B	SST-2	0.092	0.119	0.002	0.001

Table 28: Strongest-dose minus zero-dose degradation for all three Llama cells and both random directions. The last two columns show mean target–interferer coherence at the endpoints, showing that this quantity also changes under the intervention.

Table 28 gives the results. We retain the opposite response on RTE despite its baseline being below majority accuracy. Loss and accuracy can move differently. These results establish neither an isolated scalar coherence intervention nor an exact threshold at $1/(K-2)$ or a general monotonic relation.

J Curvature Probes Are Diagnostics

For each target expert, the experiment evaluates

$$q_i(t, v) = \frac{L_i(\boldsymbol{\theta}_i + tv) + L_i(\boldsymbol{\theta}_i - tv) - 2L_i(\boldsymbol{\theta}_i)}{t^2} \tag{10}$$

on validation examples, with unit directions v from each other task vector and their summed interference. Step sizes are $t \in \{.25, .5, 1\}\|r_i\|$. Between probes, we restore model parameters from a snapshot. Perturbations are formed in float32 and then cast to BF16. Restoration limits accumulation drift, although numerical and finite-step effects remain. The minimum and maximum over the sampled directions and steps are denoted $q_{\min}$ and $q_{\max}$.

Family	Positive minima	Median max/min	Nonpositive tasks
Llama 8B	8/8	37.6	–

Table 29: Sampled curvature summaries for the eight Llama experts. The median max/min ratio includes only tasks with positive sampled minima.

Table 29 summarizes the probes. Even a positive sampled minimum does not bound curvature from below at every point and direction. Under genuine uniform bounds on the probed positive and negative segments, the true lower constant would be no larger than the observed minimum, and the upper constant no smaller than the observed maximum. The finite observations do not establish the reverse inequalities. Symmetric differences cancel the first-order term, whereas actual one-sided merging retains it. Llama MNLI, for example, has positive sampled symmetric differences but seven probes with lower positive-direction loss than the expert. The probes therefore cannot be used to treat the required assumptions as verified. Appendix A.3 uses the sampled ranges only as a diagnostic check of Corollary 1's condition.

K DARE Moments and Calibration Details

Independent Bernoulli DARE masks preserve expected inner products but inflate expected squared norms by $1/(1-p)$. For fixed vectors x, y and independent Bernoulli(κ) masks ξ_l, ζ_l over coordinates

and vectors, where $\kappa = 1 - p > 0$, define $x'_l = \xi_l x_l/\kappa$ and $y'_l = \zeta_l y_l/\kappa$. Then

$$\mathbb{E}\|x'\|^2 = \|x\|^2/\kappa, \qquad \mathbb{E}\langle x', y'\rangle = \langle x, y\rangle, \tag{11}$$

$$\mathrm{Var}(\|x'\|^2) = \frac{1-\kappa}{\kappa^3}\sum_l x_l^4, \qquad \mathrm{Var}(\langle x', y'\rangle) = \frac{1-\kappa^2}{\kappa^2}\sum_l x_l^2 y_l^2. \tag{12}$$

These identities follow from independence and Bernoulli moments. Table 30 reports measured post-mask geometry. They describe signed inner products; expected absolute coherence requires further conditions. Approximate coherence attenuation by κ requires nonzero retained norms and adequate concentration of numerator and denominator. When an unmasked inner product is zero or nearly cancels, errors must be described additively: random masks can create nonzero cosine even when the unmasked cosine is zero. If a few coordinates carry most of the energy, the required concentration may also fail. Direct weight-space measurements at three Pythia sizes (70M, 1.4B, and 6.9B) show nearly unchanged scale ordering and approximately this coherence reduction, showing that DARE can substantially change directions without equalizing task-vector norms.

Pythia	p	Norm inflation	Mean μ'/μ	Scale corr.
70M	0.5	1.414	0.501	1.000000
70M	0.7	1.826	0.301	1.000000
70M	0.9	3.162	0.100	1.000000
70M	0.95	4.472	0.050	0.999999
70M	0.99	9.997	0.013	0.999998
1.4B	0.5	1.414	0.500	1.000000
1.4B	0.7	1.826	0.300	1.000000
1.4B	0.9	3.162	0.100	1.000000
1.4B	0.95	4.472	0.050	1.000000
1.4B	0.99	10.001	0.010	1.000000
6.9B	0.5	1.414	0.500	1.000000
6.9B	0.7	1.826	0.300	1.000000
6.9B	0.9	3.163	0.100	1.000000
6.9B	0.95	4.473	0.050	1.000000
6.9B	0.99	10.004	0.010	0.999998

Table 30: Post-mask geometry measurements at three Pythia sizes. Only weight geometry is measured; this experiment includes no task evaluation. Correlations rounded to six decimals do not assert exact invariance. The observed coherence ratios are close to κ, but do not establish a general multiplicative law.

Inverse-norm calibration. The coefficients $\lambda_i = \|\boldsymbol{\tau}_i\|^{-1}/(K^{-1}\sum_{j\in S}\|\boldsymbol{\tau}_j\|^{-1})$ equalize component magnitudes while keeping the mean coefficient at one. Each rescaled component has the harmonic mean of the K norms. The largest vectors shrink most; in a pair, the smaller vector can at most double. With Λ, η, and $c_i(1)$ as in Section 7 and H the harmonic mean of the norms,

$$\underbrace{\|c_i(1)\boldsymbol{\tau}_i\| = \|c_j(1)\boldsymbol{\tau}_j\| = \tfrac{\Lambda}{\sqrt{2+2\eta}}}_{\text{fixed total norm}}, \qquad \underbrace{\|\lambda_i\boldsymbol{\tau}_i\| = H,\ \ \lambda_{\text{small}} = \tfrac{2n_{\text{large}}}{n_{\text{small}}+n_{\text{large}}} < 2}_{\text{mean-one inverse norm (pair)}}. \tag{13}$$

This differs from the fixed-total-norm equalization of Appendix E, which stretches the smaller vector 4.7–50.3-fold. Among five matched methods, inverse-norm coefficients give the highest average metric on Qwen and exceed uniform TA in all three families. Llama inverse-norm averages 0.447, below TIES (0.481) and SVC (0.609); Mistral inverse-norm likewise trails TIES (0.377 versus 0.381). An association between scale and damage therefore does not make equalization an optimal merging rule.

Table 32 summarizes the calibration, and Table 33 gives the pairwise gains. Matched pairwise calibration gains are 0.051 for Llama, 0.109 for Mistral, and 0.101 for Qwen. These positive means conceal losses for individual tasks and pairs. At the task level, inverse norm changes the larger-norm task's metric by a mean of -0.191 (28/28 pairs negative), -0.023 (8/10), and -0.130 (13/15) for Llama, Mistral, and Qwen, with worst drops from -0.05 (Mistral) to -0.60 (Llama). The smaller-norm task has the larger degradation in 47 of 53 pairs under uniform TA but in 20 of 53 under inverse norm.

Family	TA	TIES	DARE	InvNorm	SVC
Llama 8B	0.306	0.481	0.370	0.447	0.609
Mistral 7B	0.337	0.381	0.207	0.377	0.475
Qwen 7B	0.260	0.254	0.262	0.532	0.293

Table 31: Ancillary full-merge average primary metric. Llama merges all eight tasks; Mistral and Qwen use their QC-valid task sets. Comparisons are matched within the experiment and use one training seed. The averages combine different task metrics.

Family	TA	InvNorm	Best other	Pairwise InvNorm−TA
Llama 8B	0.306	0.447	SVC 0.609	+0.051
Mistral 7B	0.337	0.377	SVC 0.475	+0.109
Qwen 7B	0.260	0.532	SVC 0.293	+0.101

Table 32: Mean-one inverse-norm calibration (InvNorm) versus uniform TA: full-merge average primary metric, the best of TIES/DARE/SVC, and the matched pairwise mean gain (summary of Tables 31 and 33). Averages combine different task metrics from one training seed.

As a post-hoc sensitivity check, we clip inverse-norm coefficients to a floor and renormalize them. At floors 0.1–0.3, Llama's mean of 0.452–0.475 remains below SVC. The floors are not selected by independent validation, and the zero minimum metrics do not certify retention.

L Relation to Existing Diagnostics and Controls

Prediction targets. Our directional outcome D_i and contrast $a = (D_i - D_j)/2$ distinguish the two members of a merge. Demystifying Mergeability instead predicts the pair's average normalized retention and includes a symmetric min/max norm ratio among its features (Zhou et al., 2026b). The symmetry of a predictor must match the outcome it predicts. Equations 3–4 are elementary mean/contrast identities. We use them to quantify the observed 64.1% contrast share and explain the pooled predictor comparison. Target-conditioned geometry can also distinguish tasks: SVC uses task-specific projections onto merged singular directions, while the collapse study uses target-specific representation distances (Li et al., 2026; Cao et al., 2026). Our findings for absolute coherence do not evaluate these target-conditioned predictors.

Closest predictive comparisons. Will it Merge? relates base-model knowledge to retention averaged over partners and tests merging rules beyond its main KnOTS experiments (Rahamim et al., 2026). Our individual-expert score gap differs from its base-model prior. MergeProbe proposes directional prediction from geometry and training metadata, with simulator-based and small-pilot evidence and synthetic tabulated results (Tang et al., 2026). SimMerge learns task-conditioned operator and merge-plan selection from similarity signals (Bolton et al., 2026). Our screening selects candidate pairs under a fixed retention budget; it does not select operators or optimize coefficients for a required task set. Comparing these richer predictors on common checkpoints remains open.

Exposure and training interventions. Table 34 compares magnitude and direction interventions with our allocation test. Self Positioning links data imbalance to task-vector norms and optimizes magnitude and direction for text embeddings (Li et al., 2024). Overtraining studies vary training duration (Horoi et al., 2026), while the safety-collapse study retrains an expert on balanced data (Choudhary et al., 2026). We ask whether norm ratio adds predictive information to data/step

Family	Pairs	Mean gain	Positive pairs	Min gain	Max gain
Llama 8B	28	0.051	16	-0.132	0.729
Mistral 7B	10	0.109	9	-0.019	0.466
Qwen 7B	15	0.101	11	-0.182	0.488

Table 33: Pairwise inverse-norm minus uniform-addition mean metric from matched evaluations in the calibration experiment.

Approach	Quantity or intervention	Relation to our allocation test
DisTaC (Yoshida et al., 2026)	Shrinks an oversized task vector and distills to preserve confident predictions; includes scaling-only controls.	Our total update norm is fixed and both isolated component effects are measured.
DO-Merging (Zheng et al., 2025)	Decouples column magnitudes and directions of LoRA updates; analyzes a matrix discrepancy surrogate.	Whole-vector allocation and measured task loss are different objects.
NAN / FroM (Si et al., 2025; Li et al., 2025)	NAN derives inverse model-weight-norm coefficients; FroM uses powers of layer task-vector norms.	Neither coefficient rule is the fixed-total-norm path in Equation 5.
Task Alignment (de Jorge et al., 2026)	Selects merges using representation discrepancy; also tests norm normalization.	Our matched controls identify isolated and joint loss changes for a specified parameter intervention.
TSV / SVC (Gargiulo et al., 2025; Li et al., 2026)	Alters layerwise singular directions or singular values.	Global norm equalization does not implement these spectral operations.
Reading the Gate (Zhu, 2026)	Factorial activation cross-terms and norm-matched activation interventions.	Our four-corner difference is in task loss under parameter-coefficient changes.
DMS (Bossi, 2026)	Searches module policies and redistributes module scales while preserving their mean.	A mean coefficient budget differs from a fixed merged-update norm.

Table 34: Norm, direction, and interaction refer to different mathematical objects across methods. The 13/16 harm result applies to the allocation rule tested here and does not generalize to all magnitude-aware merging. The DMS description follows the thesis executive summary.

ratios under identical held-group splits. This observational comparison does not imply that extending every expert's training improves merging. The pairwise shrink experiment tests neither FedNova-style step-normalized coefficients (Tao et al., 2025) nor DisTaC's distillation-based conditioning.

Interpretation of controls and theory. The finite-difference interaction in Equation 8 is a standard four-corner construction. The matched evaluations show why its terms must be interpreted carefully: opposite interaction changes can accompany the same direction of joint loss change. Our two-sided bounds likewise apply Taylor's formula to the actual expert-to-merge displacement, retaining the residual gradient. This complements gradient matching, tangent-space analyses, adaptive projective merging, and training-dependent loss bounds (Daheim et al., 2024; Ortiz-Jimenez et al., 2023; Wei et al., 2025; 2026). The upper bound constrains capacity. The lower bound quantifies damage under an explicit lower-curvature assumption; it does not certify a capacity law for real networks.

Coefficient budgets and evaluation goals. The scaling-law formulation uses equal-normalized effective updates, $\alpha_{i,K} = c/K$, and analyzes expected cross-entropy (Wang et al., 2026). Our exhaustive sweep fixes each TA coefficient at one and evaluates constituent-task metric degradation. In-the-wild merging additionally targets broad performance over heterogeneous checkpoints whose training objectives can overlap (Hitit et al., 2026). Because the coefficient budgets and evaluation goals differ, different task-count trends need not conflict.