PREDICTING WHEN MODEL MERGING BEATS REVERTING TO THE INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging is typically evaluated by whether the merged model outperforms the two fine-tuned models it combines. We show that this comparison is misleading whenever both fine-tuned models score below their shared initialization on the reported metric, as is common after self-training, continual fine-tuning, or fine-tuning for accuracy. Interpolating a single one of them back toward the initialization, by the same distance the merge moves, can then match or exceed the merge. We propose this interpolation, the matched regression-to-ancestor control, as a baseline that merging evaluations should report. We also derive a criterion that predicts which of the two will score higher from the initialization and the two fine-tuned models alone, before either candidate is evaluated. Both candidates lie in the plane spanned by the two task vectors, where a first-order approximation, which we test and whose error we bound, reduces the comparison to a single inequality with no fitted parameters. On a synthetic self-training testbed with exactly known structure, the criterion correctly predicts the better candidate on 28 of 28 lineage pairs. We further fine-tune three experts from each of six base models spanning three families and evaluate every pair on four benchmarks, for 72 cells in total. The control outperforms the merge in 97% of the cells in which the initialization outscores both experts, compared with 44% of the remaining cells. On third-party fine-tunes of Llama-3.1-8B, the control recovers a median 69% of the gain of a merge that outperforms both of its parents. The criterion does not uniformly favor the control: on the standard 8-task CLIP benchmark, which we reproduce, the best expert outscores the initialization, the criterion accordingly predicts no advantage for the control, and the merge outperforms it by 12.5 points. The control requires a single interpolation, and the criterion selects the better of the two candidates in 86% of cells, compared with 32% for the common practice of always selecting the merge. Code is available at `https://anonymous.4open.science/r/merge-plane-geometry-4591`.

1 INTRODUCTION

Weight-space merging has become a standard technique. Averaging fine-tuned checkpoints improves robustness and combines skills (Izmailov et al., 2018; Wortsman et al., 2022a), task arithmetic edits capabilities by adding and subtracting task vectors (Ilharco et al., 2023), and a family of operators reduces interference between them (Yadav et al., 2023; Yu et al., 2024; Matena & Raffel, 2022). The standard evidence of a successful merge is that it outperforms its parents.

This evidence is valid when each parent outperforms the shared initialization on the reported metric. It is not valid when both parents underperform the initialization on that metric, a situation that is common: recursive self-training on synthetic corpora degrades the tails of the generated distribution (Shumailov et al., 2024; Alemohammad et al., 2024), narrow fine-tuning degrades capabilities it does not target (Kirkpatrick et al., 2017; Luo et al., 2025), and fine-tuning for accuracy degrades zero-shot robustness, which is precisely why interpolating back toward the zero-shot model is itself a method (Wortsman et al., 2022b). In each of these settings, moving a single parent back toward the shared initialization improves the reported metric by itself. A merge that outperforms its parents has therefore not been shown to achieve anything that a simple interpolation could not.

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.

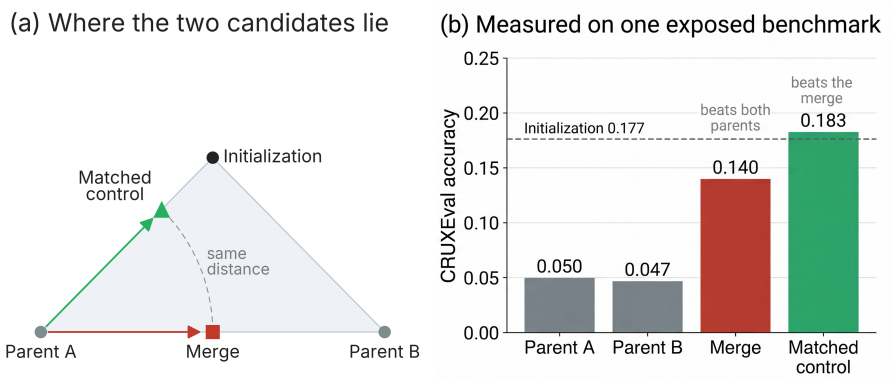


Figure 1: The regression-to-ancestor confound. (a) Schematic of the plane spanned by the two task vectors, drawn with equal task-vector norms at the case study's cosine of 0.0112. The uniform merge and the matched control leave parent A by the same distance, the merge toward the other parent and the control toward the shared initialization. (b) CRUXEval accuracy in the case study of Section 7 (Qwen2.5-1.5B-Instruct, two narrow fine-tunes). The merge outperforms both parents, which the standard evaluation counts as success, yet interpolating parent A toward the initialization by the same distance scores higher still. The benchmark is exposed, since the initialization outscores both parents; across all 33 exposed cells of Section 7 the control outperforms the merge in 97% of them.

We term this the regression-to-the-ancestor confound (Figure 1). The control that removes it is straightforward: interpolate one parent toward the shared ancestor by the same weight-space distance that the merge travels from that parent, and require the merge to outperform it. This paper addresses why the comparison resolves as it does, when it reverses, and how far the confound extends.

The plane. The entire comparison is two-dimensional (Figure 1a). With parents $\theta_a = \theta_0 + \tau_a$ and $\theta_b = \theta_0 + \tau_b$, every linear merge, every task arithmetic point, every interpolation toward the ancestor and the matched control lie in $\mathcal{P} = \theta_0 + \mathrm{span}\{\tau_a, \tau_b\}$, where a first-order model makes each behaviour's margin the barycentric combination of its margins at the three anchors. The uniform merge then carries ancestor weight zero and the matched control carries sibling weight zero: two edges of one triangle leaving the same parent. Matched distance fixes the exchange rate at $\alpha^\star = \|\tau_b - \tau_a\|/2\|\tau_a\|$, at equal norms $\sqrt{(1-\rho)/2}$, the exact conjugate of the averaging attenuation $\sqrt{(1+\rho)/2}$ used in the merging literature, so one parameter governs both sides of the comparison in opposite directions.

Contributions.

1. **A confound, made exact, and shown to be structural.** We formalize the matched regression-to-ancestor control and derive the master inequality that decides the comparison behaviour by behaviour, with its threshold $\rho \geq 1 - \beta^2/2$ in the sibling's retention β. This threshold explains why the confound is common rather than incidental: every one of the 52 task-vector cosines we measure, on a ground-truthed backbone, on fine-tuned experts and on released community checkpoints, lies below it.
2. **Why matching on *distance* is appropriate, and when the predictions can be trusted.** Because both competitors leave the shared parent by the same distance, the comparison is blind to the isotropic part of the curvature and only anisotropy can move it. This property keeps the theory usable despite an imperfect model of behaviour: a predicted sign is *certified* once the predicted gap exceeds what the anisotropy alone can shift, so errors in the fitted accuracy levels cannot overturn it. The resulting rule selects the better of the two candidates *before either is evaluated* in 0.86 of cells.
3. **Which merging evaluations are exposed to the confound.** A grid of 6 base models across three families, together with community checkpoints of Llama-3.1-8B and Qwen2.5-7B trained by third parties: the sign of the ancestral surplus separates the cells in which the control prevails from those in which it does not, the parameter-free prediction agrees in

sign on 0.82 of 136 comparisons, and the absorbed fraction the merging literature reports is itself a predicted quantity. The criterion does not uniformly favor the control: on the 8-task CLIP benchmark, which we reproduce, we recover the aggregate score that its normalisation conceals and find the setting *unexposed*, with the merge outperforming the control by +0.125.

Setting. We consider a practitioner who holds a strong ancestor checkpoint, without its training data, and two descendants that have drifted from it. The ancestor is then itself a legitimate candidate, which makes the matched control the appropriate baseline; Section 7 studies when it is not.

2 Related Work

Merging and interpolation. Weight averaging (Izmailov et al., 2018; Wortsman et al., 2022a), task arithmetic (Ilharco et al., 2023), interference-reducing operators (Yadav et al., 2023; Yu et al., 2024; Matena & Raffel, 2022) and search over merge recipes (Akiba et al., 2025) now form a large literature (Yang et al., 2026), almost all of it assuming the endpoints share a basin (Frankle et al., 2020; Ainsworth et al., 2023); Stoica et al. (2024) is the exception, and merges models that do not. Interpolation toward the initialization is already a method for robustness (Wortsman et al., 2022b), and our control performs exactly this interpolation. Our contribution is to show that it is also the baseline a sibling merge must exceed, and that it is not reported where the parents are degraded. Yuan et al. (2025) interpolate a self-improved model back toward its base inside the loop, using our control as a remedy rather than as a baseline; with no sibling, the comparison studied here does not arise.

Collapse and self-consuming training. Training on generated data is studied because public human text is finite: Villalobos et al. (2024) project its exhaustion within the decade. Recursive training on generated data narrows the learned distribution and destroys tails (Shumailov et al., 2024; Alemohammad et al., 2024; Briesch et al., 2023), with analytic characterizations (Dohmatob et al., 2024b;a; Seddik et al., 2024; Dohmatob et al., 2025), stability results (Bertrand et al., 2024), and two mitigations, accumulating rather than replacing (Gerstgrasser et al., 2024) or correcting (Gillman et al., 2024). That literature addresses how to prevent collapse; we use collapse as the cleanest instance of a regime in which both parents underperform their ancestor.

Self-improvement and forgetting. Training on filtered self-generations improves models (Zelikman et al., 2022; Yuan et al., 2024; Guo et al., 2025; Cobbe et al., 2021), though whether it adds capability or re-weights it is contested (Yue et al., 2025). Catastrophic forgetting (Kirkpatrick et al., 2017; Luo et al., 2025) and the low intrinsic dimension of fine-tuning updates (Aghajanyan et al., 2021; Hu et al., 2022) make the confound widespread.

3 The Merge Plane

The results below are exact rather than asymptotic, and they reduce the comparison to a single inequality whose two sides can be measured separately before either candidate is evaluated. Sections 5 to 7 test these results rather than merely illustrate them.

Setup. A shared ancestor θ_0 and two descendants $\theta_a = \theta_0 + \tau_a$, $\theta_b = \theta_0 + \tau_b$, with task-vector cosine $\rho = \langle \tau_a, \tau_b \rangle / (\|\tau_a\| \|\tau_b\|)$. We track a finite set of identifiable behaviours, which we call modes; on the synthetic backbone a mode is an exactly classifiable structural class, on the language-model testbeds a rare fact or a rare arithmetic template. Write $\mathcal{P} = \theta_0 + \mathrm{span}\{\tau_a, \tau_b\}$ with coordinates $\theta(u, v) = \theta_0 + u\tau_a + v\tau_b$, so $\theta_0 = (0, 0)$, $\theta_a = (1, 0)$, $\theta_b = (0, 1)$.

Proposition 1 (The plane contains the whole comparison)**.** *Uniform averaging, task arithmetic with any coefficients, spherical interpolation between the two endpoints, interpolation toward the ancestor, and the matched regression-to-ancestor control all lie in $\mathcal{P}$. Consequently the weight-space comparison is a comparison of directions in a two-dimensional plane at matched distance from a common parent, and its outcome does not depend on the ambient dimension.*

TIES, DARE and Fisher-weighted merging are coordinate-wise nonlinear and therefore leave $\mathcal{P}$; Section 6 quantifies this departure, which marks the boundary of Proposition 1.

Assumption 1 (First-order mode margins)**.** For each mode i there is a margin $s_i(\theta)$, negative when the model does not produce the mode and non-negative when it does, that is affine on $\mathcal{P}$: $s_i(\theta_0 + \Delta) = s_i^0 + \langle g_i, \Delta \rangle$. We instantiate $s_i(\theta) = \log \hat{p}_\theta(i) - \log \tau$, the log produced frequency of mode i minus the log detection threshold, base ten throughout, so a margin is measured in decades.

Proposition 2 (Barycentric margins)**.** *Under Assumption 1, for every* (u, v)*,*

$$s_i(u, v) = (1 - u - v)\, s_i^0 + u\, s_i^a + v\, s_i^b. \tag{1}$$

The uniform merge has $s_i^{\text{merge}} = \frac{1}{2}(s_i^a + s_i^b)$*, carrying ancestor weight zero, and the ancestor interpolation* $\theta_{\text{prior}}(\alpha) = (1 - \alpha, 0)$ *has* $s_i^{\text{prior}} = (1 - \alpha)s_i^a + \alpha s_i^0$*, carrying sibling weight zero.*

Assumption 1 is the only approximation in the argument, and its error is one-signed.

Proposition 3 (The first-order model errs in one direction, by a computable amount)**.** *Let* $\tilde{s}_i(u, v)$ *be the margin of the* mixture *of the three anchors, which averages frequencies where* (1) *averages log frequencies, and write* $w = (1 - u - v, u, v)$*. For every* (u, v) *with* $w \geq 0$*,*

$$0 \;\leq\; \tilde{s}_i(u, v) - s_i(u, v) \;\leq\; D_i(u, v) := \max_{k:\, w_k > 0} s_i^k - \sum_k w_k s_i^k, \tag{2}$$

and the gap is at least $D_i(u, v) - \log(1/w_{\max})$*, where* $w_{\max}$ *is the weight of the anchor attaining the maximum. Both bounds are computed from the three anchor measurements alone.*

Consequently (1) never overstates a frequency or a producible-mode count. Because D_i is the disagreement among the anchors, the error vanishes where they agree and is largest on precisely the modes that one parent has lost and the other retains, which are the modes on which the confound operates. The bracket is narrow, exactly $\log 2$ wide at the uniform merge, and Corollary 9 converts it into an a priori envelope on the thresholded count; all proofs are in Appendix A.

The exchange rate. The control is defined by matching the distance it travels from the parent from which it is constructed. Since $\|\theta_{\text{merge}} - \theta_a\| = \frac{1}{2}\|\tau_b - \tau_a\|$ and $\|\theta_{\text{prior}}(\alpha) - \theta_a\| = \alpha\|\tau_a\|$,

$$\alpha^\star = \frac{\|\tau_b - \tau_a\|}{2\|\tau_a\|}, \qquad \text{and at equal norms} \qquad \alpha^\star = \sqrt{\tfrac{1-\rho}{2}}. \tag{3}$$

Corollary 4 (Conjugate geometry)**.** *With* $A(\rho) = \sqrt{(1 + \rho)/2}$ *the factor by which averaging attenuates two equal-norm task vectors,* $\alpha^\star(\rho)^2 + A(\rho)^2 = 1$*. The single parameter* ρ *sets both sides of the comparison in opposite directions: as the drifts align, averaging becomes lossless and the matched control is simultaneously confined near the parent; as they become orthogonal, averaging attenuates and the control is permitted to travel far toward the ancestor.*

This matters because the merging literature treats $A(\rho)$ alone as the mechanism, whereas it accounts for only half of it: the deficit shrinks at high ρ mainly because the control is constrained to remain near the parent, not because averaging attenuates less.

Proposition 5 (Master inequality)**.** *Under Assumption 1, the uniform merge has at least the margin of the matched control built from parent* a *on mode* i*, that is* $s_i^{\text{merge}} \geq s_i^{\text{prior}}$*, exactly when*

$$\underbrace{\tfrac{1}{2}\left(s_i^b - s_i^a\right)}_{\text{recombination surplus } \Delta R_i} \;\geq\; \alpha^\star \quad \underbrace{\left(s_i^0 - s_i^a\right)}_{\text{ancestral surplus } \Delta A_i}. \tag{4}$$

Consequently the modes the merge produces and the control does not all satisfy (4)*, the modes the control produces and the merge does not all violate it, and the difference in producible-mode counts is the difference of those two set sizes.*

The count statement makes (4) a decision rule rather than an identity: it bounds which modes can change status, while the measured outcome records how many do. Section 5 uses this property to convert the three anchor measurements into a prediction of the endpoint.

Equation (4) therefore summarizes the entire comparison. Its left-hand side is the contribution of the sibling, and its right-hand side is the contribution of the ancestor, discounted by the cost of reaching it. Each regime we study corresponds to a statement about which side is larger.

Proposition 6 (Matching on distance cancels isotropic curvature)**.** *Suppose s_i is twice differentiable on $\mathcal{P}$, with gradient g_i and Hessian H_i at θ_a. Because θ_{merge} and $\theta_{\text{prior}}(\alpha^\star)$ lie at the* same *distance d from θ_a, by construction of $\alpha^\star$,*

$$s_i^{\text{merge}} - s_i^{\text{prior}} = \langle g_i, \delta_m - \delta_c \rangle + \tfrac{1}{2} d^2 \big(\hat{m}^\top H_i \hat{m} - \hat{c}^\top H_i \hat{c} \big) + O(d^3), \tag{5}$$

with steps $\delta_m = d\hat{m}$ and $\delta_c = d\hat{c}$ along the unit merge and ancestral directions. If H_i is isotropic on $\mathcal{P}$ the second-order term vanishes identically*, whatever its magnitude; in general it is at most $\frac{1}{2} d^2 (\mu_{\max} - \mu_{\min})$ and depends only on the anisotropic part of H_i. Under Assumption 1 the first term is exactly* (4)*; when the anchors are themselves read off a curved surface, at distances $2d$ and $d/\alpha^\star$ rather than d, that substitution adds a further term computable from the same Hessian.*

Matching on distance is therefore not an arbitrary fairness convention: it is precisely the choice that makes the comparison insensitive to the isotropic part of the curvature, so that only anisotropy can affect it. It also explains why the plane should be used to *choose* between two candidates rather than to *search* for a third: a search compares points at different distances and is therefore sensitive to the full Hessian rather than only to its anisotropy. Section 5 measures both effects.

Corollary 7 (A sharp threshold in the sibling's retention)**.** *Suppose the sibling retains a mode that a lost at a fraction $\beta \geq 0$ of the ancestral surplus, $s_i^b - s_i^a = \beta(s_i^0 - s_i^a)$; $\beta > 1$ indicates that the sibling exceeds the ancestor on that mode, as occurs in our measurements. Then the merge outperforms the matched control on that mode if and only if $\alpha^\star \leq \beta/2$, that is, if and only if*

$$\rho \;\geq\; 1 - \tfrac{\beta^2}{2}. \tag{6}$$

A sibling that retains the mode perfectly requires $\rho \geq \frac{1}{2}$; a sibling that has itself degraded raises the threshold sharply. Both ρ and β are measurable, so (6) *is falsifiable.*

This also explains why the confound is common rather than incidental. Across every pair we measure, on the synthetic backbone, on the grid of fine-tuned experts and on released community checkpoints, the cosine runs from 0.0002 to 0.130 with median 0.098, and every one of the 52 pairs sits below $\frac{1}{2}$. Independently trained task vectors in a space of this dimension are close to orthogonal, so the threshold in (6) *is essentially never met, even by a sibling that retains every mode.*

Any number of parents. None of the above requires $k = 2$, and neither does practice: Yu et al. (2024) merge three fine-tunes of one backbone, and community merges of four or more models are common. With parents $\theta_j = \theta_0 + \tau_j$, the uniform merge $\frac{1}{k}\sum_j \theta_j$ still carries ancestor weight exactly zero, since $1 - k \cdot \frac{1}{k} = 0$; the exchange rate becomes $\alpha_a^\star = \|\frac{1}{k}\sum_{j \neq a}(\tau_j - \tau_a)\| / \|\tau_a\|$, which at equal norms and common pairwise cosine is $\sqrt{(k-1)(1-\rho)/k}$; the attenuation is $\sqrt{(1 + (k-1)\rho)/k}$, and the conjugate identity of Corollary 4 still holds *exactly* for every k. Equation (4) becomes $\frac{1}{k}\sum_{j \neq a}(s_i^j - s_i^a) \geq \alpha_a^\star (s_i^0 - s_i^a)$, and (6) becomes

$$\rho \;\geq\; 1 - \tfrac{k-1}{k}\beta^2, \tag{7}$$

which *falls* as parents are added, from $1 - \beta^2/2$ at two toward $1 - \beta^2$ in the limit. Section 7 tests this at $k = 3$: the k-parent inequality is predictive, but the comparative reading of the decreasing threshold does not survive a control for composition.

Proposition 8 (Ancestor optimality, and the scope of the confound)**.** *If the ancestor produces every mode ($s_i^0 \geq 0$ for all i) then it attains the largest producible-mode count any parameter vector can attain, so no merge, control, or other method can exceed it on that metric. Every weight-space recombination method in such a regime is therefore competing for a fraction of an advantage that already belongs to the ancestor, and a comparison against the parents alone cannot separate the two; the matched-distance constraint is what makes the question well posed. Conversely, when each parent dominates the ancestor on its own reported metric, $\Delta A_i < 0$ there, the right-hand side of* (4) *is negative, and the confound does not arise.*

Proposition 8 delimits the scope of the confound: it is a property of the reported metric, not of the models or of their size, so the same pair of merged checkpoints can be exposed on a metric that both fine-tunes degrade and unexposed on one that each improves. Matching on weight-space L_2 is itself a choice, and Appendix A reports the matching-free version. Under Assumption 1 it also has a closed-form best step at matched distance, $x^\star \propto G^{-1}\nabla$ from the three anchors, which Appendix A.12 states and bounds. Section 7 tests the scope claim on one pair of checkpoints and two metrics.

4 EXPERIMENTAL SETUP

A ground-truthed backbone. The primary testbed is a probabilistic grammar over 48 tokens with $M = 300$ mutually exclusive valid modes, 30 head and 270 in a Zipf tail, whose true distribution and mode set are known exactly and whose classification is exact and injective, so survival and validity are measured without estimation error. A 3.19M-parameter decoder (Vaswani et al., 2017) trained from scratch as θ_0 reaches validity 0.999 and all 300 modes. Eight lineages fork from it and collapse by recursive self-training in the replace regime, which applies no filtering and so isolates neutral drift: each generation samples $20{,}000$ strings and trains on them for 500 AdamW steps at learning rate 10^{-4}; detection uses $\tau = 5 \times 10^{-5}$ over 2×10^5 samples. The collapse generation $t^\star = 27$, at which roughly 45% of the tail is gone, and every other hyperparameter were frozen on the ancestor and one sanity lineage before any merge outcome was inspected.

Language models. We replicate on Qwen2.5 under full fine-tuning (Yang et al., 2024) in four collapse settings (fact recall at 0.5B, the same with a deeper 250-subject tail, rare-template arithmetic at 0.5B, and fact recall at 1.5B), each fine-tuning the base into an ancestor covering the tail, forking lineages, and collapsing each on its own generations at roughly 45% tail-coverage loss.

Two instantiations of the margin. Assumption 1 posits a margin that is affine on the plane, and the observable that plays this role differs by testbed: on the synthetic backbone $\log_{10} \hat{p} - \log_{10} \tau$, a log frequency, which is where the assumption is tested; on the benchmark testbeds the accuracy itself, because accuracy is what the merging literature reports. On benchmarks we use the model only for the sign of a *difference* and never for a level, which Proposition 11 shows is the weaker requirement; Appendix A.14 measures a plane grid of accuracies and locates that boundary.

Controls and statistics. The matched control interpolates one parent toward the ancestor to the distance the merge travels from it, and the merge must outperform the stronger of the two. Where the task vectors differ substantially in norm, that distance can exceed the parent's own, so $\alpha^\star > 1$ would carry the control *past* the ancestor; we clip it at the ancestor, so that it travels less than the merge, which biases the comparison in favor of the merge. We also evaluate stronger variants of the control: interpolation toward an earlier checkpoint of the lineage or toward an exponential moving average, an oracle two-parent merge, and the best point on the matched-distance circle. Since pairs share lineages (Aronow et al., 2015), all endpoints carry a dyadic cluster bootstrap, resampling lineages and weighting each pair by its two endpoints' multiplicities (Owen, 2007). All geometry is double precision; merges and controls are formed in float32 before casting.

5 TESTING THE MODEL, AND PREDICTING THE CONFOUND

The theory rests on one approximation, Assumption 1, so we test it before using it. For six lineage pairs we evaluate a grid of $\theta(u, v)$, $u, v \in \{0, 0.25, 0.5, 0.75, 1\}$, plus the two exact matched-control points, measuring every mode's produced frequency at each; the three anchors are measured once and reused, and every other point is held out from (1) (Figure 3 in Appendix D).

The barycentric prediction holds. Over six pairs and 160 to 195 modes per pair measurable at all three anchors, the barycentric prediction of log frequency reaches a pooled R^2 of 0.924 on held-out grid points and 0.969 inside the convex hull, which contains the merge and both control rays. The thresholded mode count is less accurate: mean absolute error 20.2 modes overall and 47.3 at the merge, against a sampling-noise level of 4 to 6 modes, so the pre-registered gate is not met; the residual behaves as Proposition 3 requires. It is positive in 0.7517 of the 21,600 in-simplex mode observations, its mean lies inside the bracket (2) at all 72 held-out points, and Corollary 9's a priori envelope of 61.0 modes contains the measured count error everywhere (Appendix D). The missed gate is therefore not an unexplained defect but a censoring bias the theory predicts and bounds: Assumption 1 is accurate on the quantity it is stated about and conservative at the threshold by a computable margin, so every prediction below overstates the merge's disadvantage.

It predicts the confound with no fitted parameter. Using only the three anchor measurements and $\alpha^\star$ from the task-vector norms, the predicted merge-minus-best-matched-control endpoint over

all 28 lineage pairs correlates with the measurement at Pearson 0.85 ($p = 8 \times 10^{-9}$) and agrees on the sign in 28 of 28, placing the merge below the control in all of them as the measurement does. Its magnitude is conservative, -86.6 predicted against -42.1 measured.

The conservatism is due to curvature and is computable. Fitting a quadratic to the same grid with both competitors held out of the fit, 0.46 of the curvature is isotropic and therefore invisible to the matched comparison by Proposition 6. Adding the second-order term to the anchor prediction takes it from Pearson 0.807 to 0.900, and on the 202 modes far from the detection floor from 0.970 to 0.996, cutting the bias from -0.0113 to -0.0010. We retain the first-order prediction, which requires three evaluations rather than fourteen, with its loss of accuracy now quantified (Appendix D).

6 The Exchange Rate Governs the Deficit Across Regimes

Corollary 4 makes a structural prediction the attenuation reading alone does not: as $\rho \to 1$ the parents coincide, the control may travel no distance, and the deficit must vanish, so it is a function of $\alpha^\star$ through the origin, not merely correlated with $A(\rho)$. We test this on the four regimes with double-precision geometry and a matched-control outcome, 46 lineage pairs.

Across regimes. Using the exact $\alpha^\star$ of (3), the normalized merge-minus-control deficit regresses on $\alpha^\star$ at Pearson -0.608 ($p = 8 \times 10^{-6}$, 46 pairs) against $+0.527$ for the attenuation. The two are monotone in each other, so the discriminating test is the functional form: a one-parameter law through the origin explains $R^2 = 0.33$, while the attenuation parameterization needs an intercept and predicts a deficit of -0.10 at $\rho = 1$, where the theory requires zero. The reparameterization yields the correct value at $\rho = 1$ rather than a stronger correlation (Figure 4).

Across depths, with the geometry held fixed. Sweeping the ancestral surplus alone, at collapse generations 5 to 35 with $\alpha^\star$ held between 0.666 and 0.679, the measured deficit deepens from -15.5 to -72.8 modes and the parameter-free prediction tracks it without refitting: pooled over 24 points, Pearson 0.93 ($p = 4 \times 10^{-11}$), conservative throughout. Both sides of (4) are thus confirmed separately (Figure 5).

Regimes in which the theory predicts a reversal. Proposition 5 reverses when ΔA shrinks or $\alpha^\star$ falls, and both reversals are observed: under output filtering the comparison flips sign, from -0.040 coverage ($[-0.122, 0.000]$) to $+0.126$ ($[0.067, 0.178]$), while raising the retention budget contracts the deficit from -51.2 modes at pure replacement to -3.4 at full accumulation (Appendix D).

The threshold. Corollary 7 sharpens this into $\rho \geq 1 - \beta^2/2$. On the backbone β averages 0.996 over the 28 pairs (range 0.84 to 1.27), putting the threshold at $\rho \geq 0.504$ against a measured cosine of 0.116; the merge underperforms the control in all 28 pairs, as the corollary requires, and the master inequality holds on only 9% of contested modes. Section 7 tests its exact form, $\beta \geq 2\alpha^\star$.

Departures from the plane. Proposition 1 is verified numerically: uniform averaging has an out-of-plane fraction of 4×10^{-7} and recovered coordinates $u = v = 0.5000$. The sparsifying operators leave the plane, TIES by 0.60, DARE by 0.69, DARE-TIES by 0.75, and all three are worse than uniform averaging against the matched control (Table 3), so departure is not by itself an advantage. Retraining on the pooled generations leaves it almost entirely, 0.99 orthogonal, and still falls short of the control, consistent with Proposition 8 (Figure 7).

7 Which Merging Evaluations Are Exposed

Proposition 8 states that the confound is a property of the reported metric rather than of the models. We test this at three scales: one checkpoint pair; a grid of 6 base models over three families, 0.5B to 7B, all fine-tuned by us; and community checkpoints trained by third parties.

Setup. For each base model we fine-tune three narrow experts on GSM8K, CLUTRR and CRUX-Eval and form all three pairs, evaluating each on those tasks and on MATH-500, which is never a

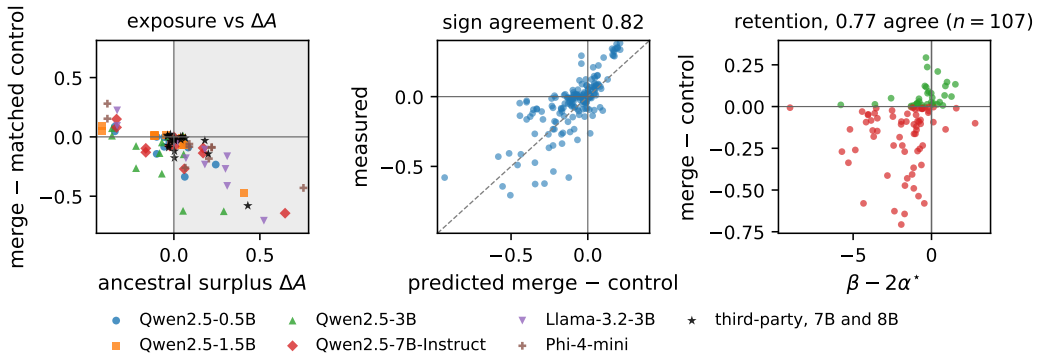


Figure 2: Every (configuration, benchmark) cell of the grid and of the third-party experiment. Left: the sign of the ancestral surplus ΔA, fixed before any merge or control accuracy was read, determines whether the matched control outperforms the merge; the shaded half is the exposed one. Middle: the barycentric prediction of merge minus control, computed from the three anchor accuracies and $\alpha^\star$ with nothing fitted, against the measurement. Right: the retention corollary in its exact unequal-norm form on cells with a positive ancestral surplus; the merge prevails to the right of $\beta = 2\alpha^\star$.

fine-tuning target. A cell is *exposed* if and only if the base model outperforms the better parent on it, which depends on the base and the parents alone.

Case study and replication. With Qwen2.5-1.5B-Instruct and two near-orthogonal experts ($\rho = 0.0112$), the merge reaches 0.402 on the multi-task axis against the base's 0.243, but the matched control reaches 0.370, 82% of its gain, and on the forgetting axis the merge underperforms the control. Rerunning it within the grid with independent seeds reproduces the exposure label in 4 of 4 cases.

The grid. Across 72 cells the control outperforms the merge on 0.97 of the 33 exposed cells and 0.44 of the 39 unexposed ones (Fisher exact $p = 5.1 \times 10^{-7}$, descriptive only, since cells inside one base model share parents), and in 6 of 6 base models with cells of both kinds (Figure 2; Table 2 breaks it down by base model). The effect does not attenuate with scale: at the largest base model, `Qwen2.5-7B-Instruct`, the control outperforms the merge on 6/6 exposed cells and 3/6 unexposed cells, and the sign prediction agrees in 0.875 of 24, above the pooled 0.82.

Parameter-free prediction. From the three anchor accuracies and $\alpha^\star$, (4) predicts merge minus control on every cell. Over 136 comparisons the sign agrees in 0.82 ($p = 4.3 \times 10^{-15}$; Spearman 0.78). The absorbed fraction is also predicted, not merely measured: it is $\alpha^\star \Delta A / \Delta R$ from the same three anchors, and over the 66 parent directions in which the model predicts, and the measurement confirms, a gain for the merge, predicted and measured shares agree at Spearman 0.73 ($p = 3.3 \times 10^{-12}$). The predicted median 1.30 against a measured 0.85 is conservative, as Propositions 3 and 6 require, and the merge prevails almost exactly where β exceeds $2\alpha^\star$ (Figure 2, right).

Third-party checkpoints. We repeat the measurement, training nothing, on community fine-tunes of a released base: three of Llama-3.1-8B-Instruct and three of Qwen2.5-7B-Instruct, trained by third parties on their own data. Their task vectors are 1 to 3% of the base norm and nearly orthogonal, the least favorable case for a merge according to (6). 7 of the Llama arm's 9 informative cells are exposed, and merging a released math fine-tune with a released tool-use fine-tune outperforms *both* parents on GSM8K, CRUXEval and MATH-500, while the matched control still recovers a median 0.69 of the merge's gain. These are measured at 1024 new tokens rather than the 256 used elsewhere, because the shorter budget truncates the *parents* rather than the merge and thereby inflates the ancestral surplus (Appendix D.7). The arm therefore supports claims about magnitude but not about sign.

Three parents. The same experts admit a uniform three-way merge and its matched controls at the $k = 3$ rate over 26 cells: the k-parent inequality is predictive, with sign agreement 0.70 over 76 ($p = 3.8 \times 10^{-4}$), while our pre-registered comparative test of (7) passes at $p = 0.048$, but we do not claim it because it is confounded by composition (Appendix D.6).

When the comparison is not degenerate. Under pure neutral drift the lineages only lose modes, so the ancestor is an upper bound and there is no reason to recombine. Self-training in practice is performed to acquire new capability, a different regime. With acquisition, where each lineage gains modes the ancestor lacks so $\Delta A_i < 0$ there, the merge reaches the Pareto frontier of the matched-control sweep in all six pairs, which it never does under neutral drift (Appendix D).

The canonical benchmark. A criterion that disfavored every merge would be uninformative, so we evaluate the setting on which much of this literature is built: the 8-task CLIP ViT-B/32 benchmark of Ilharco et al. (2023), fine-tuning every encoder ourselves and reproducing its published accuracies before reading any endpoint. The quantity that their normalisation conceals is a single parent's score on the aggregate axis: the best reaches 0.553 against zero-shot's 0.480, a surplus of -0.073, so the setting is *not* exposed and the confound cannot arise. Consistently, the merge reaches 0.645 against the best control's 0.520, which absorbs -0.369 of its gain; the merge's advantage here reflects recombination rather than reversion (Appendix F).

Levels, differences, and model selection. Proposition 6 says a matched comparison sees only the anisotropy of the curvature while a search sees all of it, and the grid confirms it: levels are predicted to 0.1159 and 0.0821 but their *difference* to 0.0869, a ratio of 0.75 against the 1.41 of independent errors ($p = 0.006$). As a search procedure it fails, outperforming both candidates in only 0.17 of 18 configurations; as a choice between the two available candidates it is effective: (4) is computable before either is evaluated and selects the better in 0.86 of cells, at regret 0.0088 against 0.1235 for always merging. It also indicates when a prediction can be trusted: a sign is certified once the predicted gap exceeds what the anisotropy alone can move (Proposition 11), agreement rising 0.564 to 0.974 across quartiles of that gap, and abstaining below a leave-one-model-out threshold answers 0.410 of 156 comparisons at 0.909 accuracy (Appendix A.13).

Comparison with simpler decision rules. We pre-registered a set of simpler decision rules and scored each on these 72 cells. The heuristics implicit in the literature perform poorly: orthogonality, the premise of interference reduction, decides at 0.542 (McNemar $p = 6.6 \times 10^{-5}$) and always merging at 0.319. The two rules nearest to it both derive from our analysis (Appendix E).

8 Discussion

Whenever the models being merged underperform their shared initialization on the reported metric, we recommend reporting the matched regression-to-ancestor control alongside the merge. The control requires a single interpolation; Proposition 8 specifies when it is needed, (4) which of the two candidates to select, and Proposition 6 why distance is the appropriate quantity to match. That the confound exists is unsurprising; its strength is less so: the control outperforms the merge on 0.97 of exposed cells, and every measured cosine lies below the threshold at which the merge could prevail. Of the three most-cited merging evaluations, one is unexposed, and the largest of the other two is also unexposed once the quantity it does not report is recovered.

Three limitations remain. Assumption 1 fits accuracy levels poorly on benchmarks (0.351 against 0.924), so we use it only for signs and differences, the weaker requirement (Proposition 11); nonlinear operators enter only through their measured departure from the plane; and the collapse backbone is a 3.19M-parameter decoder.

9 Conclusion

We studied the evaluation of merged models when fine-tuning degrades the reported metric relative to the shared initialization, and showed that comparing a merge only against its parents is then confounded by regression to the ancestor. We presented the matched regression-to-ancestor control and a geometric analysis in which a single inequality, computable before either candidate is evaluated, determines which of the two prevails; it selects the better candidate in 0.86 of cells and correctly favors the merge on the unexposed vision benchmark. Extending the analysis to nonlinear operators is a natural next step.

AI USE STATEMENT

During the preparation of this manuscript, the authors used AI tools to assist with document formatting, LaTeX code development, and technical code maintenance. In particular, an AI coding agent, Antigravity, was used to support refactoring and organizing the experimental codebase for release as supplementary material, with the aim of improving clarity and reproducibility for reviewers.

ETHICS STATEMENT

This work involves no human subjects and no personal data; the synthetic backbone and both language-model tasks are constructed environments. Its methodological contribution is a control that makes merging claims harder to overstate, which we regard as beneficial to the field. The constructive result, recovering capability from a model's own generated data, could encourage greater reliance on synthetic corpora; we therefore state explicitly that in our setting its effect is comparable to injecting roughly twenty percent real data, and that it is a bounded partial substitute for real data rather than a replacement for it.

REPRODUCIBILITY STATEMENT

Every proposition and corollary is proved in full in Appendix A and is implemented and unit tested; the suites covering the plane geometry, the censoring bracket, the curvature analysis, the in-plane optimum, the k-parent forms and the released control script hold 175 tests within a repository suite of 329. The synthetic backbone is a deterministic grammar with exact, injective mode classification, and all grammar parameters, detection thresholds, learning rates and the collapse generation were frozen on the ancestor and a sanity lineage before any recombination outcome was examined; the pre-registration is reproduced in Appendix B and hyperparameters for both testbeds in Appendix C.

The complete experiment and analysis code is available as an anonymized repository at `https://anonymous.4open.science/r/merge-plane-geometry-4591` and is included as supplementary material, as a single archive containing the plane geometry and its unit tests, the synthetic backbone, the language-model and third-party pipelines, the eight-task vision benchmark, the standalone control script and the frozen configurations; the test suite can be run directly with `pytest`, and every experiment stage is resumable, so a partial reproduction does not require repeating the full grid. Every number traces to a stored per-pair result file, and a released script recomputes each reported value from those files without reading the manuscript. A traceability map names, for every proposition and corollary, the experiment that tests it, the artifact its numbers come from and the outcome, including the 7 pre-registered predictions that failed. The matched control is released as a standalone script of about a hundred lines that takes three checkpoints and writes the two matched controls, with tests asserting that it agrees with the implementation used here and that its output travels exactly the merge's distance from the parent. Every number the replication grid, the third-party experiment and the censoring bound introduce enters the manuscript as a generated macro rather than typed text, so that an outdated value produces a build error rather than a silent inconsistency. The backbone was additionally re-trained from scratch on different hardware and a different software stack and the seven-operator panel re-run on it; Appendix D reports the outcome, including the one control that did not reproduce.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Annual Meeting of the Association for Computational Linguistics*, pp. 7319–7328, 2021.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7:195–204, 2025. doi: 10.1038/s42256-024-00975-8.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations*, 2024.

Peter M. Aronow, Cyrus Samii, and Valentina A. Assenova. Cluster-robust variance estimation for dyadic data. *Political Analysis*, 23:564–577, 2015.

Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *International Conference on Learning Representations*, 2024.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. Large language models suffer from their own output: An analysis of the self-consuming training loop. *ArXiv*, abs/2311.16822, 2023.

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*, abs/2110.14168, 2021.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. In *Neural Information Processing Systems*, 2024a.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, François Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *International Conference on Machine Learning*, pp. 11165–11197, 2024b.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. In *International Conference on Learning Representations*, 2025.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269, 2020.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *Conference on Language Modeling*, 2024.

Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong Hsu, Calvin Luo, Yonglong Tian, and Chen Sun. Self-correcting self-consuming loops for generative model training. In *International Conference on Machine Learning*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

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.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence*, pp. 876–885, 2018.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521–3526, 2017.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Neural Information Processing Systems*, pp. 17703–17716, 2022.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. doi: 10.1214/07-AOAS122.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. How bad is training on synthetic data? a statistical analysis of language model collapse. In *Conference on Language Modeling*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

George Stoica, Daniel Bolya, Jakob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. ZipIt! merging models from different tasks without training. In *International Conference on Learning Representations*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, pp. 5998–6008, 2017.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of LLM scaling based on human-generated data. In *International Conference on Machine Learning*, pp. 49523–49544, 2024.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak 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 *International Conference on Machine Learning*, 2022a.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7949–7961, 2022b.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Junyang Lin, et al. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8):1–41, 2026. doi: 10.1145/3787849.

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*, pp. 57755–57775, 2024.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *International Conference on Machine Learning*, pp. 57905–57923, 2024.

Xiangchi Yuan, Chunhui Zhang, Zheyuan Liu, Dachuan Shi, Leyan Pan, Soroush Vosoughi, and Wenke Lee. Superficial self-improved reasoners benefit from model merging. In *Conference on Empirical Methods in Natural Language Processing*, pp. 5901–5921, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Neural Information Processing Systems*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Neural Information Processing Systems*, 2022.

A PROOFS

This appendix restates and proves each result of Section 3. Throughout, θ_0 is the shared ancestor, $\tau_a = \theta_a - \theta_0$ and $\tau_b = \theta_b - \theta_0$ are the two task vectors, $\rho = \langle \tau_a, \tau_b \rangle / (\|\tau_a\| \|\tau_b\|)$, and $\theta(u, v) = \theta_0 + u\tau_a + v\tau_b$.

A.1 PROPOSITION 1

Proof. Uniform averaging gives $\frac{1}{2}(\theta_a + \theta_b) = \theta_0 + \frac{1}{2}\tau_a + \frac{1}{2}\tau_b = \theta(\frac{1}{2}, \frac{1}{2})$. Task arithmetic with coefficients (λ_a, λ_b) gives $\theta_0 + \lambda_a\tau_a + \lambda_b\tau_b = \theta(\lambda_a, \lambda_b)$. Interpolation toward the ancestor from parent a gives $(1-\alpha)\theta_a + \alpha\theta_0 = \theta_0 + (1-\alpha)\tau_a = \theta(1-\alpha, 0)$, and the matched control is this point at the particular $\alpha = \alpha^\star$ of (3). Spherical interpolation between θ_a and θ_b produces, per tensor, a combination $c_1\theta_a + c_2\theta_b$ with scalar c_1, c_2; applied globally with one pair of coefficients it is the point $\theta(c_1, c_2)$ shifted by $(c_1 + c_2 - 1)$ times θ_0's own offset, hence in the affine plane through $\theta_0, \theta_a, \theta_b$ whenever the coefficients are shared across tensors. All of these are in $\mathcal{P}$. Any method whose output is an affine combination of $\theta_0, \theta_a, \theta_b$ is in $\mathcal{P}$ by definition of the affine hull; the converse fails only for operators that act coordinate-wise with data-dependent masks, which is exactly TIES, DARE and Fisher weighting. □

A.2 PROPOSITION 2

Proof. Assumption 1 gives $s_i(\theta(u, v)) = s_i^0 + \langle g_i, u\tau_a + v\tau_b \rangle = s_i^0 + u\langle g_i, \tau_a \rangle + v\langle g_i, \tau_b \rangle$. Evaluating the same expansion at $(1, 0)$ and $(0, 1)$ gives $\langle g_i, \tau_a \rangle = s_i^a - s_i^0$ and $\langle g_i, \tau_b \rangle = s_i^b - s_i^0$. Substituting, $s_i(u, v) = s_i^0 + u(s_i^a - s_i^0) + v(s_i^b - s_i^0) = (1 - u - v)s_i^0 + us_i^a + vs_i^b$. Setting $(u, v) = (\frac{1}{2}, \frac{1}{2})$ makes the coefficient of s_i^0 equal to $1 - \frac{1}{2} - \frac{1}{2} = 0$; setting $(u, v) = (1 - \alpha, 0)$ makes the coefficient of s_i^b equal to 0 and the coefficient of s_i^0 equal to α. □

Two remarks. First, the result is exact under Assumption 1 and requires no further structure: it does not depend on the dimension, on the architecture, or on the merge being small. Second, the three anchor margins are directly measurable, which is what makes the downstream predictions parameter-free.

A.3 PROPOSITION 3

Proof. Write p_k for the frequency of mode i at anchor $k \in \{0, a, b\}$ and $w = (1 - u - v, u, v)$ with $w \geq 0$, $\sum_k w_k = 1$. Since $s_i^k = \log p_k - \log \tau$, the barycentric prediction is $s_i(u, v) = \sum_k w_k \log p_k - \log \tau = \log \prod_k p_k^{w_k} - \log \tau$, the log of the weighted *geometric* mean, while the mixture margin is $\tilde{s}_i(u, v) = \log \sum_k w_k p_k - \log \tau$, the log of the weighted *arithmetic* mean. The left inequality of (2) is the arithmetic-geometric mean inequality, with equality iff all p_k with $w_k > 0$ coincide. For the right inequality, $\sum_k w_k p_k \leq \left(\sum_k w_k\right) \max_{k: w_k > 0} p_k = \max_{k: w_k > 0} p_k$, so $\tilde{s}_i \leq \max_{k: w_k > 0} s_i^k$; subtracting $s_i = \sum_k w_k s_i^k$ gives D_i. For the lower bracket, $\sum_k w_k p_k \geq w_j p_j$ for every j with $w_j > 0$, so $\tilde{s}_i \geq \max_{j: w_j > 0} \left[\log w_j + s_i^j\right]$; taking j to be the anchor attaining $\max_k s_i^k$ gives $\tilde{s}_i - s_i \geq D_i - \log(1/w_{\max})$. At the uniform merge $w = (0, \frac{1}{2}, \frac{1}{2})$, so $w_{\max} = \frac{1}{2}$ and the bracket has width exactly $\log 2$. □

A.4 COROLLARY 9

Corollary 9 (Count-error envelope)**.** *The number of behaviours the mixture produces and* (1) *does not is at most* $\#\{i : -D_i(u, v) \leq s_i(u, v) < 0\}$*, and* (1) *never over-counts. The error of the thresholded*

count therefore has an a priori bound computable from the three anchors, with no access to the merge.

Proof. Suppose the mixture produces behaviour i at (u, v) and (1) does not, that is $\tilde{s}_i \geq 0 > s_i$. By the right inequality of (2), $\tilde{s}_i \leq s_i + D_i$, so $0 \leq s_i + D_i$, that is $s_i \geq -D_i$. Hence every behaviour counted as a miss lies in $[-D_i, 0)$ under the prediction, and the number of misses is at most the size of that set. The left inequality of (2) gives $\tilde{s}_i \geq s_i$, so a behaviour the prediction produces is produced by the mixture too and there are no misses in the other direction. □

The bound stays finite when an anchor's frequency is censored at zero, which is the regime that produces the residuals of Section 5; a variance-based Jensen bound does not, because its constant is the reciprocal of the smallest frequency. The cost is that the bracket is one-sided in information: it establishes that the model understates the frequency and bounds the amount, but does not determine it exactly.

A.5 Equation (3) and Corollary 4

Proof. $\theta_{\text{merge}} - \theta_a = \frac{1}{2}(\theta_a + \theta_b) - \theta_a = \frac{1}{2}(\tau_b - \tau_a)$, so the merge sits at distance $\frac{1}{2}\|\tau_b - \tau_a\|$ from parent a. The control $\theta_{\text{prior}}(\alpha) = \theta(1 - \alpha, 0)$ sits at distance $\|\theta_{\text{prior}}(\alpha) - \theta_a\| = \|-\alpha\tau_a\| = \alpha\|\tau_a\|$. Equating gives (3). When $\|\tau_a\| = \|\tau_b\| = L$, $\|\tau_b - \tau_a\|^2 = 2L^2(1 - \rho)$, so $\alpha^\star = L\sqrt{2(1-\rho)}/(2L) = \sqrt{(1-\rho)/2}$. Then $\alpha^\star(\rho)^2 + A(\rho)^2 = \frac{1-\rho}{2} + \frac{1+\rho}{2} = 1$. □

The asymmetry of (3) should be noted: the control built from the shorter parent must traverse a larger fraction of its own task vector to cover the same distance, so the two matched controls are not mirror images, and the strongest comparison uses the better of the two, as we do throughout.

A.6 Proposition 5

Proof. By Proposition 2, $s_i^{\text{merge}} = \frac{1}{2}(s_i^a + s_i^b)$ and $s_i^{\text{prior}} = (1 - \alpha^\star)s_i^a + \alpha^\star s_i^0$. Then

$$s_i^{\text{merge}} - s_i^{\text{prior}} = \tfrac{1}{2}s_i^a + \tfrac{1}{2}s_i^b - s_i^a + \alpha^\star s_i^a - \alpha^\star s_i^0 = \tfrac{1}{2}(s_i^b - s_i^a) - \alpha^\star(s_i^0 - s_i^a),$$

which is non-negative exactly when (4) holds. Since a mode is produced iff its margin is non-negative, the merge produces mode i whenever the control does not precisely on the set where (4) holds and $s_i^{\text{merge}} \geq 0 > s_i^{\text{prior}}$; the inequality itself is the margin comparison, which is the quantity that determines the count difference summed over modes. □

A.7 Proposition 6

Proof. Write δ for a displacement from θ_a inside $\mathcal{P}$ and expand $s_i(\theta_a + \delta) = s_i^a + \langle g_i, \delta\rangle + \frac{1}{2}\delta^\top H_i \delta + O(\|\delta\|^3)$. The merge is $\delta_m = \frac{1}{2}(\tau_b - \tau_a)$ and the control is $\delta_c = -\alpha^\star\tau_a$, and $\alpha^\star$ was defined by $\|\delta_m\| = \|\delta_c\| =: d$, so $\delta_m = d\hat{m}$ and $\delta_c = d\hat{c}$ for unit $\hat{m}, \hat{c}$. Subtracting,

$$s_i^{\text{merge}} - s_i^{\text{prior}} = \langle g_i, \delta_m - \delta_c\rangle + \tfrac{1}{2}d^2\big(\hat{m}^\top H_i \hat{m} - \hat{c}^\top H_i \hat{c}\big) + O(d^3).$$

The first term is the affine part, which by Proposition 2 is exactly $\frac{1}{2}(s_i^b - s_i^a) - \alpha^\star(s_i^0 - s_i^a)$ when the anchors are read affinely, giving (5). If $H_i = \lambda I$ on $\mathcal{P}$ then $\hat{m}^\top H_i \hat{m} = \hat{c}^\top H_i \hat{c} = \lambda$ and the second-order term is zero identically, for any λ. In general both Rayleigh quotients lie in $[\mu_{\min}, \mu_{\max}]$, the extreme eigenvalues of H_i restricted to $\mathcal{P}$, so the term is at most $\frac{1}{2}d^2(\mu_{\max} - \mu_{\min})$ in absolute value; and since adding any multiple of the identity shifts both quotients equally, it depends only on the anisotropic part of H_i. □

The equality of the two step lengths is the only thing the cancellation uses. A comparison of two plane points at *different* distances from θ_a retains a term proportional to the difference of the squared distances times the isotropic curvature, which is why a search over the plane is exposed to the whole Hessian. Reading the anchors themselves off a curved surface adds a further known term, since θ_b and θ_0 sit at distances $2d$ and $d/\alpha^\star$ rather than d; that term is what Section 5 fits and reports.

A.8 ANY NUMBER OF PARENTS

Proof. With $\theta_j = \theta_0 + \tau_j$ for $j = 1, \dots, k$, the uniform merge is $\theta_0 + \frac{1}{k}\sum_j \tau_j$, whose plane coordinates are $u_j = \frac{1}{k}$, so the ancestor's barycentric weight is $1 - \sum_j u_j = 1 - k \cdot \frac{1}{k} = 0$. Its displacement from parent a is $\frac{1}{k}\sum_j \tau_j - \tau_a = \frac{1}{k}\sum_{j \neq a}(\tau_j - \tau_a)$, and the control $\theta_a - \alpha\tau_a$ moves $\alpha\|\tau_a\|$, so equating gives $\alpha_a^\star$. Substituting the barycentric margins of the two points into $s^{\text{merge}} \geq s^{\text{prior}}$ and cancelling s_i^a gives $\frac{1}{k}\sum_{j \neq a}(s_i^j - s_i^a) \geq \alpha_a^\star(s_i^0 - s_i^a)$.

For equal norms L and common pairwise cosine ρ, expanding $\|\sum_{j \neq a}(\tau_j - \tau_a)\|^2$ gives $k-1$ diagonal terms $2L^2(1-\rho)$ and $(k-1)(k-2)$ off-diagonal terms $L^2(1-\rho)$, hence $k(k-1)L^2(1-\rho)$ and $\alpha^\star = \sqrt{(k-1)(1-\rho)/k}$. Likewise $\|\frac{1}{k}\sum_j \tau_j\|^2 = L^2(1 + (k-1)\rho)/k$, so the attenuation is $\sqrt{(1 + (k-1)\rho)/k}$ and $\alpha^{\star 2} + A^2 = \frac{(k-1)(1-\rho)+1+(k-1)\rho}{k} = \frac{k}{k} = 1$. Finally, writing $s_i^j - s_i^a = \beta(s_i^0 - s_i^a)$ for every sibling makes the left side $\frac{k-1}{k}\beta(s_i^0 - s_i^a)$, so with a positive ancestral surplus the merge prevails if and only if $\alpha_a^\star \leq \frac{k-1}{k}\beta$, that is, if and only if $\rho \geq 1 - \frac{k-1}{k}\beta^2$. □

A.9 COROLLARY 7

Proof. Substituting $s_i^b - s_i^a = \beta(s_i^0 - s_i^a)$ into (4) gives $\frac{\beta}{2}(s_i^0 - s_i^a) \geq \alpha^\star(s_i^0 - s_i^a)$. On a contested mode the ancestral surplus is strictly positive ($s_i^0 \geq 0 > s_i^a$), so dividing preserves the direction and the condition is $\alpha^\star \leq \beta/2$. With $\alpha^\star = \sqrt{(1-\rho)/2}$ this is $(1-\rho)/2 \leq \beta^2/4$, that is $\rho \geq 1 - \beta^2/2$. At $\beta = 1$ the threshold is $\rho \geq \frac{1}{2}$; at $\beta = 0$ it is $\rho \geq 1$, which is unattainable, correctly encoding that a sibling with nothing to give can never carry the merge. □

A.10 PROPOSITION 8

Proof. If $s_i^0 \geq 0$ for every i then the producible-mode count at θ_0 is M, the total number of modes, which is an upper bound on the count at any parameter vector. Hence θ_0 is a global maximizer, and in particular no point of $\mathcal{P}$ exceeds it; note that this half needs no appeal to Assumption 1, only the premise that the ancestor produces everything. For the converse, if parent a dominates the ancestor on the reported metric then $s_i^a > s_i^0$ on the modes in question, so $\Delta A_i = s_i^0 - s_i^a < 0$, the right-hand side of (4) is negative, and the merge satisfies (4) there for any non-negative recombination surplus. □

The first part of the proposition is elementary by design. In a degraded-parent regime the ancestor is not merely a strong baseline but an upper bound on the merge plane, so outperforming the parents is uninformative. The matched-distance constraint makes the comparison well posed by asking how much of the ancestor's advantage a fixed step length recovers in the sibling direction relative to the ancestral direction. The second part is the scope statement tested in Section 7: the confound follows the sign of ΔA, which is a property of the reported metric.

A.11 THE MATCHING-FREE SUMMARY IN FULL

Matching on weight-space L_2 distance is a modelling choice, so we also give a formulation that requires no matching metric. The merge's *ancestral equivalent* α_{eq} is defined as the interpolation coefficient at which pure reversion toward the ancestor attains the merge's own score, $R(\theta_{\text{prior}}(\alpha_{\text{eq}})) = R(\theta_{\text{merge}})$. This definition involves no matching metric; it expresses the merge's value in units of reversion. Comparing α_{eq} with the distance budget $\alpha^\star$ that the merge actually spends recovers the same comparison, and $\alpha_{\text{eq}} < \alpha^\star$ indicates that the sibling direction makes less effective use of that budget than the ancestral direction.

A.12 THE BEST DIRECTION AT MATCHED DISTANCE

Proposition 10 (In-plane optimum)**.** *Under Assumption 1 the summed margin $\sum_i s_i$ is affine on $\mathcal{P}$ with gradient $\nabla = \big(\sum_i(s_i^a - s_i^0),\ \sum_i(s_i^b - s_i^0)\big)$ in plane coordinates. Among all steps x from θ_a with $x^\top G x = d^2$, where G is the Gram matrix of (τ_a, τ_b) and d the merge's own step length, the*

summed margin is maximised uniquely at $x^\star = d\,G^{-1}\nabla/\sqrt{\nabla^\top G^{-1}\nabla}$*, and any competitor* x *attains a fraction* $\cos_G(x, x^\star)$ *of that maximum.*

Proof. Affineness gives $\sum_i s_i(\theta_a + x) = \text{const} + \nabla^\top x$, and maximising a linear functional over an ellipsoid $x^\top G x = d^2$ is a Cauchy–Schwarz problem in the G inner product: writing $w = G^{-1}\nabla$, $\nabla^\top x = \langle w, x\rangle_G \le \|w\|_G \|x\|_G$ with equality iff $x \parallel w$. Scaling w to length d gives $x^\star$, and the attained fraction is the G-cosine. □

The merge step $(-\frac{1}{2}, \frac{1}{2})$ and the control step $(-\alpha^\star, 0)$ are two particular admissible directions, so this replaces a statement about when the ancestor cannot be beaten with a measurement of how much of the attainable gain each competitor captures. On the backbone the optimum has both coordinates negative in four of the six pairs, that is it unwinds both task vectors at once, which is neither competitor's direction.

The proposition optimises the summed margin, the linear surrogate, and not the thresholded count that our endpoints use. We pre-registered that gap as a testable claim and it fails. Evaluating $\theta_a + x^\star$ against the uniform merge and the matched control on all 6 pairs, in one run at $n_{\text{eval}} = 50000$ so that the points are comparable, the surrogate optimum exceeds both competitors in 0 of 6 pairs against a pre-registered gate of four of six. It beats the merge in 2 of 6 and the matched control in 0 of 6, at a mean of -51.7 modes and a median of -69 against the control, both far outside the four-to-six mode sampling yardstick; the sign of the measured count difference agrees with the sign of the surrogate difference in 0 of 6. The mechanism is the threshold: unwinding both task vectors raises many already-dead modes a little way without carrying them across τ, which the summed margin rewards and the count does not. We therefore state Proposition 10 for the summed margin only and leave the relationship to the count open. It concerns directions rather than evaluated models, and the direction it identifies is not recommended for deployment.

A.13 When may a sign prediction be trusted?

Appendix A.14 shows that accuracy is not an affine function of the plane coordinates, while Section 7 shows the parameter-free sign prediction agreeing with measurement far more often than chance. The two observations are reconciled by a simple inequality whose content lies entirely in the bound on its right-hand side.

Proposition 11 (Sign certificate)**.** *Let* $\widehat{D}$ *be the predicted merge-minus-control and* D *the measured one. Then* $\operatorname{sign}(\widehat{D}) = \operatorname{sign}(D)$ *whenever* $|\widehat{D}| > |\widehat{D} - D|$.

Proof. If $|\widehat{D} - D| < |\widehat{D}|$ then D lies in the open interval of radius $|\widehat{D}|$ about $\widehat{D}$, which does not contain 0 and lies entirely on one side of it. □

The inequality is elementary; what matters is the error it involves, which is the error on the *difference*, and Proposition 6 bounds that error by the *anisotropic* part of the curvature alone, because the isotropic part cancels between two steps of equal length. A sign prediction therefore survives an arbitrarily large isotropic curvature, which is roughly half of what Appendix A.14 measures on accuracy. Affineness of levels was never the requirement; a predicted gap larger than the anisotropy can move is.

$|\widehat{D} - D|$ is not knowable before both candidates are evaluated, so we instead test whether $|\widehat{D}|$, which is computable from the three anchors alone, predicts whether the sign is correct. Over the 156 comparisons of the grid and the third-party arms, it does. The certificate condition holds on 0.558 of them, and where it does not sign agreement is 0.449, indistinguishable from a coin. Agreement rises monotonically across quartiles of $|\widehat{D}|$, 0.564, 0.667, 0.821, 0.974, with Spearman 0.359 between the predicted magnitude and being correct ($p = 4.2 \times 10^{-6}$). A rule that answers only above a threshold, chosen leave-one-model-out over the 7 base models so it is never fitted on the cells it is scored on, answers on 0.410 of comparisons and is correct on 0.909 of those.

This measurement was not pre-registered. It was written in response to Appendix A.14's result and is reported as post hoc; Proposition 11 itself is a theorem and needs no experiment.

A.14 Is the margin affine when the margin is the accuracy?

Assumption 1 is tested on the synthetic backbone in its log-frequency instantiation and then used on the benchmark testbeds as an accuracy, where it had not been tested directly. This is the largest untested step between the theory and the paper's main results, so we tested it directly, pre-registering the endpoint and the gate before any point was evaluated.

For `Qwen2.5-1.5B-Instruct` and the GSM8K $\times$ CLUTRR pair we evaluated the plane grid $u, v \in \{0, 0.25, 0.5, 0.75, 1\}$ with $u + v \leq 1$, 15 points, on all four benchmarks, using the same experts, item sets and decoding budget as the grid of Section 7 so the points are comparable to it. The pair's cosine came out at 0.0142, against the 0.0112 measured independently for the same task pair in Section 7. We then fitted the affine model on the three anchors alone and predicted the remaining 12.

The gate fails. Held-out R^2 is 0.351 averaged over the four benchmarks and 0.061 at worst, against 0.924 for the backbone's log frequency, at a held-out MAE of 0.0768, or 7.7$\times$ this evaluation's own reproducibility noise. Accuracy is not an affine function of the plane coordinates at this resolution, and every benchmark prediction in this paper rests on a model that does not hold there as a model of levels.

Two properties nevertheless hold, and they are the two on which the argument depends. First, the isotropic share of the fitted curvature is 0.536, against 0.464 on the backbone, so the cancellation Proposition 6 relies on is present here at the same strength. Second, the residual is one-signed: the affine model under-predicts interior plane points on all four benchmarks, which is the direction Proposition 3 gives for the backbone and means a difference between two plane points cancels part of what a level is exposed to.

We checked that cancellation on these records directly, and report it as *post hoc*, since it was not pre-registered. Over 32 held-out point pairs at matched distance the difference MAE is 0.0672 against a level MAE of 0.0768, a ratio of 0.875 where two independent errors would give 1.41. The direction agrees with the pre-registered version of the same test on the 72 benchmark cells of Section 7 (ratio 0.75, $p = 0.006$), but on this grid it is not significant ($p = 0.241$) and one benchmark reverses. We therefore treat it as consistent with Proposition 6 and not as evidence for it, and the load-bearing version of the claim remains the one measured on the grid.

In summary, the parameter-free sign predictions of Section 7 perform better than the fit of the model that motivates them would suggest. We report this discrepancy rather than attempt to resolve it.

A.15 What the theory does not claim

Assumption 1 is a modelling assumption about behaviour on a two-dimensional slice, and it is the only place where an approximation enters. It is tested directly in Section 5 rather than argued for, and every downstream prediction is reported against its measured counterpart. The theory says nothing about merges that leave the plane, which is why the nonlinear operators and the data-space comparison are handled empirically in Section 6. It also says nothing about which of the two competitors a practitioner should prefer; it says only that a comparison that omits the control has not measured recombination.

B Pre-registration

Every experiment in this paper was pre-registered in a dated project log with its parameters, endpoints and decision gates fixed before any outcome was inspected. The log is released with the code, at the repository cited in the reproducibility statement. The essential content is as follows.

Frozen before any merge was inspected (synthetic backbone). The grammar and its Zipf calibration, the detection threshold $\tau = 5 \times 10^{-5}$, the evaluation sample size, the fixed evaluation temperature, the self-training learning rate, and the collapse generation $t^\star = 27$. The primary merge operator (uniform averaging) was pre-declared, with the remaining six operators a robustness panel corrected by a step-down procedure.

Frozen before the merge-plane experiment. The grid values, the six lineage pairs, the grid sample size, the two evaluation seeds, the rule that a mode counts as measurable only if it was sampled at least five times at all three anchors, and the gate: the first-order model is judged adequate only if its error on the producible-mode count at the merge point and along the control ray is within the measurement's own sampling noise, estimated by re-evaluating an anchor at a second seed. The sharp threshold of Corollary 7 was pre-declared as supported only if it separates the regimes with no exception, and a single exception would be reported as falsifying the sharp form while leaving Theorem 5 intact.

Frozen before the adaptation experiment. The acquisition-pool size, the fixed seed that draws it, the per-lineage disjoint split, and the adaptation fraction of each generation's training set. The experiment was declared not to require any positive outcome: its purpose is to establish whether the comparison is degenerate, and that finding is the result either way.

Frozen before the external-validity experiment. The base model, both fine-tuning corpora and their sizes, all optimization settings, the evaluation protocol and item counts, the definition of both axes, and the prediction that the control absorbs a substantial share of the merge's gain on the axis both fine-tunes degrade and approximately none on the axis each improves. It was declared in advance that if the narrow fine-tunes did not in fact degrade the second axis, the arm would be reported as inconclusive rather than the axis re-chosen.

Frozen before the replication grid. The five base models and their order, the three fine-tuning tasks and therefore all three pairs per base, the four evaluation sets (byte-identical to the case study's, verified item by item), the fine-tuning protocol (imported from the case study, not re-implemented), the disjoint CRUXEval train and evaluation split, the rule that a cell counts as exposed if and only if the ancestral surplus is positive, and five endpoints: the prevalence of the confound split by exposure; the parameter-free sign agreement and rank correlation of the barycentric prediction of merge minus control, with sign agreement and not magnitude as the endpoint because accuracy is bounded and censors; the aggregate multi-task axis; the plane search; and the decision-rule reading of the master inequality. A base model that would not train or evaluate was declared to be reported as attempted and not completed rather than dropped. Three additional views, none of which replaces an endpoint, were declared before the first analysis: the separation per base model, since cells inside one base share parents and the pooled test is descriptive; the endpoint restricted to cells whose item-level bootstrap interval excludes zero; and a sensitivity on cells whose exposure label is larger than the evaluation noise. A sixth base model, `Qwen2.5-7B-Instruct`, was added afterwards under its own pre-registration, written before any 7B number existed, which declared the grid's endpoints unchanged and set a falsification criterion of monotone weakening of the separation with scale. Its twelve cells are pooled into the grid exactly as that pre-registration specified, so the 72 cells reported here span 6 base models and not five, and the pooling is a pre-registered extension rather than a post-hoc addition.

Analyses added after the data existed, and labelled as such. Three, none of which replaces a pre-registered endpoint. The plane search's informative-benchmark sensitivity and its mean-normalised secondary objective, both added after the pre-registered search failed and reported beside the pre-registered value. The absorbed-fraction summary of the third-party arms, a re-expression of quantities the prevalence endpoint already computes, added after those arms turned out to have differences at the noise floor. Finally, two analyses in Section 7 that use no new evaluation and test consequences derived from the geometry rather than from the grid, but were run after the grid was collected: the level-versus-difference test of Proposition 6, and the prediction of the absorbed fraction.

Frozen before the third-party experiment. Both base models and all six community checkpoints, selected by verifying from their configs that each declares that exact base and matches its architecture and vocabulary, before any of them was downloaded. Three guards were declared with it: a basin check reporting each parent's task-vector norm relative to the base and the held-out token-level likelihood of every merge; a chat-template rule that evaluates each parent under both its own template and the base's and takes the better, which shrinks the ancestral surplus and so works against the effect we are testing for; and an informativeness rule dropping any cell in which every point scores below

0.05, since such a cell would otherwise count as evidence for us while carrying no signal. It was declared that a pair failing the basin check would be reported as unmergeable rather than replaced.

Limitations of the external-validity design. The forgetting axis, intended as the exposed arm, has an ancestral surplus of $+0.002$ against the better parent, inside the 0.01 evaluation noise, so its axis-level label is ambiguous even though the matched control outperforms the merge there. In addition, the GSM8K fine-tune did not improve GSM8K accuracy (0.483 to 0.453), the instruct model being already strong on it, so parent A is a weak instance of a target-task fine-tune. Neither was corrected by re-running with different settings, since doing so after observing the outcome is precisely what pre-registration is designed to prevent.

Pre-registered predictions that failed. There are seven, each reported as failing rather than re-scoped. (i) The producible-count gate for Assumption 1, missed by a wide margin and diagnosed in Section 5. (ii) The case study's prediction that the canonical multi-task axis would be unexposed, refuted by our own measurement, which broadened the claim. (iii) The comparative reading of (7), that a three-way merge should be less disadvantaged than the two-way merges of the same experts: the pre-registered comparison passes at $p = 0.048$ but is confounded by composition, and both controls for that confound return null results, so we report it as not supported (Appendix D.6). (iv) The plane search of Section 7, pre-registered as a constructive method, which outperforms both candidates in only 0.17 of 18 configurations. (v) The transfer of the in-plane optimum to the thresholded count, which outperforms both competitors in 0 of 6 pairs against a pre-registered gate of four of six (Appendix A.12). (vi) The affineness of the margin on benchmark accuracy, with a held-out R^2 of 0.351 against 0.924 on the backbone (Appendix A.14). (vii) The prediction that the matched control would still match or exceed the merge on the third-party arm at a fair generation budget (Appendix D.7).

Errors identified and corrected. An early diversity-matched baseline compared a small-sample distinct-mode count against a full-sample count, which inflated the single source's sampling temperature and produced a spurious negative; correcting to equal sample sizes reversed the sign, and the provenance and unique-recovery metrics, which do not depend on the matching choice, were positive throughout. A single-precision task-vector cosine once returned a value above one before clamping and led to an incorrect mechanism claim, which is why all geometry here is double precision.

C HYPERPARAMETERS

Table 1 lists the frozen settings. All synthetic values were fixed on the ancestor and one sanity lineage before any recombination outcome was inspected.

Table 1: Frozen experimental hyperparameters.

	Synthetic backbone	Language model	E12 external
Model	3.19M decoder (scratch)	Qwen2.5-0.5B/1.5B-Instruct	Qwen2.5-1.5B-Instruct
Adaptation	full training	full fine-tuning	full fine-tuning
Optimizer	AdamW	AdamW	AdamW
Lineages (seeds)	8	4 to 8	2 narrow fine-tunes
Generations	40	20	n/a
Samples per generation	20,000	2,000	4,000 train items
Steps per generation	500	150	600 total
Learning rate	10^{-4}	10^{-5}	10^{-5}
Batch size	256	256	8
Sampling temperature	1.0	1.0	greedy at eval
Detection threshold τ	5×10^{-5}	10^{-3}	n/a
Evaluation samples	2×10^5	8,000	400 items
Collapse generation $t^\star$	27 ($\approx 45\%$ tail lost)	per lineage ($\approx 45\%$)	n/a
Bootstrap resamples (dyadic)	10^4	10^4	n/a

The merge-plane grid of Section 5 uses $u, v \in \{0, 0.25, 0.5, 0.75, 1\}$ with 5×10^4 evaluation samples per grid point and the two exact matched-control points appended, over six lineage pairs; the anchors

Table 2: The confound by base model. A cell is exposed when the base outperforms the better parent on it, a label fixed before any merge accuracy is read; "ctrl wins" counts cells in which the matched control outperforms the merge, "median" is the median of merge minus control, and daggered rows are third-party checkpoints.

	exposed ($\Delta A > 0$)			not exposed ($\Delta A < 0$)		
base model	cells	ctrl wins	median	cells	ctrl wins	median
Qwen2.5-0.5B	4	4/4	-0.162	8	4/8	-0.002
Qwen2.5-1.5B	4	4/4	-0.068	8	1/8	$+0.013$
Qwen2.5-3B	4	3/4	-0.385	8	5/8	-0.062
Llama-3.2-3B	8	8/8	-0.207	4	2/4	$+0.027$
Phi-4-mini	7	7/7	-0.090	5	2/5	$+0.000$
Qwen2.5-7B†	5	5/5	-0.143	4	2/4	-0.033
Llama-3.1-8B†	6	4/6	-0.012	3	2/3	-0.023
all, grid	27	26/27	-0.163	33	14/33	$+0.000$
all, third-party†	11	9/11	-0.030	7	4/7	-0.023

and the merge-versus-control measurement use the full 2×10^5 samples over all 28 pairs. All task-vector geometry is computed in double precision, since single precision on full-fine-tuning task vectors of this norm overestimates ρ by up to 0.26 and can return a cosine above one before clamping.

D Supporting Measurements

This appendix collects the measurements that support the main text but are not part of its argument. All values trace to stored per-pair artifacts.

D.1 The first-order model: residuals and diagnostics

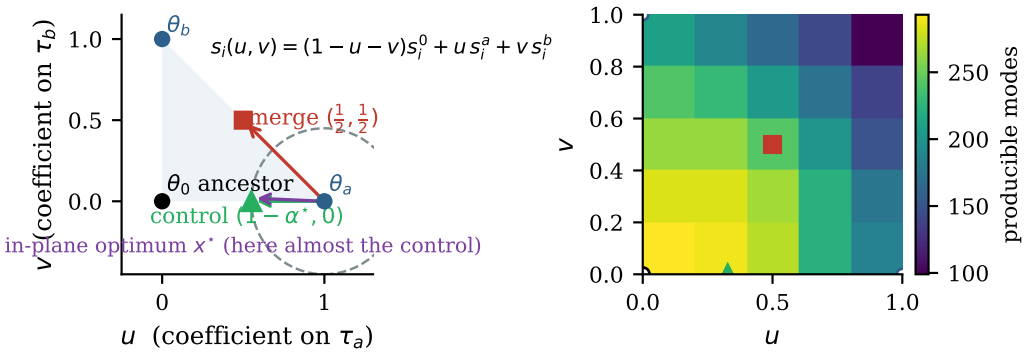


Figure 3: The merge plane. Left: the ancestor and the two parents are the three anchors of a two-dimensional affine plane that contains every linear merge and every ancestor interpolation. The uniform merge and the matched control leave the same parent along the two edges of the triangle, the dashed circle is the matched-distance constraint that fixes the exchange rate $\alpha^\star$, and the purple arrow is the in-plane optimum of Appendix A.12, computed from the anchors alone, which for this pair almost coincides with the control. Right: the measured producible-mode count over that plane for one lineage pair on the synthetic backbone, with the merge (square), the matched control (triangle), the ancestor (white circle) and the two parents marked.

Figure 6 shows the two panels of Section 5. Per pair, the barycentric R^2 of log frequency on held-out grid points is 0.899 to 0.934 over all points and 0.959 to 0.975 inside the convex hull, over 160 to 195 modes measurable at all three anchors. The producible-mode count is predicted with a mean absolute error of 15.8 to 27.2 modes per pair and a pooled bias of -8.5 modes. The residual on the merge point, broken down by how many parents still express a mode, is $+0.06$ decades where both do, $+0.25$ to $+0.35$ where exactly one does, and $+0.38$ to $+0.65$ where neither does; the sampling-noise yardstick from re-evaluating an anchor at a second seed is 1 mode for the ancestor, 4

modes for a collapsed lineage at full resolution and 6 modes at grid resolution. The parameter-free prediction of the endpoint over all 28 pairs has dyadic cluster intervals $[-107.0, -68.7]$ predicted against $[-56.0, -28.3]$ measured, with sign agreement in 28 of 28 pairs.

The censoring bracket, in full. Proposition 3 is proved for the mixture foil; a real merged network is not a mixture, so we report exactly how much of it survives on the real merges (`scripts/exp10g_censoring.py`, which re-reads the frequency files E10 already wrote and trains nothing). Over 72 held-out in-simplex grid points and 21,600 mode observations the measured residual is non-negative in 0.7517 and inside the bracket in 0.5569 (0.2169 below, 0.2262 above), so it is not a per-mode envelope. The mean residual is 0.149 decades against a mean upper bound of 0.332 and a mean bracket width of 0.335, and the point-level mean lies inside the point-level bracket at all 72 points, with Pearson 0.659 between the two across points. By parent status the mean residual and bound are 0.0375 against 0.1461 where both parents express the mode and 0.3238 against 0.5153 where neither does. The in-simplex count bias is -21.7 modes, negative at 87.5% of points; this is not in tension with the pooled -8.5 quoted above, which averages over all held-out points including extrapolations outside the simplex, where the weights are not a convex combination and (2) does not apply.

Figure 5 shows the depth sweep. Per-pair values are stored in the released artifacts.

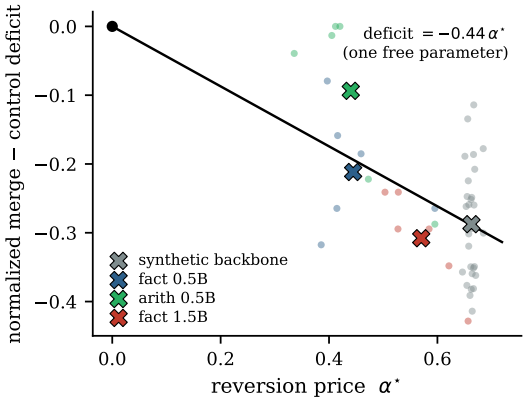


Figure 4: The exchange rate $\alpha^\star$ is the relevant variable: the normalized merge-minus-control deficit against the exact $\alpha^\star$ of (3) over 46 lineage pairs in four regimes. Small points are pairs, crosses are setting means, and the line is the one-parameter law through the origin that the theory requires. Figure 8 in Appendix D plots the two conjugate factors against ρ.

The merge's ancestral equivalent. This comparison requires no matching metric. Over the six pairs the merge's producible-mode count is matched by pure reversion at α_{eq} between 0.16 and 0.36, mean 0.264, while its own distance permits the control a coefficient of $\alpha^\star = 0.667$: the merge attains 40% of the reversion that its step length would allow.

The in-plane oracle. Among grid points within 25% of the matched distance from a parent, the best is the ancestral direction itself ($u = 0.25$, $v = 0$) in all six pairs, recovering 283 to 289 modes against the merge's 209 to 244 and the control's 275 to 289. The search is coarse, at the grid's 0.25 resolution and with a tolerance that lets the winner sit slightly beyond the matched point, so it is an existence statement about directions rather than a precise optimum: within the plane the ancestor is the optimum, and the matched-distance constraint measures what fraction of that optimum a step spent toward the sibling recovers.

The two reversals in full. Under output filtering, collapse still destroys rare templates (tail coverage 0.95 to 0.37) but the lineages stay valid, so ΔA is small on the validity coordinate, while ρ rises from 0.745 to 0.826, lowering $\alpha^\star$ from 0.357 to 0.295 in the equal-norm form (that run stored the cosine but not the norms). The coverage comparison flips from -0.040 ($[-0.122, 0.000]$, failing in four of six pairs) to $+0.126$ ($[0.067, 0.178]$, failing in none), with merged validity 0.887 against

0.998. Earlier reporting attributed the reversal to validity alone because the cosines were similar; both factors contribute, a change in ρ from 0.745 to 0.826 being a 17% change in the exchange rate.

The cross-regime law in full. We use the exact $\alpha^\star$ of (3) rather than its equal-norm form, recovering each pair's norm ratio from its stored cosine and measured attenuation; the ratios range from 0.67 to 1.66, so the equal-norm approximation is accurate on the synthetic backbone but not on the language models. Regressing the normalized merge-minus-control deficit on $\alpha^\star$ gives Pearson -0.608 ($p = 8 \times 10^{-6}$, 46 pairs) against $+0.527$ for the attenuation; the two are monotone in each other, so the discriminating test is the functional form. A one-parameter law through the origin, deficit $= -0.436\,\alpha^\star$, explains $R^2 = 0.33$, while the same two-parameter fit in the attenuation parameterization carries an intercept of -0.87 and predicts a deficit of -0.10 at $\rho = 1$, where the theory requires zero (Figure 4, Appendix D). Setting-level residuals of the one-parameter law are $+0.001$, -0.018, $+0.098$ and -0.059 across a deficit range of 0.214: a linear relationship with substantial scatter rather than a tight law, and the normalized deficit puts a mode count and a coverage fraction on one axis, so part of the residual is unit conversion. The reparameterization yields the correct value at $\rho = 1$ rather than a stronger correlation, and under it the 1.5B point, previously described as an exception to a monotone attenuation trend, falls within ordinary scatter.

D.2 The collapse process on the synthetic backbone

Under recursive self-training in the replace regime the count of producible tail modes falls in discrete steps, from 269 at the ancestor to 103 by generation forty, while the thirty head modes remain producible throughout. The loss is largely irreversible: a mode driven to deep extinction reappears in a later generation at a rate of 0.030, stable between 0.023 and 0.030 across the detection-threshold sweep. At $t^\star = 27$ the number of lost tail modes per lineage ranges from 63 to 135, and the pairwise complementarity of the lost-mode sets, measured as a symmetric-difference Jaccard index, ranges from 0.57 to 0.85 with mean 0.71. The linear-interpolation loss barrier between lineage endpoints is 0.000 and merged models retain validity 0.993, so the lineages share a basin and the weight-space results are not artifacts of incompatible endpoints.

D.3 The full operator panel

Table 3 gives the matched-control comparison for all seven merge operators on the synthetic backbone. Two corrections to previously circulated versions of this table are noted here for the record: the Fisher-weighted interval is $[-87.9, -55.7]$, not the narrower interval reported earlier, and the tuned-coefficient operator falls below the control in 21 of 28 pairs rather than in all of them. The tuned operator's mean and interval still exclude zero, so its conclusion is unchanged; only the per-pair claim was too strong.

Table 3: Every merge operator falls below the matched regression-to-ancestor control on the synthetic backbone. Values are producible modes of the merge minus the best matched control, over 28 lineage pairs, with dyadic cluster 95% intervals.

Merge operator	Merge − best control	95% CI	Pairs below control
Uniform average	-51.3	$[-62.9, -38.0]$	28/28
SLERP	-59.6	$[-72.5, -45.0]$	28/28
TIES	-69.9	$[-83.3, -56.5]$	28/28
DARE	-50.4	$[-60.8, -38.5]$	28/28
DARE-TIES	-157.1	$[-172.7, -138.1]$	28/28
Fisher-weighted	-71.7	$[-87.9, -55.7]$	28/28
Tuned coefficient	-2.9	$[-5.8, -0.7]$	21/28

The control is not a weak baseline. Interpolating a lineage toward an earlier checkpoint of itself, which requires no access to the ancestor, recovers 48.8 modes and also exceeds the merge's 45.9 on this run (but see the reproduction below); interpolating toward an exponential moving average of the trajectory recovers 25.1; and an oracle merge whose two-parent coefficient is chosen to maximize recovery reaches 46.9, still below the 97.2 of the matched ancestral control. The tuned-coefficient merge approaches the control only by selecting a coefficient near zero, which is itself reversion.

D.4 Reproduction on a re-trained backbone

The synthetic backbone was retrained in full from the unchanged frozen configuration on different hardware and a different software stack, and the seven-operator panel re-run on it. Every operator reproduces in sign and in the fraction of pairs below the control, with overlapping intervals and magnitudes about a tenth smaller across the board, consistent with the retrained lineages being slightly less collapsed at the same $t^\star$ (the ancestral control recovers 84.4 modes rather than 97.2). Uniform averaging -44.2 $[-57.2, -30.8]$ against -51.3 $[-62.6, -38.4]$, SLERP -54.5 against -59.6, TIES -56.8 against -69.9, DARE -43.9 against -50.4, DARE-TIES -142.7 against -157.1, Fisher -63.4 against -71.7, tuned -3.9 against -2.9; all seven below the control in $28/28$ pairs except the tuned operator at $24/28$ against $21/28$. The loss barrier is again 0.000 and merged validity 0.994.

One result does not reproduce. On the original run the rewind control, which interpolates a lineage toward an earlier checkpoint of itself and so needs no access to the ancestor, recovered 48.8 modes and exceeded the merge's 45.9. On the re-trained backbone it recovers 32.9 against the merge's 40.2, so it no longer exceeds the merge. The ancestral control, which is the one every claim in this paper rests on, reproduces robustly (84.4 against a merge of 40.2); the no-ancestor rewind variant does not, and should be treated as suggestive rather than established.

D.5 The replication grid: the endpoints not in the main text

Five endpoints were pre-registered for the grid and all five are reported. R1 (prevalence by exposure), R2 (the parameter-free sign and rank) and R5 (the decision rule) are in Section 7. The other two are here.

R3, the aggregate multi-task axis. Averaging the two targeted benchmarks of each pair gives 18 axes, of which 0.61 are exposed; the median fraction of the merge's gain over the better parent that the matched control also achieves is 0.79, so the case study's 0.82 is the middle of the grid rather than an outlier. This is the endpoint on which the case study's 82% was measured, and it is the one whose pre-registered prediction was refuted there: we predicted the canonical multi-task axis would be unexposed and it is not, because Proposition 8's criterion is about the aggregate, not per-task dominance.

R4, the plane search, and two sensitivities. The pre-registered objective is the worst benchmark normalised by its own best anchor, maximised over the simplex from the three anchor evaluations alone. It outperforms both competitors in 0.17 of 18 configurations. Two further analyses were added after this failure was observed and are labelled post hoc. The objective is a minimum over four ratios, so a benchmark on which every anchor scores near zero has a near-zero denominator and determines the objective by itself; restricting to benchmarks whose best anchor exceeds 0.05 and re-scoring the points the search had already selected, the plane point outperforms both in 0.28. On the mean normalised benchmark rather than the worst, an objective the search does not optimise, it outperforms both in the fraction recorded in the released artifact. We report the pre-registered value first and treat the search as a negative result.

D.6 Three parents, in full

The same experts admit a uniform three-way merge and its three matched controls at the $k = 3$ rate, 26 informative cells over 7 base models, nothing trained. The k-parent inequality remains predictive: exposure still separates, 0.90 of exposed cells against 0.56 of unexposed, and parameter-free sign agreement is 0.70 over 76 comparisons ($p = 3.8 \times 10^{-4}$). The *comparative* reading requires more care, and we report the sequence of analyses in full. Our pre-registered test, the $k = 3$ deficit against the mean of the three $k = 2$ deficits on the same experts, gives $+0.034$ with $k = 3$ better in 0.62 of cells, Wilcoxon $p = 0.048$. This meets the pre-registered criterion. With five of the seven base models it stood at $+0.040$ ($p = 0.07$), and it was at that point, after observing a positive but non-significant result, that we identified a confound in the comparison: the three-way merge always holds the benchmark's own expert while one of the three two-way pairs does not, so the average mixes configurations of different composition. This confound was not anticipated in the pre-registration, so we ran two post hoc controls. Averaging only over $k = 2$ pairs of matched composition gives $+0.001$ ($p = 0.635$). In addition, the theory's own prediction of the change, which controls composition by

construction because each configuration is predicted from its own anchors, has sign agreement 0.46 and Spearman -0.07 ($p = 0.721$), no better than chance.

We therefore report the pre-registered result but do not regard it as evidence: a nominally significant effect for which both controls return null results, and whose sign the theory's own parameter-free prediction cannot determine, does not support the claim. Equation (7) is an exact consequence of Assumption 1 and is not refuted by any measurement; what the data do not support is its practical interpretation, that adding parents makes a merge more competitive. The third analysis indicates why: adding a parent does not hold β and ρ fixed.

D.7 Sensitivity to the generation budget, and a withdrawn claim

The third-party arms were first measured at 256 new tokens, the budget frozen for every language-model evaluation in this paper. This budget is not neutral with respect to the comparison, and we report the following analysis in full because it weakens one of our own earlier claims.

The budget was already known to distort the Qwen arm, whose parents are long-chain-of-thought models, which is why that arm is reported in this appendix; the Llama arm, which the main text reports, had not been checked. Re-evaluating the entire arm at 1024 tokens, with all other settings unchanged, gives the comparison below.

budget	informative cells	exposed	control > merge	median absorbed
256 tokens	9	6	4/6	1.09
1024 tokens	9	7	4/7	0.69

The shorter budget truncated the *parents* rather than the merge. The mathematics fine-tune gains 0.543 on GSM8K and 0.133 on MATH-500 at the longer budget, and an underestimated parent produces an apparent ancestral surplus. At 256 tokens this arm reported a median absorbed fraction above one, from which we had initially concluded that the control captures the entirety of the merge's gain. At 1024 tokens the median is 0.69, the fraction at or above one falls from 0.67 to 0.17, and the merge outperforms the best matched control on MATH-500. We therefore withdraw the stronger statement. On these released checkpoints the control recovers approximately seven tenths of the merge's gain, which still warrants reporting it alongside a merge but does not render the merge redundant.

This is the only instance in which our own protocol inflated a merge's apparent gain over its parents, which is precisely the failure mode this paper examines. It also limits what the third-party evidence can support: this arm bears on magnitude but not on sign, and the prevalence and sign claims rest on the grid of Section 7, whose parents we fine-tune ourselves under a fixed protocol.

Finally, we note a reproducibility issue. One of the three released parents, `NeelRajani/Llama-3.1-8B-Instruct_SFT_mathv00.02_s45`, was publicly downloadable when the first measurement was made and now returns HTTP 401. The re-evaluation used the locally cached snapshot, checked by confirming its task-vector norm reproduces the original to five decimals. Third-party checkpoints are therefore not a stable basis for evaluation, which further argues for a control that requires only the base model and a single interpolation.

D.8 The third-party Qwen arm and a protocol limitation

The second third-party arm merges three released fine-tunes of Qwen2.5-7B-Instruct. Its results are reported here rather than in Section 7 for a reason that was apparent in the anchor evaluations before any merge was evaluated. The base scores GSM8K 0.827, CRUXEval 0.313 and MATH-500 0.210; 'OpenThinker-7B' scores 0.123, 0.003 and 0.000 and 'HuatuoGPT-o1-7B' 0.393, 0.133 and 0.010, while 'Qwen2.5-7B-CLIPPER' is comparable to the base at 0.823, 0.350 and 0.240. We do not consider the first two scores representative of those models: both are long-chain-of-thought models, and our frozen protocol allows 256 new tokens and requires the answer on a marked line, so a model that has not finished reasoning by token 256 scores zero regardless of its underlying capability. The exposure labels in this arm are therefore partly a property of our generation budget. We report the arm's endpoints, 5 of 9 cells exposed with the control beating the merge on 1.00 of them and a median absorbed fraction of 2.44, together with a bounded diagnostic declared before it was run: the base

and the two long-reasoning parents re-evaluated on a fixed 100-item GSM8K subsample at 1024 new tokens instead of 256, everything else unchanged. The diagnostic confirms the concern and bounds its size. The base goes from 0.770 to 0.890, 'OpenThinker' from 0.170 to 0.740 and 'HuatuoGPT-o1' from 0.380 to 0.860, so the base's lead falls from $+0.600$ to $+0.150$ over the first and from $+0.390$ to $+0.030$ over the second. The exposure label retains its sign for both, narrowly for the second, but most of its *magnitude* in this arm reflects the generation budget rather than the models. For this reason the arm is reported here and is not pooled into the main results. The magnitude claim in Section 7 rests on the Llama arm, whose three parents are standard supervised fine-tunes and whose results are unaffected by this limitation.

D.9 Language-model settings

Table 4 reports the four language-model settings. The pattern observed on the synthetic backbone recurs: the merge never exceeds the matched control, while retraining on pooled generations outperforms both a volume-matched and a diversity-matched single source, recovers modes the single source does not produce, and transmits them from one identifiable parent.

Table 4: The four language-model settings. Coverage is the fraction of tail modes produced; intervals are dyadic cluster 95% intervals; ρ is the double-precision task-vector cosine.

	Fact 0.5B	Arith 0.5B	Deeper-tail 0.5B	Fact 1.5B
Lineage pairs	28	6	28	6
Ancestor tail coverage	0.84	0.944	0.81	0.83
Merge does not exceed control	27/28	6/6	24/28	6/6
Crossover − volume solo	+0.090	+0.121	+0.075	+0.097
95% CI	[0.06, 0.12]	[0.09, 0.17]	[0.06, 0.09]	[0.03, 0.16]
Crossover − diversity solo	+0.110	+0.111	+0.077	+0.115
95% CI	[0.07, 0.15]	[0.06, 0.16]	[0.06, 0.09]	[0.08, 0.16]
Unique recovery (modes)	11.2	23.0	21.1	12.3
Task-vector cosine ρ	0.62	0.74	–	0.37

D.10 Data-space recombination on the synthetic backbone

Retraining a fresh model from the ancestor on the pooled generations of two lineages outperforms a volume-matched single source by 39.9 modes ($[27.8, 55.8]$) and a diversity-matched single source, which is sampled hot enough to match the pooled set's distinct-mode count at equal sample size, by 34.7 ($[10.3, 53.7]$). It recovers 34.3 modes ($[24.2, 47.0]$) that the diversity-matched source does not produce at all, and 0.90 of its recovered modes survived in exactly one parent against a chance rate of 0.67. Matching diversity on mode-distribution entropy instead of distinct-mode count gives the same conclusion ($+40.0$, $[28.7, 54.6]$; unique recovery 35.0, $[26.0, 48.0]$). A non-complementary pseudo-crossover that pools two generations of the same lineage recovers less ($+18.9$, $[4.0, 33.9]$).

D.11 Retention and positioning

Sweeping the data-retention window W from pure replacement to full accumulation, the crossover advantage falls monotonically from $+28.7$ modes to a non-significant $+2.5$, and the merge's deficit against the matched control shrinks in step from -51.2 to -3.4; both are what Proposition 5 predicts, since retention shrinks ΔA and ΔR together. Matched on data volume, the crossover attains tail coverage 0.725, which on the same axis equals injecting about twenty percent real data per generation. Pooling three lineages instead of two raises coverage from 0.744 to 0.811 and then saturates.

E Comparison with simpler decision rules

One might expect the confound to require no geometric analysis: once it is observed that the base model can outperform both parents, a simple heuristic might identify the same cases. To test this, we pre-registered six candidate decision rules before scoring any of them and evaluated each on the 72

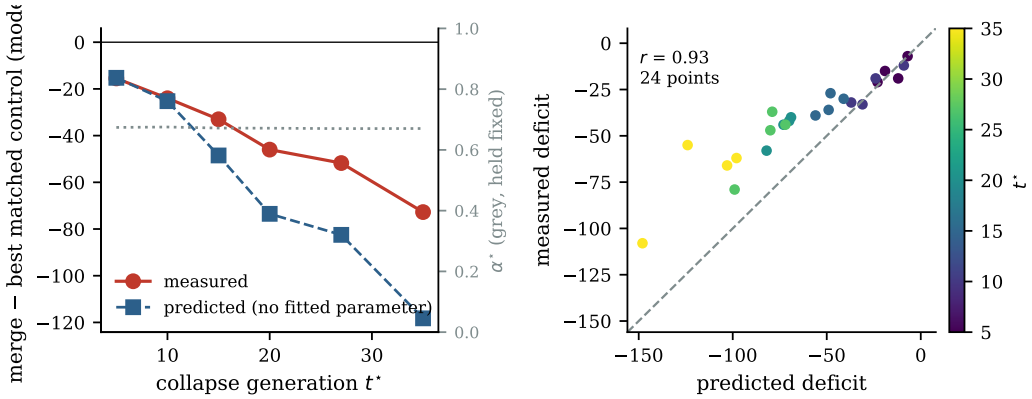


Figure 5: The other side of the master inequality, with the geometry held fixed. Sweeping the collapse generation changes the ancestral surplus while leaving $\alpha^\star$ between 0.666 and 0.679. Left: the measured deficit deepens with depth and the parameter-free prediction tracks it, drifting conservative as more modes reach zero measured frequency. Right: predicted against measured over all 24 points, coloured by depth.

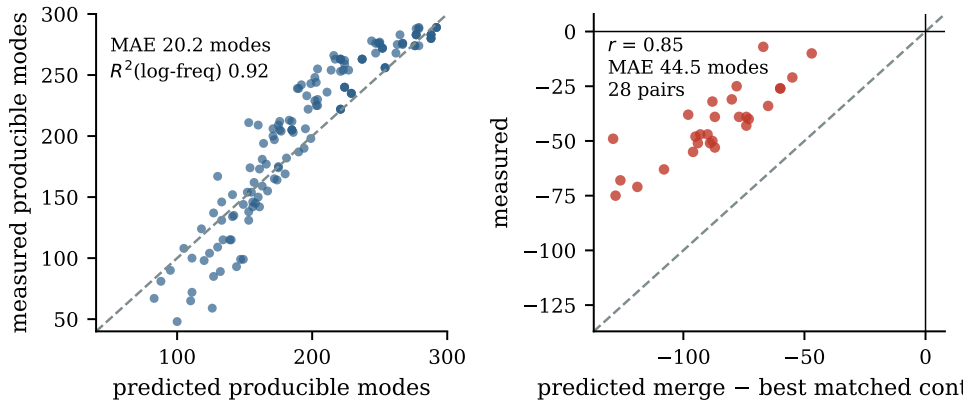


Figure 6: The first-order model is accurate enough to predict the endpoint. Left: predicted against measured producible-mode count at every held-out grid point of the merge plane. Right: the merge-minus-best-matched-control deficit predicted from the three anchor measurements and $\alpha^\star$ alone, against the measured deficit, over all 28 lineage pairs. No parameter is fitted in either panel.

grid cells under the decision convention of Section 7: each rule sees only the three anchor accuracies and the task-vector geometry, all available before either candidate is evaluated.

Two conclusions follow.

First, the heuristics implicit in the merging literature perform poorly. Orthogonality is the premise of interference reduction, yet selecting the merge when the task vectors are more orthogonal achieves 0.542, close to chance and significantly below our rule. Complementarity performs only slightly better, at 0.625. The common practice of always merging is the weakest of the six rules, at 0.319 accuracy and 0.1235 regret, an order of magnitude above ours. Current practice therefore does not make this decision reliably.

Second, the two rules that approach the master inequality are both derived from our analysis, and neither is statistically distinguishable from it at this sample size. The exposure test of Proposition 8 alone, which selects the merge if and only if the better parent outperforms the base, reaches 0.778 ($p = 0.146$); always reverting reaches 0.736 ($p = 0.078$). We therefore do not claim that the exchange rate is necessary for this decision. It improves on the coarser criterion by +0.083 in point estimate, a difference that 72 cells cannot establish as significant. What the geometry provides beyond these rules

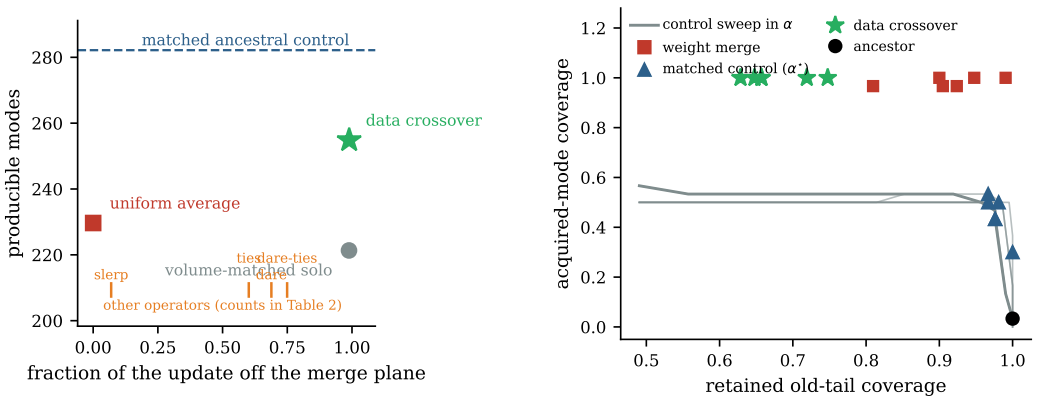


Figure 7: Left: what matters is the direction of departure from the merge plane, not the departure itself. Uniform averaging is in the plane to numerical precision; the sparsifying operators leave it and do worse; the data crossover leaves it almost entirely and does better than every merge, but under neutral drift still not better than reverting to the ancestor, which by Proposition 8 no method can exceed. Right: with acquisition the ancestor is uniquely best on the old tail and worst on everything the lineages learned, the merge sits on the Pareto frontier of the control sweep, and the crossover acquires both lineages' new modes in full.

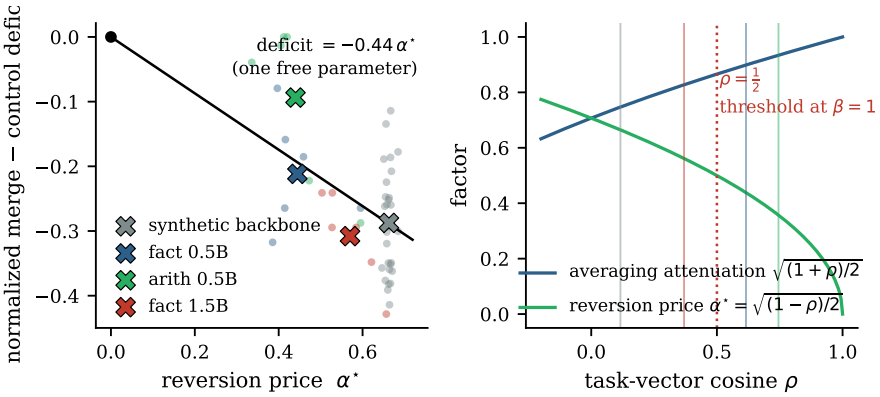


Figure 8: The exchange rate $\alpha^\star$ is the relevant variable. Left: the normalized merge-minus-control deficit against the exact $\alpha^\star$ of (3) over 46 lineage pairs in four regimes; small points are pairs, crosses are setting means, and the line is the one-parameter law through the origin that the theory requires. Right: the two conjugate factors as functions of ρ. The averaging attenuation used in the merging literature and the exchange rate available to the control are complementary; vertical lines mark the four measured regimes, and the dotted line marks the $\rho = 1/2$ threshold of Corollary 7 for a perfectly retaining sibling.

is a justification for the criterion, the threshold at which the confound must appear, and a certificate indicating when a prediction can be trusted.

F THE CANONICAL EIGHT-TASK VISION BENCHMARK

Section 7 reports that the criterion classifies this setting as unexposed; this appendix gives the measurement.

We fine-tuned all 8 CLIP ViT-B/32 image encoders ourselves under the protocol of Ilharco et al. (2023), with the zero-shot text head frozen so every model and every merge shares one head and one weight space. Before reading any endpoint we checked the reproduction against the values Ilharco et al. (2023) report for this backbone in their Table 7: zero-shot MNIST 0.484 against their 0.487,

Table 5: Rules for choosing between the uniform merge and the matched control, scored on all 72 grid cells. Accuracy is the fraction of cells on which the rule selects the candidate that scores higher; regret is the mean accuracy lost relative to an oracle. The p-values are exact McNemar tests against the master inequality.

rule, using only pre-evaluation quantities	accuracy	regret	vs. ours
master inequality (4), ours	0.861	0.0088	
exposure test of Proposition 8, ours	0.778	0.0211	0.146
always revert to the ancestor	0.736	0.0197	0.078
select the merge when the parents differ more	0.625	0.0691	0.003
select the merge when the task vectors are more orthogonal	0.542	0.0747	6.6×10^{-5}
always merge (common practice)	0.319	0.1235	2.8×10^{-9}

SVHN 0.316 against 0.315, EuroSAT 0.449 against 0.459, DTD 0.444 against 0.441, GTSRB 0.325 against 0.325 and RESISC45 0.597 against 0.607; fine-tuned EuroSAT 0.989 against 0.986, GTSRB 0.986 against 0.985, MNIST 0.996 against 0.996, SVHN 0.972 against 0.972, Cars 0.761 against 0.792 and SUN397 0.748 against 0.751. The mean pairwise task-vector cosine is 0.079.

Table 6: The eight-task benchmark. "own expert" is the model fine-tuned on that task, "other experts" the mean of the seven fine-tuned on the other tasks. Forgetting is present, with the other experts averaging 0.436 against the zero-shot model's 0.480, but each expert's gain on its own task is large enough that a single expert's eight-task *aggregate* still exceeds zero-shot, which is why the ancestral surplus is negative here.

task	zero-shot	own expert	other experts	uniform merge
Cars	0.596	0.761	0.516	0.628
DTD	0.444	0.770	0.385	0.499
EuroSAT	0.449	0.989	0.361	0.696
GTSRB	0.325	0.986	0.302	0.507
MNIST	0.484	0.996	0.496	0.847
RESISC45	0.597	0.947	0.481	0.703
SUN397	0.631	0.748	0.573	0.652
SVHN	0.316	0.972	0.374	0.627
average	0.480	0.896	0.436	0.645

The quantity that the published normalisation conceals is a single parent's score on the aggregate axis. The best is the SVHN expert at 0.553, above the zero-shot model's 0.480, giving an ancestral surplus of -0.073. By Proposition 8 the comparison is therefore not degenerate and the confound cannot arise, and the measurement agrees: the uniform merge reaches 0.645 while the best matched control reaches only 0.520, so the merge outperforms it by +0.125 and the absorbed fraction is -0.369; the control does not even reach the best parent. No individual task is exposed.

This result shows that the criterion discriminates between settings. A natural objection to a control of this kind is that it would disfavor every merge; it does not. On 33 of the 72 constructed-expert cells the ancestral surplus is positive and the control outperforms the merge on 0.97 of them, whereas on the setting on which this literature is built the surplus is negative and the merge outperforms the control. In particular, the most-cited merging result in the field is not explained by reversion to the pretrained model.

G What published merging evaluations report

Our exposure test requires two quantities on the metric a paper reports: the ancestor's score, and one parent's score on that same aggregate. We examined the three most-cited weight-merging evaluations directly. Ilharco et al. (2023), Table 3, report both for their GLUE task-vector addition: zero-shot average 57.1, fine-tuned 78.1, fine-tuned plus task vectors 78.6. The ancestral surplus is strongly negative, so by Proposition 8 that evaluation is not exposed and our control could not have absorbed its gain; we report it as a case the criterion classifies as unexposed, not as an instance of the confound.

Their eight-task image results are reported as accuracy normalised by each task's own fine-tuned model and then averaged, which makes any single parent's score on the aggregate axis unrecoverable from the paper; we therefore reproduced that setting ourselves and recovered the quantity, and it is likewise unexposed (Appendix F). Yadav et al. (2023), Table 1, report rows for the individual fine-tuned models, a multitask model and five merging methods, with no pretrained or zero-shot row; the individual row is each model on its own task, not one model on the aggregate, so neither number the test needs is present. Yu et al. (2024), Table 1, merge three Llama-2-13B fine-tunes and report two of the three parents only on their own tasks and each merge on all tasks, and their Table 2 compares merged models against the source models they were built from on a six-benchmark average; neither has a row for the pretrained backbone.

We recompute none of their results and make no claim about their conclusions. We note only what the tables show: the control requires one interpolation and two evaluations, and in two of these three cases it cannot be applied from the published results because the required quantities are not reported. Both cases that could be evaluated, the one that reports both quantities and the one we reproduced, are unexposed. This supports the scope claim, and it is why we describe the confound as a property of the reported metric rather than of merging as such.