PREDICTING AND REPAIRING MERGE COLLAPSE IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models fine-tuned from a shared base can be merged by averaging their task vectors, but some merges collapse far below the base model, and common merge operators give no warning before evaluation. We show that one statistic of the specialists' task vectors both predicts this collapse and calibrates its repair. The power that averaging removes equals the variance of the task vectors across specialists, our measure of interference. Under a working noise model, the disturbance that a merge injects grows with the merge coefficient and with interference, yielding a pre-merge score. In our experiments on twenty-two merge configurations from four model families, only destructive merges exceed a threshold on this score. We find that statistics of sign conflict between specialists, a common target of existing merge operators, are anti-predictive. We then predicted the outcomes of fourteen merges before evaluating them, and twelve predictions were correct, including the destructive outcome of a specialist pair pushed past the threshold by continued pretraining. To address this collapse, we introduce PRISM, an operator that averages the task vectors first and then soft-thresholds each layer at a level set by the layer's interference. Without data or tuning, PRISM keeps all five destructive merges above the threshold within evaluation noise of the base model, where plain averaging falls at least 14.4 points below it or collapses entirely. We apply PRISM only above the threshold and keep the plain average for merges below it, which include all fifteen harmless ones. Code will be released upon publication.

1 INTRODUCTION

Model merging combines several models fine-tuned from the same pretrained checkpoint into one model by operating directly on their weights, with no further training (Wortsman et al., 2022; Matena & Raffel, 2022; Ilharco et al., 2022). Task Arithmetic (Ilharco et al., 2023) represents each fine-tuned specialist by its task vector, the difference between its weights and those of the shared base model. A merge adds a scaled combination of task vectors back to the base, in the simplest case their average multiplied by a merge coefficient (Ilharco et al., 2023; Ortiz-Jimenez et al., 2023). Open-weight large language models (LLMs) now arrive as families of specialists for mathematics, code and other domains, built on shared bases (Yang et al., 2024a; Grattafiori et al., 2024; Yang et al., 2024b; Hui et al., 2024; Cao et al., 2026). Merging such specialists has become common practice, including in the training of production models (Team Cohere et al., 2025), and is supported by dedicated toolkits, benchmarks and studies at scale (Goddard et al., 2024; He et al., 2025; Yadav et al., 2024; Wang et al., 2026; Hitit et al., 2026).

When specialists are fine-tuned for different tasks, however, they change shared weights in different directions, and averaging cancels part of each change, an effect usually called interference (Yadav et al., 2023; Gargiulo et al., 2025). We call a merge destructive when it falls substantially below the base model from which its specialists were fine-tuned, and harmless when it stays near or above that base. The standard way to reduce interference is to discard part of each task vector before averaging (Yadav et al., 2023; Yu et al., 2024; Deep et al., 2024). Each such prune-then-merge operator sets its keep rate in advance, whatever the disagreement between the specialists. These operators, like methods that tune the merge on data (Matena & Raffel, 2022; Yang et al., 2023; Chopra et al., 2025) or derive it from the task vectors alone (Marczak et al., 2025; Cheng et al., 2025), leave open whether a merge will be destructive until it has been built and evaluated.

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.

We start from the identity underlying the ambiguity decomposition of ensembles (Krogh & Vedelsby, 1994). At every weight coordinate, the mean-squared power that averaging removes from the task vectors equals their variance across specialists. This variance is our measure of interference, and we compute it from the weights in the same pass as the average. Interference splits into cancellation, where specialists write comparable values of opposite sign, and dispersion, the remainder. Dispersion includes one-sided drift, in which one specialist moves a weight that the others leave near zero, and such drift can hold knowledge unique to that specialist (Wang et al., 2024; Lu et al., 2024). Under a working noise model, we show that the disturbance a merge injects into the base grows in proportion to the merge coefficient and to the square root of the interference. We call this estimate of the disturbance, available before merging, the interference score.

We introduce PRISM, a merge operator whose denoising threshold is set by interference, as Figure 1 illustrates. The operator averages the task vectors first and then soft-thresholds the average in each layer at the universal threshold of wavelet shrinkage (Donoho & Johnstone, 1994; Donoho, 1995). The noise level comes from the layer's interference, a floor keeps a small share of each matrix's coordinates, and one threshold scale serves all settings without data or tuning. A gate on the share of interference that comes from cancellation returns the plain average when that share is small, leaving one-sided drift intact.

Across twenty-two merge configurations from four model families, with up to eight specialists in one merge, each outcome is either destructive or harmless. Seven merges are destructive at the default coefficient, six of them between public specialists. A threshold on the interference score flags five of the seven and none of the fifteen harmless merges, whereas statistics of sign disagreement, the conflict that sign election resolves (Yadav et al., 2023), are anti-predictive. We then fixed the score and this threshold, predicted the outcomes of fourteen further merges before evaluation, and found twelve predictions correct.

Specialists are often built by continuing to pretrain a base model on domain text (Gururangan et al., 2020; Hui et al., 2024), and a practitioner who plans to merge them needs to know when more training would make the merge destructive. We therefore built two such specialists from one base model, one on Japanese web text and the other on biomedical abstracts, and tracked the pair's score as training proceeded. At a fixed learning rate, the score grew roughly as the square root of the number of training tokens. Continued training at a higher learning rate further improved each specialist on its own domain but pushed the pair past the threshold. Plain averaging then became destructive as predicted, while PRISM stayed near the base.

PRISM keeps each destructive merge above the threshold within evaluation noise of the base, where plain averaging falls far below it. The merge we study most closely combines a math specialist with a code specialist, as recent merging work does (Yu et al., 2024; Cheng et al., 2025; He et al., 2025). On this merge, PRISM leads Task Arithmetic, two prune-then-merge operators (Yadav et al., 2023; Deep et al., 2024) and a method that tunes the merge on data (Chopra et al., 2025) at a shared coefficient. When we sweep the merge coefficient, these baselines collapse at large values while PRISM degrades gradually. A merge whose coefficients are tuned by entropy minimization on unlabeled text (Yang et al., 2023) also falls far below the base at the shared coefficient. Shrinkage helps a destructive merge but can forgo the gain that plain averaging brings to a harmless one, and our recipe therefore averages plainly below the score threshold and applies PRISM above it. The score needs only the weights and minutes of CPU time, and we hope it makes a pre-merge check routine whenever public checkpoints are combined.

2 RELATED WORK

Task Arithmetic (Ilharco et al., 2022; 2023) established linear task-vector combination, and subsequent work has refined merging along three lines (Liu et al., 2026; Cheng et al., 2025; Wang et al., 2024; Gargiulo et al., 2025). Prune-then-merge methods include TIES-Merging, which combines magnitude trimming with cross-model sign election (Yadav et al., 2023), DARE, which drops entries at random and rescales the survivors (Yu et al., 2024), and DELLA, which drops entries with a probability that decreases with their magnitude (Deep et al., 2024).

Training-free approaches include model matching (Xu et al., 2024), hierarchical merging (Fu et al., 2025), and trust-region constraints (Sun et al., 2025). Fisher averaging (Matena & Raffel, 2022), AdaMerging (Yang et al., 2023), LEWIS (Chopra et al., 2025) and RAIN-Merging (Huang et al.,

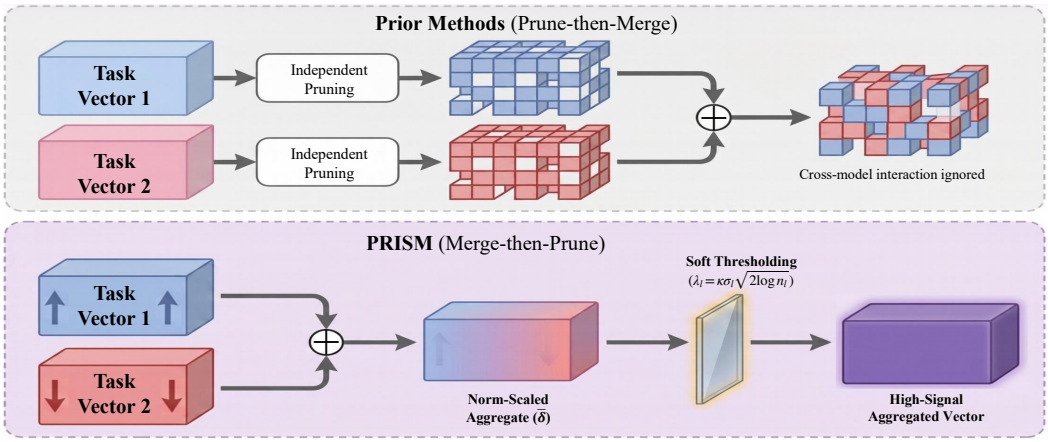

Figure 1: Merge then prune. **Top:** task vectors are pruned before aggregation. **Bottom:** PRISM averages first with norm scaling and then soft-thresholds the aggregate at the layer's interference scale. Full algorithm in Appendix A.

2026) use auxiliary data. NeuroMerging decomposes task representations into neuronal subspaces (Fang et al., 2025), and NeuronMerge groups neurons by function (Gu et al., 2025). EMR-Merging (Huang et al., 2024) uses task-specific masks, which requires knowing the task at test time.

Spectral and subspace methods use spectrum flattening (Marczak et al., 2025), task singular vectors (Gargiulo et al., 2025), or task-vector guidance for data-free interference reduction (Cheng et al., 2025). PACT preprocesses task vectors through subspace alignment and offers a randomized-SVD variant (Shi et al., 2026). Twin-Merging dynamically combines shared and exclusive knowledge according to the input (Lu et al., 2024).

We screen for collapse risk, a step absent from the merge operators evaluated here. The PRISM repair soft-thresholds the coordinates of the averaged task vector in linear time, rather than pruning each task vector or computing singular value decompositions. It sets the threshold from the weights alone, through an exact variance identity and a noise model.

3 Interference in Task-Vector Merging

3.1 Interference is the cross-model variance

Let θ_{base} denote the parameters of a pretrained base model and θ_k those of specialist $k \in \{1, \dots, K\}$, fine-tuned from it. The task vector $\tau_k = \theta_k - \theta_{\text{base}}$ has value $\tau_k(i)$ at coordinate i. A merge rescales each task vector by a weight $s_k > 0$, set by norm scaling in Section 4.3 or to one in a raw merge, and averages them into $\bar{\delta} = \frac{1}{K}\sum_k s_k\tau_k$. At coordinate i the total power $P(i) = \frac{1}{K}\sum_k (s_k\tau_k(i))^2$ splits into signal power $S(i) = \bar{\delta}(i)^2$ and *interference* $I(i) = P(i) - S(i)$, the reduction in mean-squared power due to averaging.

$$I(i) = \underbrace{\frac{1}{K}\sum_{k=1}^{K}\big(s_k\tau_k(i)\big)^2}_{P(i)} - \underbrace{\bar{\delta}(i)^2}_{S(i)} = \mathrm{Var}_k\big(s_k\tau_k(i)\big) \geq 0, \tag{1}$$

where Var_k is the population variance over the K specialists. Equation 1 is the ambiguity decomposition of Krogh & Vedelsby (1994) applied to the scaled task vectors, with the base model as the target. This variance measures task-vector disagreement, which we use as a proxy for task interference (Yadav et al., 2023; Gargiulo et al., 2025), and is computed alongside the mean. Interpreting it as noise requires a model. Throughout, *drift* means the displacement $s_k\tau_k$ of a specialist from the base. Identical scaled task vectors give zero interference regardless of the size or effect of their shared drift. Aligned vectors with unequal magnitudes can still have positive variance, as can *one-sided* drift, present in one specialist and absent in the others.

Its normalized read-out is the *Coherence Ratio*, $\mathrm{CR}(i) = S(i)/(P(i) + \epsilon) \in [0, 1]$ with $\epsilon = 10^{-10}$, a continuous generalization of sign agreement that stays meaningful when task vectors differ in

magnitude. For a layer l with n_l merged coordinates, the means $\overline{P}_l$, $\overline{S}_l$ and $\overline{I}_l$ of P, S and I over those coordinates give $\overline{\mathrm{CR}}_l = \overline{S}_l/(\overline{P}_l + \epsilon)$ and an interference scale

$$\sigma_l^2 = \overline{I}_l = \overline{P}_l - \overline{S}_l \simeq \overline{P}_l(1 - \overline{\mathrm{CR}}_l). \tag{2}$$

The last expression omits the stabilizer term $-\epsilon\overline{\mathrm{CR}}_l$. We compute σ_l from $\overline{P}_l - \overline{S}_l$. The standard deviation $\sigma_l = \sqrt{\overline{I}_l}$ has weight units and can set a threshold on task-vector coordinates. Computing these statistics takes about two CPU-minutes on a 7B model. Sign agreement tracks $\overline{\mathrm{CR}}_l$ almost exactly at the layer level ($r = 0.998$) but is a dimensionless fraction, so the threshold uses σ_l to supply the weight scale, as shown in Appendix E.

3.2 What a merge injects

The merge applies $c\,\bar{\delta}$, where c is an additional global multiplier on the already-averaged task vector. Our default $c = 1/K$ further shrinks that mean, while $c = 1$ applies it in full.

Proposition 1 (Injected interference)**.** *Assume that, at a coordinate of layer l, the scaled task vectors share a signal and have independent, zero-mean disturbances of common variance v_l. For a nonnegative coefficient c, adding $c\,\bar{\delta}$ introduces disturbance standard deviation $c\sqrt{v_l/K}$ and power $c^2 v_l/K$. Using $\overline{I}_l$ as an empirical proxy for v_l gives the screening scale $c\sqrt{\overline{I}_l/K}$ and power $c^2\overline{I}_l/K$, both available before merging.*

Argument, with the full version in Appendix B. Independence removes cross-covariance terms, giving disturbance variance v_l/K for the mean. Scaling by c multiplies this variance by c^2. Substituting $\overline{I}_l$ for v_l is our modeling choice. □

The mean update is an absolute perturbation of the base weights. Its estimated disturbance power grows as c^2 at fixed interference, the regime that the coefficient sweep of Section 5.3 examines.

3.3 Cancellation and dispersion

Treating all interference as noise may suppress useful specialist knowledge (Wang et al., 2024; Lu et al., 2024). Low coherence has two causes. *Cancellation* is specialists writing comparable magnitudes with opposing signs so the aggregate nearly vanishes. *Dispersion* includes one specialist writing a large value while the others sit near zero, task-exclusive drift that can carry useful signal. With $\mathrm{Pos}(i) = \sum_k \max(s_k\tau_k(i), 0)$ and $\mathrm{Neg}(i) = \sum_k \max(-s_k\tau_k(i), 0)$ the positive and negative mass at coordinate i, the cancellation power is

$$C(i) = \min\Big(\tfrac{2}{K}\min\big(\mathrm{Pos}(i), \mathrm{Neg}(i)\big)^2,\ I(i)\Big), \qquad D(i) = I(i) - C(i) \geq 0, \tag{3}$$

so that $I = C + D$. Summing over all merged coordinates yields the *empirical cancellation ratio* $\rho = \big(\sum_l n_l\overline{C}_l\big)/\big(\sum_l n_l\overline{I}_l + \epsilon_g\big) \in [0, 1]$, where $\overline{C}_l$ is the mean of C over layer l and $\epsilon_g = 10^{-12}$ is a separate gate stabilizer. High ρ marks a conflict-dominated merge and low ρ a dispersion-dominated one, and empirically ρ spans 0.003–0.720 across four families. The interference scale sets the initial threshold, and ρ determines whether shrinkage is applied.

4 Repairing a Dangerous Merge with Prism

Prism averages first and shrinks the aggregate using the interference scale σ_l of Eq. 2, as in Figure 1.

4.1 Interference-calibrated soft thresholding

Unlike prune-then-merge methods (Yadav et al., 2023; Deep et al., 2024), Prism averages first and then denoises the consensus with the Donoho–Johnstone universal soft threshold (Donoho & Johnstone, 1994), used here as a tuning-free default,

$$\tilde{\delta}(i) = \mathrm{sgn}(\bar{\delta}(i))\max(0,\ |\bar{\delta}(i)| - \lambda_l), \quad \text{where} \quad \lambda_l = \kappa\sigma_l\sqrt{2\log n_l}. \tag{4}$$

Substituting Eq. 2 gives $\lambda_l = \kappa\sqrt{2\overline{I}_l \log n_l}$ with the threshold scale κ fixed at 1.0. At fixed total power, high-coherence layers ($\overline{\mathrm{CR}}_l \to 1$) have smaller interference scales and thresholds. A coordinate survives only when its aggregate magnitude exceeds λ_l, and its magnitude is then reduced by λ_l in Algorithm 1. Prism shrinks the aggregate once and then scales the resulting vector by c.

When fewer than 0.1% of a projection matrix's coordinates survive, a keep-rate floor lowers the threshold to the smallest magnitude among its top 0.1% and applies soft thresholding again.

4.2 THE ρ-GATE

Calibrating by $\sigma_l = \sqrt{\bar{I}_l}$ treats interference as destructive conflict. For one-sided drift, where one specialist moves while the others do not, σ_l can be large even when the update carries useful specialist signal. The cancellation ratio ρ of Section 3.3 measures how much interference comes from opposing signs, so we gate on it with a threshold τ_g,

$$\lambda_l = \begin{cases} \kappa\, \sigma_l \sqrt{2 \log n_l} & \text{if } \rho \geq \tau_g, \\ 0 & \text{if } \rho < \tau_g, \end{cases} \qquad \tau_g = 0.1. \tag{5}$$

Under substantial cancellation ($\rho \geq \tau_g$), PRISM applies soft thresholding. Under dispersion ($\rho < \tau_g$), it sets $\lambda_l = 0$, leaves the mean unchanged, and yields the Task-Arithmetic model with the same scaling and coefficient. Fallback examples from two families include the raw Llama-3.1-8B $K=3$ merge at $\rho = 0.021$, the Mistral-7B $K = 2$ merge at $\rho = 0.093$ and Tulu-3 with its own direct preference optimization (DPO) variant at $\rho = 0.003$. With the reported ratios, any threshold strictly between 0.093 and 0.229 makes the same decisions as the default $\tau_g = 0.1$ on every setting here.

4.3 MERGE PROCEDURE AND SCOPE

The merged model is $\theta_{\text{merged}} = \theta_{\text{base}} + c\tilde{\delta}$ with norm scaling $s_k = \min_j \|\tau_j\|_2 / \|\tau_k\|_2$, where the norms run over the merged coordinates. We apply PRISM to the multilayer perceptron (MLP) projections (`gate/up/down_proj`) and leave attention at its base values in the default comparison, since MLP layers are the primary locus of domain-specific knowledge, holding key–value memories with localized knowledge neurons (Geva et al., 2022; Dai et al., 2022), where interference should concentrate. Restricting PRISM to the MLP projections stays within 0.1pp of the six-task mean obtained by extending it to attention, which merges $2.3\times$ more keys, as reported in Appendix G.

5 EXPERIMENTS

5.1 SETUP

As in MergeBench (He et al., 2025), we vary model family and specialist domain, and we also vary the number of merged specialists. We merge specialists from two 7–8B families, Qwen2.5 (Yang et al., 2024a) and Llama-3.1 (Grattafiori et al., 2024), adding Mistral-7B and DeepSeek-7B for the breadth study. The main setting pairs Qwen2.5 Math (Yang et al., 2024b) with Coder (Hui et al., 2024) at $K = 2$, extended to $K = 3$ (Math, Coder, Instruct) and $K = 1$. For the shared task-vector protocol, we norm-scale by $s_k = \min_j \|\tau_j\|_2 / \|\tau_k\|_2$ before averaging. By default, we apply an additional global coefficient $c = 1/K$ to that average under Appendix B's convention.

Baselines are Task Arithmetic (Ilharco et al., 2023), TIES (Yadav et al., 2023), DELLA (Deep et al., 2024), DARE (Yu et al., 2024), LEWIS (Chopra et al., 2025) with the activation-based variant and weight-sparsity fallback compared in Appendix D.7, AdaMerging (Yang et al., 2023), Greedy Soup (Wortsman et al., 2022), Fisher-Weighted Averaging (Matena & Raffel, 2022) and the 2025 data-free Iso-C (Marczak et al., 2025) and WUDI-Merging (Cheng et al., 2025). Headline task-vector comparisons use the shared coefficient, with baseline settings and sweeps summarized in Appendix D.9, while PRISM stays at its data-free default $\kappa = 1.0$. SLERP (Shoemake, 1985) and Model Stock (Jang et al., 2024) use their published recipes without our coefficient or norm scaling.

Our evaluation spans six benchmarks, GSM8K (Cobbe et al., 2021), ARC-Challenge (Clark et al., 2018), MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), HellaSwag (Zellers et al., 2019) and WinoGrande (Sakaguchi et al., 2020), under `lm-evaluation-harness` (Gao et al., 2024), plus WikiText-2 perplexity (Merity et al., 2016), with GSM8K under strict-match extraction. Every reported mean is the average of the six task scores shown, and differences are absolute percentage points (pp). Generation metrics on GSM8K, MATH-500 (Lightman et al., 2023), HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) use greedy decoding with one sample for each problem. A six-task mean carries a propagated standard error of at most 0.8pp, so we treat sub-pp gaps as ties under Appendix D.4's uncertainty estimate.

Table 1: Qwen2.5-7B Math+Coder base checkpoints. Generation: GSM8K 5-shot, MATH-500 4-shot, harness HumanEval/MBPP pass@1 with completion prompts. LEWIS uses activation-based six-task/HumanEval/MBPP results and weight-sparsity GSM8K/MATH generation. Bold: best merged; underline: second best; †: calibration for activation-based LEWIS.

	Six-task evaluation							Generation			
Method	GSM8K	ARC	MMLU	TQA	HellaS.	WinoG.	**Avg.**	GSM8K	MATH	HumanE.	MBPP
Math specialist	84.4	49.2	60.2	47.9	58.8	68.4	61.5	83.8	53.6	34.1	56.2
Coder specialist	81.8	49.0	69.5	49.8	66.4	72.0	64.8	82.4	41.4	59.8	68.4
Task Arithmetic	22.0	47.0	55.7	50.0	55.4	66.6	49.5	23.0	5.8	13.4	27.4
TIES	67.2	**53.4**	64.3	50.2	60.6	71.0	61.1	65.8	24.6	17.7	43.0
DELLA	65.2	51.6	64.2	50.3	59.8	70.4	60.3	64.0	22.0	17.7	42.4
LEWIS†	43.0	44.2	69.1	50.7	62.8	66.6	56.1	74.2	15.4	26.2	52.6
PRISM (ours)	**81.4**	51.4	**74.3**	**55.2**	**67.4**	**77.0**	**67.8**	**82.4**	**30.2**	**51.2**	**62.8**

Table 2: Qwen2.5-1.5B Math+Coder instruction-tuned checkpoints. Columns, marks and LEWIS variants follow Table 1.

	Six-task evaluation							Generation			
Method	GSM8K	ARC	MMLU	TQA	HellaS.	WinoG.	**Avg.**	GSM8K	MATH	HumanE.	MBPP
Task Arithmetic	4.6	32.4	35.7	**49.4**	43.4	57.0	37.1	4.0	1.8	0.0	0.0
TIES	18.8	37.6	41.2	46.1	50.6	61.2	42.6	19.0	3.8	11.6	16.2
LEWIS†	49.4	40.2	57.8	46.4	55.4	64.8	52.3	35.6	13.6	28.0	38.6
PRISM (ours)	**59.6**	**42.6**	**61.8**	45.5	**59.0**	**66.4**	**55.8**	**59.0**	**26.6**	**35.4**	**43.0**

5.2 MERGING UNDER SEVERE CONFLICT

In Table 1, PRISM leads Task Arithmetic, TIES and DELLA on the six-task mean by at least 6.7pp and exceeds calibration-guided LEWIS without using calibration data. Task Arithmetic falls to 13.4 on HumanEval, whereas PRISM leads every baseline on all four generation metrics, by at least 25.0pp on HumanEval. With instruction-tuned specialists, PRISM again leads these baselines in Appendix D.4 and posts the best training-free coding scores in Appendix H.

Model scale. Table 2 repeats the setting on Qwen2.5-1.5B. Task Arithmetic collapses outright, and PRISM leads the strongest baseline by 3.5pp. Appendix D.6 adds Instruct+Coder-Instruct merges at 3B and 14B, where PRISM gains 0.8 and 0.7pp over the base.

Vendor-native coding evaluation. Under vendor-native EVALPLUS evaluation, Task Arithmetic, TIES, DELLA, DARE, LEWIS and SLERP score at most 0.3 on all four coding metrics in Table 3, while Model Stock reaches 0.6. PRISM retains coding capability across these metrics. Section 5.4 compares PRISM with the base and a norm-matched control on these metrics.

5.3 ROBUSTNESS TO THE SCALING COEFFICIENT

Proposition 1 motivates a c^2 increase in injected interference power under its working model. PRISM shrinks the update before applying the coefficient. In Figure 2, PRISM holds a 64.8–68.0% mean across the sweep. At $c = 0.7$, the four evaluated baselines sit 19 to 33 points below the base at perplexities from 100 to 961. In the generation evaluation at $c = 1.0$, Task Arithmetic, TIES and DELLA score zero on GSM8K, while LEWIS reaches 17.4 in Appendix D.11. The best baseline in the plotted sweep, LEWIS at 67.6 ($c = 0.3$), is tied within evaluation uncertainty with the 67.8 that untuned PRISM posts at the shared default.

5.4 IS IT JUST CONSERVATIVE SHRINKAGE?

We test whether generic shrinkage (Yu et al., 2024) explains the preserved capability with a control matched to PRISM's update norm. Table 3 reports each update's norm relative to the plain average. The baseline operators and community defaults write updates with 46 to 285% of that norm, while PRISM writes 2.2%. For the matched control, one scalar multiplies every coordinate of the plain average to reach the same norm.

The norm-matched control avoids collapse at both evaluated coefficients and stays close to the unmerged base on the six-task mean, perplexity and all four coding metrics. Its norm, however, is

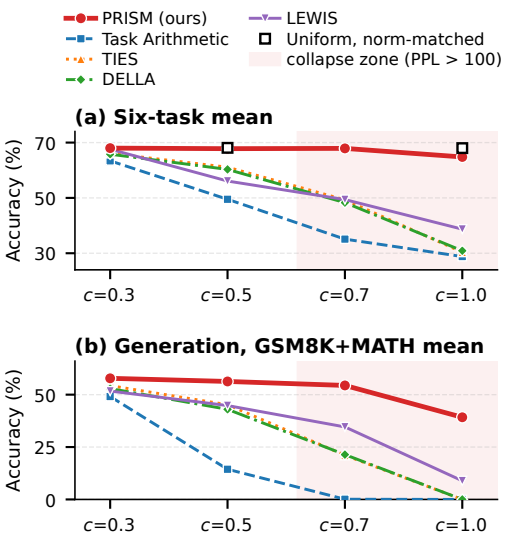

Figure 2: Coefficient sweep on Qwen2.5-7B Math+Coder: **(a)** six-task evaluation and **(b)** separately prompted generation. Squares mark the norm-matched control.

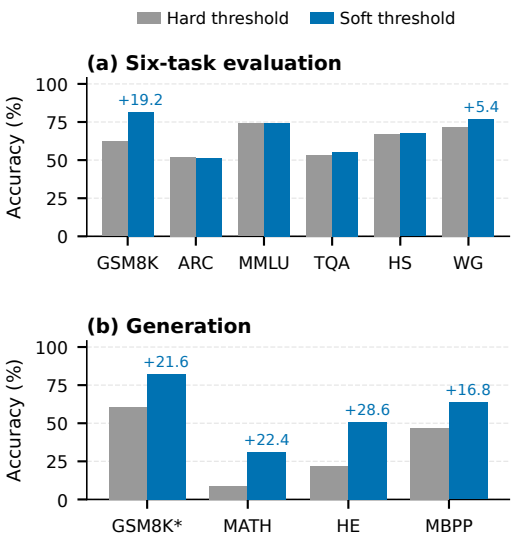

Figure 3: Soft and hard thresholding at the same interference scale on Qwen2.5-7B Math+Coder. Labels mark gaps above five points; GSM8K* uses the generation-format evaluation.

Table 3: Qwen2.5-7B Math+Coder. Coding: vendor-native EvalPlus pass@1 at $c = 0.5$ where applicable, not comparable with Table 1. LEWIS: activation-based Mean/PPL, weight-sparsity fallback for Norm/Coords/Coding. Bold: above the unmerged base. §: rescaled; †: calibration for activation-based LEWIS; SLERP and Model Stock: no coefficient.

	Norm	Coords	Mean		PPL	Coding (EvalPlus)			
Method	$\|\delta\|/\|\bar{\delta}\|$	%	$c=0.5$	$c=1$	$c=1$	HumanEval	HumanEval+	MBPP	MBPP+
Base model, unmerged	0	0.0	68.0	68.0	8.6	62.8	57.9	69.8	59.5
Task Arithmetic	1.00	99.7	49.5	28.8	4.6×10^5	0.0	0.0	0.0	0.0
TIES	0.80	34.8	61.1	30.4	1.1×10^4	0.0	0.0	0.0	0.0
DELLA	0.82	34.9	60.3	30.9	1.6×10^4	0.0	0.0	0.0	0.0
DARE (drop 0.9)§	2.85	19.0	30.0	29.8	7.2×10^6	0.0	0.0	0.0	0.0
LEWIS†	0.55	5.0	56.1	38.7	5.9×10^3	0.0	0.0	0.3	0.0
SLERP (t=0.5)	1.27	99.9	30.0		1.0×10^6	0.0	0.0	0.0	0.0
Model Stock	0.46	99.6	53.2		35.0	0.6	0.6	0.0	0.0
Uniform, norm-matched	0.022	85.1	68.1	68.0	8.7	62.2	57.3	69.6	59.0
Hard threshold, same λ	0.126	0.1	63.4	57.2	18.4	51.8	48.2	60.1	51.6
PRISM	0.022	0.1	67.8	64.8	11.1	**65.9**	**59.1**	**71.2**	**61.1**

taken from PRISM, whose thresholds set that norm from each layer's interference without data or tuning. Where the control spreads this norm across 85.1% of the coordinates, PRISM concentrates it on the 0.1% retained by thresholding. PRISM's coding scores are higher than both the control's and the base's on all four EvalPlus metrics, reaching 65.9 on HumanEval against 62.2 and 62.8.

5.5 Data-dependent baselines need data to reach parity

AdaMerging reaches 67.9 at its best point, a tie with PRISM's 67.8, after entropy minimization on unlabeled C4 text and a coefficient sweep. Greedy Soup adds neither specialist and returns the base model, and Fisher-weighted averaging falls to 47.5 in Appendix D.8. Among data-free competitors, Iso-C reaches 66.6 when tuned away from its collapsing default, and WUDI-Merging diverges in every configuration of Appendix D.10. SLERP at $t = 0.5$ (Shoemake, 1985; Goddard et al., 2024) writes 1.27× the norm-scaled average and collapses. Model Stock (Jang et al., 2024) applies 2.4× interpolation shrinkage toward the base and still loses 14.8pp.

5.6 Ablations and mechanism

On Qwen $K = 2$, soft thresholding beats hard truncation at the same threshold by 4.4pp on the six-task mean in Table 3. Truncation's cost falls almost entirely on tasks that produce text, at 19.2pp

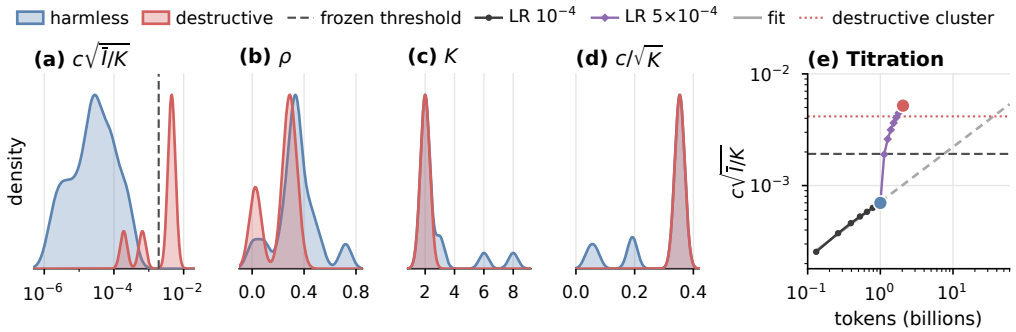

Figure 4: **(a)** Densities of the interference score by outcome. **(b–d)** Densities of ρ, K and $c/\sqrt{K}$ overlap. **(e)** At the lower learning rate, the score grows roughly as the square root of specialist training tokens. Large markers are evaluated merges.

Table 4: The interference score exceeds the threshold of 1.9×10^{-3} only for destructive merges. Means are six-task accuracies, and self-SFT specialists are fine-tuned by us from the base. ⋆Outcome predicted before evaluation. ^{g}Gated, so PRISM equals Task Arithmetic (TA).

Setting	K	c	ρ	$c\sqrt{\overline{I}/K}$ ($\times10^{-3}$)	Base	TA	PRISM
naive averaging destroys the model							
Qwen2.5-1.5B ja + med, constructed⋆	2	0.50	0.342	5.19	56.1	35.8	55.9
Qwen2.5-1.5B Math + Coder	2	0.50	0.246	4.94	56.3	37.1	55.8
Qwen2.5-7B Math + Coder	2	0.50	0.293	4.25	68.0	49.5	67.8
Qwen2.5-7B R1-Distill + Coder⋆	2	0.50	0.291	4.24	68.1	49.7	67.7
Qwen2.5-7B Math-PRM + Coder⋆	2	0.50	0.285	4.17	68.1	53.7	67.6
Qwen2.5-7B Math + Finance⋆	2	0.50	0.036^{g}	0.653	68.1	37.0	37.0^{g}
Qwen2.5-7B Coder + Coder-Instruct⋆	2	0.50	0.013^{g}	0.190	68.1	39.3	39.3^{g}
merging is harmless							
DeepSeek-7B math + chat	2	0.50	0.348	0.228	46.8	48.1	46.6
Qwen2.5-7B self-SFT at 10× learning rate⋆	2	0.50	0.349	0.227	68.1	68.8	68.2
Llama-3.1-8B R1-Distill + Tulu-3⋆	2	0.50	0.346	0.089	59.3	62.6	59.6
Qwen2.5-7B Aloe medical + Math⋆	2	0.50	0.279	0.082	68.1	68.0	68.1
Qwen2.5-7B HuatuoGPT + R1-Distill⋆	2	0.50	0.349	0.076	68.1	68.9	68.1
Qwen2.5-7B Math + Coder + Instruct	3	0.33	0.443	0.042	68.0	68.6	68.1
Mistral-7B MetaMath + Code	2	0.50	0.093^{g}	0.030	57.4	59.9	59.9^{g}
Qwen2.5-7B self-SFT, $K=2$	2	0.50	0.333	0.027	68.1	68.4	68.2
Mistral-7B SaulLM legal + BioMistral⋆	2	0.50	0.322	0.025	57.3	57.8	57.4
Qwen2.5-7B Math + EVA roleplay⋆	2	0.50	0.229	0.019	68.1	68.1	68.1
Llama-3.1-8B Swallow ja + Guard⋆	2	0.50	0.290	0.012	60.2	60.0	59.7
Llama-3.1-8B OpenMath2 + Hermes-3 + Guard	3	0.33	0.435	0.008	60.2	60.0	59.7
Qwen2.5-7B off-the-shelf ×6	6	0.17	0.516	0.005	68.1	68.0	68.2
Qwen2.5-7B self-SFT, $K=8$	8	0.13	0.720	0.003	68.1	68.0	68.0
Llama-3.1-8B Tulu-3 + Tulu-3-DPO⋆	2	0.50	0.003^{g}	0.002	59.3	64.5	64.5^{g}

on GSM8K and up to 28.6pp on HumanEval against 1.4pp averaged over the five multiple-choice benchmarks in Figure 3. Across threshold scales from $\kappa=0.25$ to 3, the six-task mean varies by at most 0.5pp, and the default scores highest in Table 6. Replacing soft thresholding with magnitude pruning of the average at coherence-adapted layer rates lowers the mean by 3.7pp at 7B and 6.9pp at 1.5B, and Appendix D.5 shows that the gap closes at $K=3$, where interference is lower.

Appendix C examines the layer-retention pattern. Where the calibrated threshold rather than the floor sets the keep rate ($\kappa=0.5$), layer retention p_l tracks coherence data-free at $r=0.91$. Low coherence mixes sign conflict with task-exclusive drift, which the split of Eq. 3 separates, and the ρ-gate returns the plain average when dispersion dominates. Appendix F shows that the coordinates removed by thresholding are predominantly conflict elements. In our low-rank simulation using SVD projections of full task vectors, PRISM stays within 0.6pp of its full-weight six-task mean across $r\in\{8,\dots,128\}$ in Appendix D.12.

5.7 THE SAME SCALE SUPPORTS PRE-MERGE SCREENING

The same $\overline{I}$ that sets the threshold defines the pre-merge score $c\sqrt{\overline{I}/K}$ motivated by Proposition 1. We test whether this score separates destructive and harmless merges.

Observed discrimination. Table 4 and Figure 4(a–d) cover twenty-two configurations across four families and $K\in[2,8]$. Seven are destructive (perplexity >100 or mean >10pp below base). Fifteen are harmless (within 1pp or better), with no intermediate cases. The score exceeds

the threshold of 1.9×10^{-3} for five of the seven destructive merges and for none of the harmless ones. We compute $c\sqrt{\overline{I}/K}$ from each evaluation's scaled task vectors and convention, with $\overline{I}$ the parameter-count-weighted MLP mean interference, detailed in Appendix J.

Rival statistics. Interference, total power and signal power reach AUC 0.98 or higher, and interference also sets the repair's variance scale. Mean incoherence $1-\overline{\mathrm{CR}}$ and ρ are anti-predictive.

The gate. For raw Llama-3.1-8B $K=3$, gating ($\rho=0.021$, $\lambda=0$) raises ungated PRISM from 59.6 to Task Arithmetic's 62.0 (base 60.2), as Table 19 shows. On Mistral-7B $K=2$ at $\rho=0.093$, the gate likewise returns Task Arithmetic's 59.9, 2.5pp above the base, in Table 4.

A prospective test. With the statistic, convention and threshold fixed, twelve of the fourteen predictions made before evaluation were correct across three model families, including the harmless and destructive endpoints of the constructed pair. On the other two merges predicted to be destructive, R1-Distill+Coder and Math-PRM+Coder, naive averaging falls 18.4 and 14.4pp below the base in Table 4, whereas PRISM keeps both within 0.5pp of the base.

A controlled titration. We continued pretraining two Qwen2.5-1.5B specialists on disjoint Japanese web text and biomedical abstracts. At learning rate 10^{-4}, eight measured checkpoints follow an approximate square-root law through a harmless endpoint at one billion tokens for each specialist in Figure 4. Extrapolated, the fitted law crosses the threshold near 7.5 billion tokens and reaches 4.17×10^{-3}, the lower edge of the destructive cluster in Table 4, near 36 billion. Continuing at $5\times$ the rate reaches that cluster within a further billion tokens and improves own-domain perplexity (Japanese 16.7 $\rightarrow$ 12.2, PubMed 7.2 $\rightarrow$ 7.0). As predicted before evaluation, the merge at this endpoint is destructive. Naive averaging loses 20.3pp against the base, while PRISM stays within 0.3pp of it.

The recipe. Reported PRISM results above evaluate the denoising operator, including its $\rho<\tau_g$ bypass gate. Our recipe uses naive averaging below the fixed score threshold and Eq. 4's gated shrinkage above. All fifteen harmless configurations in Table 4 scored below the threshold and therefore keep the plain average, which scores 62.6 against the base's 59.3 on Llama-3.1-8B R1-Distill with Tulu-3. The twenty-eight public pairs without a code specialist that we screened across three families all scored below half the threshold, as Appendix J details.

6 CONCLUSION

We have shown that the variance of task vectors across specialists both predicts which merges collapse and calibrates their repair. Across twenty-two merges from four model families, the interference score built on this variance exceeded its threshold only for destructive merges, and twelve of the fourteen predictions made before evaluation were correct. PRISM, which soft-thresholds the average of the task vectors, kept above-threshold destructive merges near the base without data or tuning. Because the score grows with training, model developers can track it to keep their specialists mergeable.

AI USE STATEMENT

A large language model assisted with LaTeX tables, and language polishing. The authors reviewed and verified all outputs and take full responsibility for the paper.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. This work merges publicly released checkpoints and specialists we trained from public base models, and it evaluates the merges on public benchmarks. It involves no human subjects and collects no personal data. The continued pretraining uses public Japanese web text from FineWeb-2 and PubMed abstracts, as described in Appendix J. Merging can combine the weaknesses of its constituents along with their capabilities: pruning parameter updates could remove fragile safety alignment or amplify biases that conflict across specialists. Our screen and evaluations target capability benchmarks and do not measure safety behavior, so merged models need separate safety and alignment evaluation before deployment (Lee et al., 2022).

Reproducibility Statement

Algorithm 1 in Appendix A gives the complete PRISM procedure, and Sections 3 and 4 define the quantities it computes. Appendix B gives the full argument for Proposition 1, including its working noise model. Section 5.1 specifies the model families, baselines and evaluation protocol. Appendix D.9 lists the baseline settings and sweeps, Appendix G the merged parameter scope and memory footprint, Appendix I the EvalPlus scaffold commit, prompt and decoding settings, and Appendix D.4 the standard errors used to treat small gaps as ties. Appendix J details the breadth study and the prospective test, together with the full continued-pretraining recipe, including dataset splits, sequence packing and optimizer settings.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Ruisheng Cao, Mouxiang Chen, Jiawei Chen, Zeyu Cui, Yunlong Feng, Binyuan Hui, Yuheng Jing, Kaixin Li, Mingze Li, Junyang Lin, et al. Qwen3-Coder-Next technical report. *arXiv preprint arXiv:2603.00729*, 2026. URL https://arxiv.org/abs/2603.00729.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 10121–10143. PMLR, 2025. URL https://proceedings.mlr.press/v267/cheng25h.html.

Hetarth Chopra, Vidhi Rambhia, and Vikram Adve. Lewis (layer wise sparsity) – a training free guided model merging approach. *arXiv preprint arXiv:2503.03874*, 2025. URL https://arxiv.org/abs/2503.03874.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

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

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. Della-merging: Reducing interference in model merging through magnitude-based sampling. *arXiv preprint arXiv:2406.11617*, 2024. URL https://arxiv.org/abs/2406.11617.

David L Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3): 613–627, 1995.

David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. doi: 10.1093/biomet/81.3.425. URL https://doi.org/10.1093/biomet/81.3.425.

Zitao Fang, Guodong Du, Shuyang Yu, Yifei Guo, Yiwei Zhang, Yiyao Cao, Jing Li, Ho-Kin Tang, and Sim Kuan Goh. To see a world in a spark of neuron: Disentangling multi-task interference for training-free model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15720–15740, 2025. doi: 10.18653/v1/2025.emnlp-main.793. URL https://aclanthology.org/2025.emnlp-main.793/.

Zichuan Fu, Xian Wu, Yejing Wang, Wanyu Wang, Shanshan Ye, Hongzhi Yin, Yi Chang, Yefeng Zheng, and Xiangyu Zhao. Training-free llm merging for multi-task learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33111–33124, 2025. doi: 10.18653/v1/2025.acl-long.1588. URL https://aclanthology.org/2025.acl-long.1588/.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness. Zenodo, 2024. URL https://zenodo.org/records/12608602. Version v0.4.3.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18695–18705, 2025. doi: 10.1109/CVPR52734.2025.01742. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gargiulo_Task_Singular_Vectors_Reducing_Task_Interference_in_Model_Merging_CVPR_2025_paper.html.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, 2022. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's mergekit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 477–485, 2024. doi: 10.18653/v1/2024.emnlp-industry.36. URL https://aclanthology.org/2024.emnlp-industry.36/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Wangyun Gu, Qianghua Gao, Li-Xin Zhang, Xu Shen, and Jieping Ye. Neuronmerge: Merging models via functional neuron groups. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9015–9037, 2025. doi: 10.18653/v1/2025.findings-acl.471. URL https://aclanthology.org/2025.findings-acl.471/.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

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*, volume 38, 2025. doi: 10.52202/085713-3374. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/91f7f71cb04699f387dc863da42a1fe3-Abstract-Datasets_and_Benchmarks_Track.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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://openreview.net/forum?id=6zSIyrqS7J.

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

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging. *Advances in Neural Information Processing Systems*, 37:122741–122769, 2024. URL https://arxiv.org/abs/2405.17461.

Zhehao Huang, Yuhang Liu, Baijiong Lin, Yixin Lou, Zhengbao He, Hanling Tian, Tao Li, and Xiaolin Huang. Rain-merging: A gradient-free method to enhance instruction following in large reasoning models with preserved thinking format. *arXiv preprint arXiv:2602.22538*, 2026. URL https://arxiv.org/abs/2602.22538.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024. URL https://arxiv.org/abs/2409.12186.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29262–29277, 2022. doi: 10.52202/068431-2122. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/bc6cddcd5d325e1c0f826066c1ad0215-Abstract-Conference.html.

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. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model stock: All we need is just a few fine-tuned models. In *Computer Vision – ECCV 2024*, volume 15102 of *Lecture Notes in Computer Science*, pp. 207–223. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72784-9_12. URL https://link.springer.com/chapter/10.1007/978-3-031-72784-9_12.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, 1994. URL https://proceedings.neurips.cc/paper_files/paper/1994/hash/b8c37e33defde51cf91e1e03e51657da-Abstract.html.

Jungseob Lee, Midan Shim, Suhyune Son, Chanjun Park, Yujin Kim, and Heuiseok Lim. There is no rose without a thorn: Finding weaknesses on blenderbot 2.0 in terms of model, data and user-centric approach, 2022. URL https://arxiv.org/abs/2201.03239. arXiv preprint arXiv:2201.03239.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21558–21572, 2023. doi: 10.52202/075280-0943. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract.html.

Zehao Liu, Kun Li, and Wei Zhou. Globa: Rethinking parameter conflicts in model merging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23954–23962, 2026. doi: 10.1609/aaai.v40i28.39572. URL https://ojs.aaai.org/index.php/AAAI/article/view/39572.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. Twin-merging: Dynamic integration of modular expertise in model merging. In *Advances in Neural Information Processing Systems*, volume 37, pp. 78905–78935, 2024. doi: 10.52202/079017-2505. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8fcd17eb91bae20d9826786d7d6be799-Abstract-Conference.html.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 43177–43199. PMLR, 2025. URL https://proceedings.mlr.press/v267/marczak25a.html.

Michael S. Matena and Colin A. Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17703–17716, 2022. doi: 10.52202/068431-1287. URL https://doi.org/10.52202/068431-1287.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. URL https://arxiv.org/abs/1609.07843.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2023.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. doi: 10.1609/aaai.v34i05.6399. URL https://doi.org/10.1609/aaai.v34i05.6399.

Ningyuan Shi, Zhipeng Zhou, Hao Wang, Chunyan Miao, and Peilin Zhao. Pact: Preserving anchored cores in task-vectors for model merging. *arXiv preprint arXiv:2606.18627*, 2026. URL https://arxiv.org/abs/2606.18627.

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 245–254, 1985. doi: 10.1145/325334.325242. URL https://doi.org/10.1145/325334.325242.

Wenju Sun, Qingyong Li, Wen Wang, Yangliao Geng, and Boyang Li. Task arithmetic in trust region: A training-free model merging approach to navigate knowledge conflicts. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5178–5187, 2025. doi: 10.1145/3746027.3755789. URL https://doi.org/10.1145/3746027.3755789.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, et al. Command A: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*, 2025. URL https://arxiv.org/abs/2504.00698.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 50268–50287. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24k.html.

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 *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=VkTfx8FJoq.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya. Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 23965–23998. PMLR, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Zhengqi Xu, Ke Yuan, Huiqiong Wang, Yong Wang, Mingli Song, and Jie Song. Training-free pretrained model merging. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5915–5925, 2024. doi: 10.1109/CVPR52733.2024.00565. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xu_Training-Free_Pretrained_Model_Merging_CVPR_2024_paper.html.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115, 2023. doi: 10.52202/075280-0310. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. What matters for model merging at scale? *arXiv preprint arXiv:2410.03617*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a. URL https://arxiv.org/abs/2412.15115.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024b. URL https://arxiv.org/abs/2409.12122.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. *arXiv preprint arXiv:2310.02575*, 2023. URL https://arxiv.org/abs/2310.02575.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 57755–57775. PMLR, 2024. URL https://proceedings.mlr.press/v235/yu24p.html.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. doi: 10.18653/v1/P19-1472. URL https://doi.org/10.18653/v1/P19-1472.

A Full PRISM Algorithm

Algorithm 1 gives the complete PRISM-Denoise procedure summarized in Section 4 and illustrated in Figure 1.

The floor is checked separately for each MLP projection matrix. Numerical safeguards clamp the threshold variance at 10^{-12} and the logarithm's element count at 2; the gate denominator uses $\epsilon_g = 10^{-12}$. Norm scaling uses a 10^{-10} denominator stabilizer and returns unit scales when all task vectors vanish. Layer coherence also uses 10^{-10} in its denominator; $\overline{I}_l = \overline{P}_l - \overline{S}_l$ is computed directly, with negative roundoff clipped to zero.

B Denoise-Once-Then-Amplify: Full Argument

Proposition 1 in Section 3.2 combines an exact scaling identity with a working noise model. Equation 1 defines $I = P - S$ as the empirical variance across scaled task vectors.

For the working model, write $s_k\tau_k(i) = \mu(i) + \eta_k(i)$, with zero-mean disturbances that are uncorrelated across specialists and have a common variance v_l at a coordinate of layer l. Then $\mathrm{Var}(K^{-1}\sum_k \eta_k(i)) = v_l/K$. Applying c after this average gives standard deviation $c\sqrt{v_l/K}$ and variance $c^2 v_l/K$. Replacing the unknown v_l by the measured $\overline{I}_l$ gives the plug-in screening scale

Algorithm 1 PRISM-Denoise: Interference-Calibrated Soft Thresholding

Require: Base θ_{base}, specialists $\{\theta_k\}_{k=1}^{K}$, threshold scale κ, coefficient c, gate τ_g
Ensure: Merged model θ_{merged}

// Step 1: Task Vectors and Norm Scaling
$\tau_k \leftarrow \theta_k - \theta_{\text{base}}$ for all k; restrict to merged MLP coordinates
$s_k \leftarrow \min_j \|\tau_j\|/\|\tau_k\|$ for all k (or $s_k = 1$ for raw merges)
// Step 2: Aggregation, Layer Interference, and Cancellation
for each MLP layer l, containing n_l merged coordinates **do**
  $\bar{\delta}_l \leftarrow \frac{1}{K}\sum_k s_k \tau_{k,l}$
  $\sigma_l \leftarrow \sqrt{\overline{I}_l} = \sqrt{\overline{P}_l - \overline{S}_l}$ (empirical interference scale)
  $\overline{C}_l \leftarrow \text{mean}_i \ \min\big(\frac{2}{K}\min(\text{Pos}(i), \text{Neg}(i))^2,\ I(i)\big)$
  $\lambda_l \leftarrow \kappa\sigma_l\sqrt{2\log n_l}$
end for
// Step 3: Bypass Thresholding When Cancellation Is Low
$\rho \leftarrow \big(\sum_l n_l \overline{C}_l\big)/\big(\sum_l n_l \overline{I}_l + \epsilon_g\big)$
if $\rho < \tau_g$ **then**
  $\lambda_l \leftarrow 0$ for all l (Task Arithmetic with the same s_k and c)
end if
// Step 4: Soft Thresholding with a Keep-Rate Floor
$\tilde{\delta} \leftarrow 0$; $p_{\min} \leftarrow 0.1\%$
for each MLP layer l **do**
  for each projection matrix m in layer l, containing $n_{l,m}$ coordinates **do**
    $\tilde{\delta}_{l,m} \leftarrow \text{sgn}(\bar{\delta}_{l,m})\max(0, |\bar{\delta}_{l,m}| - \lambda_l)$ (elementwise)
    if $\|\tilde{\delta}_{l,m}\|_0/n_{l,m} < p_{\min}$ **then**
      $k_{l,m} \leftarrow \max(\lfloor p_{\min} n_{l,m} \rfloor, 1)$
      $q_{l,m} \leftarrow$ the $k_{l,m}$-th largest magnitude in $|\bar{\delta}_{l,m}|$
      $\tilde{\delta}_{l,m} \leftarrow \text{sgn}(\bar{\delta}_{l,m})\max(0, |\bar{\delta}_{l,m}| - q_{l,m})$
    end if
  end for
end for
return $\theta_{\text{base}} + c\tilde{\delta}$

$c\sqrt{\overline{I}_l/K}$ and its squared value $c^2\overline{I}_l/K$. In the implementation the mean already divides by K: $c = 1$ writes the full average, and our default $c = 1/K$ shrinks it further.

For any fixed update, multiplication by c scales its squared norm by c^2, and PRISM thresholds the average before this multiplication. In the coefficient sweep of Section 5.3, baselines deteriorate at $c \geq 0.7$, with the worst reaching perplexity 961 at $c = 0.7$, while PRISM retains low perplexity across the tested sweep.

C MECHANISM ANALYSIS

Figure 5 examines the Qwen2.5-7B Math+Coder merge at $\kappa = 0.5$. Here the calibrated threshold sets the keep rate above the 0.1% floor: the mean is 2.66% with instruction-tuned specialists in panel a and 2.3% with base specialists in Table 6. At the default $\kappa = 1$, the floor binds in every layer and retention is approximately flat. At $\kappa = 0.5$, early layers with $\overline{\text{CR}}_l \approx 0.68$ keep 4–6% of coordinates, while middle layers with $\overline{\text{CR}}_l \approx 0.63$ keep about 2%. Retention p_l and coherence have Pearson $r = 0.91$ in panel b.

Across the settings in panel c, the largest advantage occurs on Math+Coder, whose sign agreement is 58% and disagreement 42% and where the compared baselines fall below the base. For fixed layer power and dimension, $\text{CR} \to 1$ implies $\sigma_l \to 0$ up to the numerical stabilizer, and a zero threshold gives Task Arithmetic with the same scaling.

Coherence mixes two sources of disagreement, which the split of Section 3.3 separates. Ignoring the stabilizer, mutual conflict ($\tau_1 = +10, \tau_2 = -10$) gives $\text{CR} = 0$, while a task-exclusive update ($\tau_1 = +10, \tau_2 = 0$) gives $\text{CR} = 0.5$. Shrinking the latter can remove useful specialist signal, and the

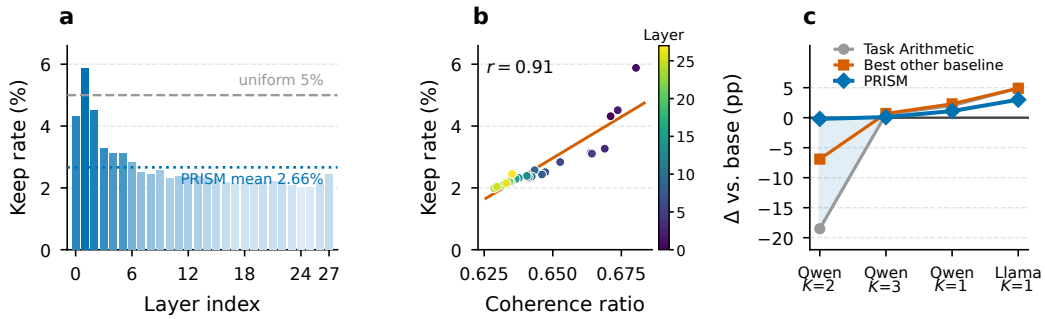

Figure 5: Mechanism on Qwen2.5-7B Math+Coder ($K=2$, instruction-tuned specialists, $\kappa=0.5$). **(a)** Layer-wise retention, bars shaded by coherence. **(b)** Retention against coherence, colored by depth. **(c)** Distance from the base model across settings ordered by interference.

ρ-gate of Section 4.2 returns the plain average when dispersion dominates. Interference-based scoring exceeds activation-norm scoring by 1.3pp without calibration data, as Appendix D.7 details.

D Extended Results

D.1 Calibration Granularity

Table 5 compares global, row-wise and layer-wise estimates of the threshold scale in Eq. 4, together with a sign-agreement substitute. Their six-task means lie within evaluation uncertainty, and their perplexities range from 9.0 to 9.1. Global and layer-wise estimates stay close at larger coefficients, with 67.8 against 67.9 at $c=0.7$ and 65.0 against 64.8 at $c=1.0$. The same holds on the six- and eight-specialist merges of Section 5.7, with 67.9 against 68.2 and 68.2 against 68.0. We retain the layer-wise default because it lets the threshold vary across depth.

Table 5: Calibration-granularity variants on Qwen2.5-7B Math+Coder ($K=2$, $c=0.5$). The variants use interference or sign agreement to calibrate a soft threshold.

Variant	Mean	PPL↓
Global-λ (single σ)	68.1	9.0
SA-calibrated σ (sign agreement)	67.8	9.0
Row-wise σ (empirical Bayes)	67.8	9.1
PRISM (layer-wise CR, default)	67.8	9.0

D.2 Threshold Scale and Scope

Table 6 sweeps the threshold scale κ and the merged scope on Qwen2.5-7B $K=2$. The 0.1% floor binds in every layer for $\kappa \geq 1$, so $\kappa=1$, 2 and 3 give identical merged weights and share one row.

With the floor disabled at $\kappa=1$, the calibrated threshold keeps 0.01% of coordinates and writes 1.4% of the average's norm. The six-task mean is 68.6 at perplexity 8.9, within evaluation uncertainty of the floored operator. On HumanEval, HumanEval+, MBPP and MBPP+, the EvalPlus scores fall from the floored 65.9/59.1/71.2/61.1 to 63.4/57.3/70.6/60.8, so the floor adds 0.3 to 2.5pp on each coding metric.

Table 6: Threshold-scale and scope ablations of PRISM on Qwen2.5-7B $K=2$. *floor*: the 0.1% keep-rate floor binds in every layer. **Bold** = best in each block.

Ablation	Variant	PPL↓	GSM	ARC	MMLU	TQA	HS	WG	Mean
Threshold scale κ	$\kappa=0.25$ (keep 22.6%)	11.6	81.2	**56.6**	72.0	53.9	66.6	74.2	67.4
	$\kappa=0.5$ (keep 2.3%)	9.5	79.0	53.2	74.1	54.2	**67.8**	75.4	67.3
	$\kappa\in[1,3]$ (floor, default)	**9.0**	**81.4**	51.4	**74.3**	**55.2**	67.4	**77.0**	**67.8**
Scope	MLP + Attention	**9.0**	**83.0**	51.2	74.2	**55.5**	**67.6**	76.0	**67.9**
	MLP-only (default)	**9.0**	81.4	**51.4**	**74.3**	55.2	67.4	**77.0**	67.8

D.3 Interference Distribution

We examine the distribution on which the threshold of Eq. 4 acts. The layer statistic $\sigma_l^2 = \overline{I}_l$ averages interference across $n_l \sim 10^8$ coordinates in the Qwen2.5-7B Math+Coder merge. Across its 28 MLP layers, the sampled standardized aggregate has excess kurtosis 0.44 and skewness near zero, with mild heavy-tail deviations beyond $|z| > 3$ on a Q-Q plot. Layer coherence correlates negatively with mean interference ($r = -0.41$).

Under the common-component, independent-disturbance model, the empirical variance with divisor K has expectation $(K-1)/K$ times the disturbance variance, including for $K \in \{2, 3\}$. A constant rescaling of σ_l can be offset algebraically by κ before the floor is applied. Using σ_l rather than $\sigma_l/\sqrt{K}$ changes the threshold by $\sqrt{K}$, at most about 2.8 over the studied range through $K = 8$. The κ sweep in Table 6 spans a factor of 12, wider than any of these convention changes, and the six-task mean stays between 67.3 and 67.8 across it.

At $K = 2$, an element-level $\sigma(i)$ rests on a single squared difference. Multiplying this unstable local scale by the universal factor $\sqrt{2 \log n} \approx 6.2$ risks excessive shrinkage. The row-wise estimate already matches the layer-wise six-task mean in Table 5.

D.4 Specialist Checkpoints and Evaluation Uncertainty

Table 7 compares instruction-tuned and base specialist checkpoints, with benchmark-level standard errors from binomial proportion variance at the shared task-specific sample sizes. Both specialist pairs are merged into Qwen2.5-7B with task vectors taken against it, and the instruction-tuned pair consists of Math-7B-Instruct and Coder-7B-Instruct. PRISM leads the listed merges with both pairs, and Task Arithmetic scores 6.5 points higher with instruction-tuned specialists.

Individual benchmarks carry standard errors of 1.7 to 2.2pp, with MMLU at 0.4pp. Propagating the six task-specific errors gives at most 0.8pp for a six-task mean and roughly 1.1pp for a difference between two means.

Table 7: Benchmark scores with standard errors for Qwen2.5-7B Math+Coder merges of instruction-tuned and base specialists; the base-specialist merges are those of Table 1. The final column propagates the six benchmark errors into the standard error of the mean.

Method	GSM8K	ARC	MMLU	TQA	HS	WG	Mean
Instruction-tuned specialists							
Task Arith.	47.8 ± 2.2	48.4 ± 2.2	59.5 ± 0.4	53.0 ± 2.0	57.8 ± 2.2	69.8 ± 2.1	56.0 ± 0.8
TIES	69.4 ± 2.1	51.4 ± 2.2	65.4 ± 0.4	52.8 ± 2.0	62.6 ± 2.2	70.0 ± 2.1	61.9 ± 0.8
DELLA	70.4 ± 2.0	52.2 ± 2.2	65.0 ± 0.4	51.7 ± 2.0	62.2 ± 2.2	70.6 ± 2.0	62.0 ± 0.8
LEWIS (wt.-sp.)	77.4 ± 1.9	54.6 ± 2.2	70.6 ± 0.4	50.7 ± 2.0	66.2 ± 2.1	72.0 ± 2.0	65.2 ± 0.8
PRISM	81.2 ± 1.7	51.4 ± 2.2	74.3 ± 0.4	55.0 ± 1.9	67.8 ± 2.1	75.8 ± 1.9	67.6 ± 0.7
Base-model specialists							
Task Arith.	22.0 ± 1.9	47.0 ± 2.2	55.7 ± 0.4	50.0 ± 2.0	55.4 ± 2.2	66.6 ± 2.1	49.5 ± 0.8
TIES	67.2 ± 2.1	53.4 ± 2.2	64.3 ± 0.4	50.2 ± 2.0	60.6 ± 2.2	71.0 ± 2.0	61.1 ± 0.8
DELLA	65.2 ± 2.1	51.6 ± 2.2	64.2 ± 0.4	50.3 ± 2.0	59.8 ± 2.2	70.4 ± 2.0	60.3 ± 0.8
LEWIS (act.)	43.0 ± 2.2	44.2 ± 2.2	69.1 ± 0.4	50.7 ± 2.0	62.8 ± 2.2	66.6 ± 2.1	56.1 ± 0.8
PRISM	81.4 ± 1.7	51.4 ± 2.2	74.3 ± 0.4	55.2 ± 1.9	67.4 ± 2.1	77.0 ± 1.9	67.8 ± 0.7
Base model	82.8 ± 1.7	51.4 ± 2.2	74.3 ± 0.4	55.6 ± 1.9	67.0 ± 2.1	77.0 ± 1.9	68.0 ± 0.7

D.5 PRISM-Denoise against PRISM-Prune

Table 8 compares PRISM-Denoise with PRISM-Prune, which averages the scaled task vectors and keeps the largest-magnitude coordinates of each layer at a rate of 5% scaled by the layer's coherence relative to the whole model. With two specialists, Denoise leads on the six-task mean by 3.7pp at 7B and 6.9pp at 1.5B, and the variants tie at $K = 3$, where interference is lower. At $K = 1$ Denoise equals Task Arithmetic because $\lambda_l = 0$, while Prune keeps 5% of the task vector and trails by 3.2pp.

D.6 Model Scale

Table 9 evaluates merges of the Qwen2.5 Instruct and Coder-Instruct models at 3B and 14B. The interference scores of both merges lie far below the threshold of Table 4, and PRISM gains 0.8 and 0.7pp over the base while Task Arithmetic stays within 0.1pp of it.

Table 8: PRISM-Denoise against PRISM-Prune across settings (%). **Bold** marks the better variant, with ties sharing the mark.

		Harness							Generation			
Setting	**Variant**	**GSM8K**	**ARC**	**MMLU**	**TruthQA**	**HellaSwag**	**WinoGr.**	**Mean**	**GSM8K**	**MATH**	**HumanEval**	**MBPP**
Qwen 7B $K=2$	Denoise	**81.4**	51.4	**74.3**	**55.2**	**67.4**	**77.0**	**67.8**	**82.4**	**30.2**	**51.2**	**62.8**
	Prune	73.8	**53.6**	70.2	50.8	65.2	71.0	64.1	73.8	15.6	31.1	39.6
Qwen 1.5B $K=2$	Denoise	**59.6**	**42.6**	**61.8**	45.5	**59.0**	**66.4**	**55.8**	**59.0**	**26.6**	**35.4**	**43.0**
	Prune	35.4	40.0	53.4	**46.2**	53.8	64.8	48.9	35.6	13.8	21.3	37.4
Qwen 7B $K=3$	Denoise	**83.4**	51.6	**74.3**	55.6	**66.8**	**76.8**	**68.1**	**83.6**	37.2	57.3	**64.2**
	Prune	83.0	**52.0**	**74.3**	**55.8**	66.4	76.6	68.0	82.6	**38.0**	**57.9**	63.6
Llama $K=1$	Denoise	**71.8**	53.6	**68.9**	**51.3**	70.6	**74.0**	**65.0**	**73.4**	**28.6**	**53.7**	**52.4**
	Prune	59.0	**54.0**	67.1	46.7	**71.2**	73.0	61.8	58.8	25.0	40.2	52.2

Table 9: Instruct+Coder-Instruct merges of Qwen2.5 at 3B and 14B ($K=2$, $c=0.5$, six-task mean, %). Both merges score far below the threshold of Table 4. **Bold** marks the best score in each row.

Model	$c\sqrt{\overline{I}/K}$ ($\times 10^{-3}$)	**Base**	**TA**	**PRISM**
Qwen2.5-3B	0.076	61.5	61.4	**62.3**
Qwen2.5-14B	0.067	72.2	72.1	**72.9**

D.7 LEWIS VARIANTS AND THE SCORING ABLATION

In Table 10, activation-based LEWIS trails the weight-sparsity variant by 8.8pp on the Qwen2.5-7B Math+Coder merge of Table 1. The gap narrows to 0.8pp on the Llama-3.1-8B single-specialist merge and closes on the Qwen2.5-7B three-specialist merge, where interference is lower. Table 11 compares activation-norm and interference-based threshold scales on the same Math+Coder merge while retaining soft thresholding. Denoising with the activation-based scale scores 10.4pp above activation-based pruning, and PRISM's interference-based scale adds a further 1.3pp without calibration data.

Table 10: Activation-based and weight-sparsity LEWIS on three merges (six-task mean, %). **Bold** marks the better variant in each merge.

Setting	**Method**	**Mean (%)**
Qwen $K=2$	LEWIS (act.)	56.1
	LEWIS (wt.-sp.)	**64.9**
	Base Model	68.0
Llama $K=1$	LEWIS (act.)	60.6
	LEWIS (wt.-sp.)	**61.4**
	Base Model	60.0
Qwen $K=3$	LEWIS (act.)	**68.1**
	LEWIS (wt.-sp.)	**68.1**
	Base Model	68.0

D.8 DATA-DEPENDENT BASELINES

Table 12 lists the data-dependent baselines on Qwen $K=2$. AdaMerging learns layer-wise coefficients by minimizing prediction entropy on 128 unlabeled C4 texts with 100 steps of simultaneous perturbation stochastic approximation (SPSA). It reaches parity with PRISM only at its tuned coefficient and falls to 51.6 at the shared $c=0.5$. Greedy Soup, which adds a specialist only when it lowers perplexity on 50 WikiText-2 texts, rejects both specialists and returns the base model. Fisher-weighted averaging, which estimates a diagonal Fisher on 100 WikiText-2 texts, falls to 47.5.

D.9 BASELINE FAIRNESS PROTOCOL

Table 13 lists the settings and sweeps used for every baseline across our experiments. PRISM uses $\kappa=1.0$ without setting-specific tuning. The main task-vector comparisons fix $c=1/K$ after averaging, and Section 5.3 reports the coefficient sweep. The GSM8K and MATH-500 generation evaluations fix the TIES trim at $k=0.2$ and use weight-sparsity LEWIS, which needs no calibration data, whereas activation-based LEWIS computes its layer importance from 128 WikiText-2 texts.

Table 11: Scoring ablation on Qwen2.5-7B Math+Coder base specialists ($K=2$, six-task mean).

Scoring Method	Mean (%)	PPL↓
PRISM-Denoise (interference σ)	**67.8**	**9.0**
LEWIS-CTP (activation σ + denoise)	66.5	11.9
LEWIS-Act (activation + prune)	56.1	24.2

Table 12: Data-dependent baselines on Qwen2.5-7B Math+Coder ($K=2$, six-task mean). PRISM uses no data.

Method	Data used	Mean (%)
Base model, unmerged	none	68.1
AdaMerging (tuned, $c=0.1$)	128 C4 texts	67.9
AdaMerging ($c=0.5$)	128 C4 texts	51.6
Greedy Soup	50 WikiText-2 texts	68.1
Fisher-weighted averaging	100 WikiText-2 texts	47.5
PRISM ($c=0.5$)	none	67.8

D.10 WHY BASELINES FAIL ON QWEN $K=2$

The Math+Coder task vectors are large and conflict in sign on about 42% of elements. This weakens sign consensus for TIES and DELLA and motivates examining the size of the update written by each method. At $c=0.5$, DARE collapses at both tested drop rates: PPL 7.9×10^5 with a six-task mean of 30.0 at $p=0.9$, and PPL 128 with 39.2 at $p=0.5$. Its $1/(1-p)$ rescaling enlarges the surviving coordinates, and at $p=0.9$ its update reaches $2.85\times$ the norm of the plain average in Table 3.

Iso-C and WUDI-Merging, the two 2025 data-free methods, were evaluated over the grids in Table 13. Iso-C at its default coefficient $c=1/K=0.5$ collapses with PPL 3.5×10^4 and a six-task mean of 29.3. At $c=0.25$ it gives PPL 36.4 and a mean of 44.3; at $c=0.1$ it reaches PPL 11.6 and a mean of 66.6, 1.2pp below untuned PRISM. Its scores are $78.4/54.6/71.4/55.0/64.8/75.2$ on GSM8K, ARC, MMLU, TruthfulQA, HellaSwag and WinoGrande. WUDI-Merging diverged in all four evaluated configurations, with six-task means of 29.6 to 30.3. MLP scope at learning rate 10^{-5} gave PPL 2.1×10^7, and all linear layers at 10^{-5} gave 1.6×10^6. On the instruction-tuned versions of these specialists, all linear layers gave 1.3×10^6 at 10^{-5} and 2.2×10^6 at 10^{-6}.

Community recipes are our evaluations on the same specialist pair, using each recipe's own rule without our external coefficient or norm scaling. SLERP at $t=0.5$ (Shoemake, 1985), as used in the mergekit recipe (Goddard et al., 2024), is numerically close to the full unscaled average because the shared base dominates each tensor. It writes $1.27\times$ the norm-scaled mean and collapses at PPL 1.0×10^6. Model Stock (Jang et al., 2024) derives $t\approx0.41$ from the observed mean task-vector cosine of 0.26, giving $2.4\times$ shrinkage, and still loses 14.8pp at perplexity 35. PRISM, by comparison, writes 2.2% of the average's norm in Table 3. Model Stock's angle-derived rule does not directly calibrate the absolute update magnitude, consistent with the limitations of direction statistics measured in Section 5.7.

D.11 SCALING-COEFFICIENT SWEEP

Table 14 gives the numbers behind Figure 2: the six-task harness mean and the four generation benchmarks at every coefficient. The unmerged base has a harness mean of 68.1 and scores 83.8, 38.2, 56.7 and 63.8 on GSM8K, MATH-500, HumanEval and MBPP. At $c=0.7$, baseline perplexities are 961, 100, 116 and 105 for Task Arithmetic, TIES, DELLA and LEWIS, and at $c=1.0$ they are 4.6×10^5, 1.1×10^4, 1.6×10^4 and 5.9×10^3.

D.12 LOW-RANK SPECIALISTS

LoRA represents an update as a low-rank product (Hu et al., 2022). A trained adapter can be materialized as $\tau_k = (\alpha/r)\mathbf{B}_k\mathbf{A}_k$ before merging in full weight space. The experiment in Table 15 instead simulates low-rank updates by retaining the top-r singular components of each full Math+Coder task vector. Across ranks from 8 to 128, PRISM stays within 0.6pp of its full-weight six-task mean of 67.8 and outperforms Task Arithmetic at every rank on all three measures.

Table 13: Baseline settings and sweeps across all experiments. The LEWIS data entry applies to the activation-based variant.

Method	Hyperparameters	Data used
Task Arithmetic	$c \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$	none
TIES	c (as above); trim $k \in \{0.1, 0.2\}$	none
DELLA	c; $\epsilon = 0.3$; trim $k = 0.2$	none
LEWIS	c; keep rate $\in \{0.05, 0.1\}$	128 WikiText-2 texts
Fisher	c; diagonal Fisher	100 WikiText-2 texts
Greedy Soup	inclusion by perplexity	50 WikiText-2 texts
AdaMerging	$c \in \{0.1, 0.2, 0.3, 0.5\}$; 100 SPSA steps; layer-wise coeffs	128 C4 texts
Iso-C	$\alpha \in \{0.10, 0.25, 0.50\}$	none
DARE	drop rate $p \in \{0.5, 0.9\}$ at $c = 0.5$; $c = 1.0$ at $p = 0.9$	none
WUDI-Merging	scope $\in$ {MLP, all linear}; lr $\in \{10^{-5}, 10^{-6}\}$; 300 iters	none
PRISM (ours)	none (fixed $\kappa = 1.0$)	none

Table 14: Scaling-coefficient sweep on Qwen2.5-7B Math+Coder: six-task harness mean and generation accuracy (%). LEWIS: activation-based harness mean, weight-sparsity generation. **Bold** = best merged in each block.

		Harness	Generation			
c	Method	Mean	GSM8K	MATH-500	HumanEval	MBPP
0.3	Task Arith.	63.4	69.2	29.0	33.5	48.8
	TIES	66.1	78.4	30.2	42.1	53.8
	DELLA	65.8	78.2	27.6	42.7	53.4
	LEWIS	67.6	80.8	22.6	41.5	55.0
	PRISM-Denoise	**68.0**	**83.0**	**32.6**	**51.2**	**64.8**
0.5	Task Arith.	49.5	23.0	5.8	13.4	27.4
	TIES	61.1	65.8	24.6	17.7	43.0
	DELLA	60.3	64.0	22.0	17.7	42.4
	LEWIS	56.1	74.2	15.4	31.7	41.4
	PRISM-Denoise	**67.8**	**82.4**	**30.2**	**51.2**	**62.8**
0.7	Task Arith.	35.1	0.2	0.0	1.2	1.0
	TIES	49.3	32.2	10.0	6.7	13.6
	DELLA	48.3	33.2	9.6	11.6	11.0
	LEWIS	49.4	62.8	6.4	23.2	25.8
	PRISM-Denoise	**67.9**	**81.8**	**27.0**	**47.0**	**61.2**
1.0	Task Arith.	28.8	0.0	0.0	0.0	0.0
	TIES	30.4	0.0	0.0	0.0	0.0
	DELLA	30.9	0.0	0.0	0.0	0.0
	LEWIS	38.7	17.4	0.4	1.8	1.2
	PRISM-Denoise	**64.8**	**65.8**	**12.6**	**39.0**	**57.2**

E Coherence Ratio and Sign Agreement

For $K = 2$ with equal scaling and nonzero task-vector power, neglecting the stabilizer gives $\mathrm{CR}(i) = (\tau_1(i) + \tau_2(i))^2 / \big(2(\tau_1(i)^2 + \tau_2(i)^2)\big)$. When both magnitudes are equal and nonzero, this equals binary sign agreement; unequal magnitudes give intermediate values. At the layer level, coherence and sign agreement are highly correlated ($r = 0.998$ on Qwen2.5-7B $K = 2$), an empirical finding rather than a consequence of averaging alone over $n_l \sim 10^8$ coordinates. Coherence relates the observed interference to power through $\sigma_l^2 = \overline{I}_l = \overline{P}_l - \overline{S}_l \simeq \overline{P}_l(1 - \overline{\mathrm{CR}}_l)$, where the final expression neglects the stabilizer. This supplies the chosen threshold scale; a sign-agreement fraction alone does not determine the magnitude of interference.

F Conflict and Exclusivity

Section 5.6 notes that low coherence mixes two phenomena. We make this operational with an element-wise exclusivity score,

$$e(i) = \frac{\max_k |s_k \tau_k(i)|}{\sum_k |s_k \tau_k(i)| + \epsilon}, \tag{6}$$

For nonzero coordinates and negligible ϵ, $e(i)$ lies in $[1/K, 1]$: values near 1 indicate concentration in one specialist, while values near $1/K$ indicate similar magnitudes across specialists. Equal

Table 15: Low-rank SVD simulation on Qwen2.5-7B Math+Coder ($K=2$) across projection ranks: six-task harness mean, GSM8K generation accuracy and perplexity for PRISM-Denoise and Task Arithmetic (TA), with the merge's coherence ratio.

		Mean		GSM8K (gen.)		PPL↓	
Rank	CR	PRISM	TA	PRISM	TA	PRISM	TA
$r=8$	0.529	**67.7**	65.1	**81.8**	78.2	**10.1**	14.2
$r=16$	0.524	**67.9**	65.2	**83.8**	78.4	**9.9**	14.5
$r=32$	0.520	**68.4**	64.8	**84.2**	77.4	**9.8**	14.9
$r=64$	0.519	**68.0**	64.2	**83.4**	73.6	**9.5**	16.0
$r=128$	0.521	**68.3**	63.2	**83.6**	72.6	**9.2**	17.3
Full fine-tuning	0.630	**67.8**	49.5	**82.4**	23.0	**9.0**	46.7

magnitudes imply symmetric conflict only within the low-CR subset. When every task-vector entry is zero, $e(i) = 0$. Partitioning the low-CR mass of Qwen2.5-7B Math+Coder by this score shows that thresholded elements are predominantly conflict elements, while many exclusive elements retain a nonzero consensus and survive.

Why the threshold uses the total interference. Calibrating with cancellation alone, $\sigma_l^2 \leftarrow \overline{C}_l$, gives 67.5 at $c=0.5$, 67.0 at $c=0.7$ and 60.8 at $c=1.0$ on Qwen2.5-7B $K=2$. At the last coefficient, PRISM scores 64.8. Since $\overline{C}_l \leq \overline{I}_l$, this substitution lowers the nominal threshold and can retain a larger update. The observed loss at aggressive scaling supports retaining $\sigma_l^2 = \overline{I}_l$ for the threshold and using ρ as a binary gate.

Norm scaling with heterogeneous specialists. The min-norm rule equalizes scaled task-vector norms at the smallest input norm. With strongly heterogeneous specialists, this shrinks the whole update: on the Llama-3.1-8B $K=3$ merge, the rule keeps every method within 0.5pp of the base's 60.2 in Table 4. Table 19 therefore uses unscaled task vectors. The Qwen2.5 Math+Coder specialists have comparable norms, with $s_k \in [0.84, 1.0]$.

G Implementation Details

The default matched comparison applies PRISM and its baselines to the MLP projections (`gate_proj`, `up_proj`, `down_proj`). Attention projections (`q/k/v/o_proj`), embeddings, layer norms and the output head retain their base values. Scope ablations and extended recipes state their exceptions. MLP layers can store domain knowledge as key–value memories (Geva et al., 2022; Dai et al., 2022), which motivates this scope. MLP-only merging stays within 0.1pp of the attention-inclusive variant's six-task mean in Table 6 with 2.3× fewer merged keys, 84 against 196. The norm-scaling factors on Qwen2.5-7B Math+Coder are 0.84 and 1.0.

We load models one at a time, with specialist weights and intermediates on CPU; the peak GPU footprint is a single 7B model, about 15 GB in bfloat16.

H Coding Generation

Table 16 reports coding results with the instruction-tuned specialists of Section 5.2. PRISM-Denoise matches the base model's 63.4 on MBPP and scores highest on both benchmarks among the merges that use no calibration data.

I Coding Evaluation with EvalPlus

We evaluate the same model checkpoints with the EvalPlus framework (Liu et al., 2023), adapting the QwenCoder-2.5 evaluation scaffold in the QwenLM/Qwen3-Coder repository.[1] Generation uses vLLM with a fixed ChatML prompt whose assistant turn is pre-seeded with ` ```python `, greedy decoding, 2048 new tokens and `max_model_len`=4096. Scoring applies `evalplus.sanitize` and `evalplus.evaluate` to the base and augmented (plus) splits. The recipe differs from the `lm-eval-harness` evaluation in Section 5.2 in its prompt, code extraction and test suites. Tables 3 and 17 report the merged models and specialist references under this recipe.

[1] QwenLM, `eval_plus/` scaffold, commit `33bc6aa`.

Table 16: Coding results with instruction-tuned Qwen2.5-7B Math+Coder specialists ($K = 2$), pass@1 (%). *Gray italic* marks specialist reference scores. LEWIS uses its weight-sparsity variant. **Bold** best merged, underline second best; †needs calibration data.

Method	MBPP	HumanEval
Base Model	63.4	57.3
Math Specialist	*0.0*	*48.8*
Coder Specialist	*70.0*	*81.7*
Task Arith.	34.8	11.6
TIES	45.1	18.3
DELLA	45.7	18.9
LEWIS	45.7	35.4
AdaMerging† ($c=0.1$)	62.8	**53.0**
PRISM-Prune	45.1	31.7
PRISM-Denoise	**63.4**	48.8

Under this recipe, Task Arithmetic, TIES, DELLA, DARE, LEWIS, SLERP and Model Stock produce degenerate generations, such as runs of zeros, multilingual token noise, empty code blocks or repeated prompt text, and none of them solves more than one problem on any split.

Table 17: EVALPLUS pass@1 (%) of the unmerged base, PRISM and the two specialists on Qwen2.5-7B Math+Coder ($K=2$, base specialists). Table 3 reports the baselines under the same recipe.

Model	HumanEval	HumanEval+	MBPP	MBPP+
Base model, unmerged	62.8	57.9	69.8	59.5
PRISM-Denoise	**65.9**	**59.1**	**71.2**	**61.1**
Math Specialist	*66.5*	*63.4*	*60.3*	*52.9*
Coder Specialist	*64.6*	*60.4*	*72.0*	*60.6*

J BREADTH EXPERIMENTS: DETAILS

Configurations and threshold. Table 4 reports twenty-two evaluated configurations, seven destructive and fifteen harmless. We placed the threshold of 1.9×10^{-3} in the gap between the harmless and destructive scores of the nine unstarred configurations, which were measured first. Twelve of the fourteen predictions made before evaluation were correct.

Three settings extend the Qwen2.5-7B Math+Coder merge along orthogonal axes: a different base family (Llama-3.1-8B, $K=3$) in Table 19, an off-the-shelf merge of six community specialists with conflicting objectives (Qwen2.5-7B $K=6$: Math, Coder, Finance, Thai, Safety, Roleplay), and a matched-recipe self-SFT scaling curve to $K=8$ in Table 20.

Screened public pairs. The screen scored twenty-eight public pairs without a code specialist, fourteen on Qwen2.5-7B, eleven on Llama-3.1-8B and three on Mistral-7B. Their specialists span mathematics and reasoning, medicine, finance, law, long-context modeling, general post-training, safety classification, role-play, and continued pretraining in Japanese, Thai, Southeast Asian and European languages. The scores range from 0.002×10^{-3} to 0.853×10^{-3}, all below half the threshold, and the largest belongs to Qwen2.5-7B R1-Distill+Math-PRM.

Continued-pretraining recipe. The titration uses full-parameter causal-language-model training from `Qwen/Qwen2.5-1.5B`. The Japanese run uses the `train` split of `HuggingFaceFW/fineweb-2`, configuration `jpn_Jpan`; the biomedical run uses titles and abstract content from `MedRAG/pubmed`, also `train`. We shuffle the streams with a 10,000-document buffer, insert EOS between documents, and pack 4096-token sequences into an effective batch of 64 sequences. Both phases use fused AdamW ($\beta_1=0.9$, $\beta_2=0.999$, $\epsilon=10^{-8}$), weight decay 0.01, gradient-norm clipping at 1.0, bfloat16 computation, and 100 warmup steps followed by a constant learning rate. At 10^{-4}, the measured steps are $500, 1000, \ldots, 3500, 3815$, totaling 1.00 billion tokens for each specialist. The higher-rate phase loads each step-3815 checkpoint's weights, restarts the optimizer and step count, and uses 5×10^{-4}. Before shuffling, it skips the first one million Japanese or four million PubMed documents. Its measured steps are $500, 1000, \ldots, 4000$; step 4000 adds 1.05 billion tokens to each specialist. At every measured step, the Japanese and biomedical checkpoints are scored as one merge with $K=2$, $c=0.5$ and min-norm scaling. The own-domain

perplexities in Section 5.7 use the FineWeb-2 Japanese `test` split and a PubMed `train` slice after nine million documents.

The first titration endpoint. At one billion tokens and learning rate 10^{-4}, the merge is harmless at a score of 0.699×10^{-3}, with six-task means of 55.4 for Task Arithmetic, 56.2 for PRISM and 56.1 for the base.

Rival predictors. Table 18 compares the magnitude statistics $c\sqrt{x/K}$ for $x \in \{\overline{I}, \overline{P}, \overline{S}\}$, using parameter-count-weighted layer means, with two direction statistics, the mean incoherence $1 - \overline{\mathrm{CR}}$ and the cancellation ratio ρ. Inverting the definition of $\overline{\mathrm{CR}}_l$ gives total power as $\overline{P}_l = (\overline{I}_l + \epsilon\,\overline{\mathrm{CR}}_l)/(1 - \overline{\mathrm{CR}}_l)$, and signal power is $\overline{S} = \overline{P} - \overline{I}$. All three magnitude statistics reach AUC 0.98 or higher, and total and signal power separate all twenty-two outcomes. Both direction statistics place destructive merges below many harmless ones.

Table 18: Rival pre-merge statistics on all configurations of Table 4. AUC treats larger scores as more destructive. Margin is the smallest destructive value divided by the largest harmless one.

Statistic	AUC	Margin	Result
$c\sqrt{\overline{I}/K}$ (used)	0.98	0.83×	predictive
$c\sqrt{\overline{P}/K}$ (total power)	1.00	21.2×	separates
$c\sqrt{\overline{S}/K}$ (signal power)	1.00	23.3×	separates
$1 - \overline{\mathrm{CR}}$ (mean incoherence)	0.08	$<0.01\times$	anti-predictive
ρ (cancellation ratio)	0.25	0.02×	anti-predictive

Table 19: Llama-3.1-8B $K = 3$ merge of OpenMath2, Hermes-3 and Llama-Guard with unscaled task vectors (six-task mean, %). The data-free ρ-gate detects the dispersion-dominated regime at $\rho = 0.021$ and returns Task Arithmetic. **Bold** = best.

Method	Mean (%)
Base Model	60.2
Task Arithmetic	**62.0**
DELLA	61.4
TIES	61.2
LEWIS	60.3
PRISM-Denoise (no gate)	59.6
PRISM-Denoise (+ ρ-gate)	**62.0**

$K = 6$ and self-SFT scaling. On the six-specialist merge, the base, Task Arithmetic and PRISM score 68.1, 68.0 and 68.2, consistent with an interference score of 0.005×10^{-3}.

The matched-recipe curve in Table 20 probes expert count under a common training recipe: as K grows from 2 to 8, ρ rises monotonically ($0.34 \to 0.70$), while every method remains within a point of the base. Neither increasing K nor increasing ρ indicates harm within this curve.

Table 20: Matched-recipe self-SFT scaling on Qwen2.5-7B with $c = 1/K$ and unscaled task vectors, merging $K = 2$ to 8 specialists trained from one base under an identical recipe.

Method	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$
Base Model (ref.)	68.1	68.1	68.1	68.1	68.1	68.1	68.1
Task Arithmetic	68.7	68.5	68.0	68.1	68.1	68.1	68.0
TIES	68.4	68.2	68.2	68.1	67.9	68.1	68.2
PRISM-Denoise (+ ρ-gate)	67.9	68.1	68.1	68.1	68.0	68.0	68.1
ρ (cancellation ratio)	0.34	0.42	0.46	0.55	0.60	0.65	0.70