WHEN ARE NEARBY EXPERTS COMPOSABLE? CROSS-TASK MODEL MERGING IN NEURAL THICKETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Random perturbations around a shared checkpoint can produce experts for distinct tasks, but shared origin does not ensure that their gains survive composition. We formulate expert-pair selection for joint base-relative retention, measured by strict joint success and worst-task gain with unconditional abstention accounting. An exact finite-difference decomposition and an arbitrarily small orthogonal counterexample show why proximity and orthogonality alone are insufficient. We instantiate this formulation with quality-constrained spectral selection, a fixed-budget framework combining cross-task screening, layerwise pair geometry, separate validation, and Base fallback. Across six checkpoints, 21 task pairs, and five independent perturbation pools, the procedure improves strict joint success over Top-1 averaging throughout and leads the four-method model-scale comparison on four of six checkpoints. For the five models with at least 1.5B parameters, matched quality-only, global-cosine, and spectral ranking attain 49.1%, 50.5%, and 53.1% joint success. Under this matched protocol, the spectral selector also yields the highest worst-task gain and Weak Pareto rate and the lowest regression rate. Mean worst-task gain reaches $+0.61$ percentage points, while 17.1% regression and 13.0% abstention show that composability remains conditional. Project page: https://anonymous.4open.science/r/Thickets-Merge/.

1 INTRODUCTION

A useful merge must retain multiple capabilities: improving programming while degrading mathematics can raise an average score without achieving joint retention. Methods include geometric averaging in Model Stock (Jang et al., 2025), task arithmetic in MetaGPT (Zhou et al., 2024), sparse updates in DARE and Model Breadcrumbs (Yu et al., 2024; Davari & Belilovsky, 2025), adaptive coefficients in AdaMerging (Yang et al., 2024), and task-specific masks and rescalers in EMR-Merging (Huang et al., 2024). MergeBench examines retention tradeoffs in domain-specialized LLMs (He et al., 2025), motivating taskwise assessment of whether one merged checkpoint retains the distinct capabilities supplied by its constituent experts.

Random experts make this question concrete. Neural Thickets identifies task-improving models near a pretrained checkpoint through RandOpt's sample–select–ensemble procedure (Gan & Isola, 2026b). SubZero and LOZO instead use structured random perturbations to estimate iterative optimization updates (Yu et al., 2025; Chen et al., 2025); Iterative RandOpt adds repeated sampling and distillation (Gan & Isola, 2026a). We study composition of separately selected task experts from a shared candidate population. A shared checkpoint provides a common parameterization for combining updates, but does not establish that their gains survive addition.

Geometric structure offers useful signals. Task Singular Vectors addresses task-subspace interference (Gargiulo et al., 2025), KnOTS aligns low-rank updates (Stoica et al., 2025), isotropic merging separates common and task-specific subspaces (Marczak et al., 2025), and Essential Subspace Merging connects update selection to output shifts (Li et al., 2026). Functional evaluation remains essential: individually beneficial updates can interact adversely or damage another task. We study *joint retention under a specified expert-selection and merging procedure*.

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.

The reference and objective matter. Pareto-oriented coefficient search examines tradeoffs among capabilities (Khalifa et al., 2024); beating selected specialists or increasing an average need not improve both tasks over Base. We distinguish strict joint success, Weak Pareto improvement, and worst-task gain, counting abstentions in the unconditional evaluation. An exact finite-difference identity separates individual effects from interactions, and an arbitrarily small orthogonal counterexample establishes a geometric non-guarantee. Composability is therefore evaluated for a specified selection and merging procedure.

We instantiate this formulation with quality-constrained spectral selection, a fixed-budget expert-pair selection framework: cross-task scores screen quality, layerwise subspace overlap ranks pairs, and separate validation calibrates merges or returns Base. Subspace-guided expert selection (Skorobogat et al., 2025), rewarded-perturbation consolidation (Zhang et al., 2026), and evolutionary merging recipes (Akiba et al., 2025) provide complementary precedents. Matched quality-only and global-cosine controls assess pair ranking while keeping candidate screening, coefficient calibration, validation budget, and decision rule fixed throughout the comparison.

We evaluate six instruction-tuned Qwen2.5 (Qwen et al., 2025), Llama 3.1 (Grattafiori et al., 2024), and OLMo 3 (Team Olmo et al., 2025) checkpoints across 21 task pairs and five independent pools. Ours exceeds Top-1 throughout and leads the four-method comparison on four checkpoints; Direct Search leads Qwen 7B and TIES leads OLMo 3. For the five $\geq$1.5B models, strict joint success is 53.1% (quality-only: 49.1%; cosine: 50.5%), and mean worst-task gain is $+0.61$ pp.

Our central contribution is joint-retention pair selection for local perturbation experts: deciding which independently selected experts around a shared checkpoint can be composed to improve both tasks over Base.

- **Problem formulation.** We formulate cross-task composability as joint base-relative retention rather than average improvement or specialist-relative performance, with unconditional accounting of abstentions.
- **Failure characterization.** We characterize why local proximity and simple orthogonality are insufficient through an exact interaction decomposition, an arbitrarily small orthogonal counterexample, and radius identities.
- **Decision procedure.** We introduce bounded expert-pair selection combining functional screening, layerwise subspace ranking, coefficient calibration, and validation with abstention. Evaluation spans six checkpoints, seven tasks, 21 pairs, and five independent perturbation pools, with matched controls and unconditional outcome accounting.

2 Related Work

Random search and expert consolidation. Evolution strategies, MeZO, and Black-Box Tuning use random perturbations for optimization (Salimans et al., 2017; Malladi et al., 2023; Sun et al., 2022). We instead compose separately selected perturbation experts. RandOpt ensembles local experts; Iterative RandOpt distills them (Gan & Isola, 2026b;a). CoRP combines reward/compatibility weighting, validation, and population-informed search (Zhang et al., 2026); evolutionary and Pareto methods search checkpoint combinations (Akiba et al., 2025; Khalifa et al., 2024). Our target is joint base-relative retention across tasks.

Weight merging and merge geometry. Checkpoint averaging (Wortsman et al., 2022), task arithmetic (Ilharco et al., 2023), and importance/statistics-based rules (Matena & Raffel, 2022; Jin et al., 2023) combine given updates. Comparing random and fine-tuned experts additionally requires matched quality, displacement, and cost. Connectivity and transfer motivate shared-base composition (Frankle et al., 2020; Neyshabur et al., 2020); Git Re-Basin and tangent-space arithmetic address alignment and functional disentanglement (Ainsworth et al., 2023; Ortiz-Jimenez et al., 2023). Task-vector and subspace methods provide geometric signals (Gargiulo et al., 2025; Li et al., 2026).

Mergeability and merge selection. *Will it Merge?* quantifies fine-tuned-model mergeability and links it to base knowledge (Rahamim et al., 2026). *Demystifying Mergeability* studies method-

and partner-dependent predictors for fine-tuned CLIP pairs, including left/right task-vector singular-subspace overlap (Zhou et al., 2026). SimMerge uses functional/structural similarity to select checkpoint subsets, operators, and order (Bolton et al., 2026). CABS+ combines conflict-aware sparsification and adaptive weights; its Relative Synergy Score guides model selection (Liu et al., 2026). Zhou et al. (2026) report positive predictive associations for leading right-subspace alignment; our ranking favors lower overlap after functional screening. We therefore interpret overlap direction as protocol-dependent. Prediction assesses a given merge; our decision jointly specifies separately selected local perturbation experts from one shared population, strict improvement on both tasks over Base, and pair selection under bounded validation with explicit abstention.

3 From Expert Quality to Composability

Candidates and selected experts. Let $\theta_0 \in \mathbb{R}^d$ be a shared checkpoint and $\theta_i = \theta_0 + \delta_i$ a sampled candidate. For task t, $S_t(\theta; D)$ is a higher-is-better score on data D. Selection data D_t^{sel} determine a task-specific pool $\mathcal{E}_t$ and Top-1 index e_t; separate validation data calibrate merges, and held-out data determine final gains. The shared checkpoint may already be instruction tuned. Candidates are sampled around the shared checkpoint and subsequently identified by their resulting parameter updates and task scores; selection and merging operate on these realized candidates rather than on seed identifiers alone.

Merging and reference models. For tasks A, B, arithmetic merging gives

$$\theta_{ij}(\alpha, \beta) = \theta_0 + \alpha\delta_i + \beta\delta_j, \qquad i \in \mathcal{E}_A,\ j \in \mathcal{E}_B. \tag{1}$$

Top-1 averaging uses $\alpha = \beta = 1/2$; calibration selects coefficients on validation data. Deployment uses one checkpoint and one response per example. On held-out data, define

$$g_t(\theta) = S_t(\theta) - S_t(\theta_0), \qquad w(\theta) = \min\{g_A(\theta), g_B(\theta)\}. \tag{2}$$

For accuracy, gains are percentage points (pp). Strict joint success (SJ) requires $w > 0$. Weak Pareto improvement requires $w \geq 0$ and $\max(g_A, g_B) > 0$; a two-task tie does not qualify. Regression means $w < 0$. Separately, *specialist dominance* compares each task with its selected expert rather than Base, and can hold even when a task remains below Base.

Unconditional reporting. SJ is the percentage of successful pair–pool evaluations. Mean worst-task gain (WG) averages w; mean gain averages $(g_A + g_B)/2$. If no merge passes validation, the procedure returns Base and records abstention. Such outcomes remain in every denominator and contribute zero to gain averages, but are neither joint successes, Weak Pareto improvements, nor regressions. Validation acceptance and held-out success are distinct events.

Conditional composability. For a fixed candidate/selection distribution Q, merge rule M, and margins $m_t \geq 0$, define

$$\kappa(Q, M, \mathbf{m}) = \Pr_{z \sim Q}\left[\forall t,\ g_t\big(M(z)\big) > m_t\right]. \tag{3}$$

Here, z denotes a sampled selection instance under Q. Our finite evaluation measures protocol-specific frequency over selected tasks and pools. It does not estimate the volume density of successful models in weight space or establish that a different pairing rule has the same success rate.

4 Why Nearby Experts Can Conflict

Merging changes direction and radius. For $u = \alpha\delta_i$ and $v = \beta\delta_j$,

$$\|u + v\|^2 = \alpha^2 \|\delta_i\|^2 + \beta^2 \|\delta_j\|^2 + 2\alpha\beta\langle\delta_i, \delta_j\rangle. \tag{4}$$

Averaging two orthogonal updates of norm r gives radius $r/\sqrt{2}$, whereas adding them gives $\sqrt{2}r$. Fixed coefficients therefore do not isolate directional compatibility. Global norm matching also does not equalize layerwise sensitivity.

Orthogonality does not remove cross-task effects. Near-zero cosine in high dimensions does not imply that two updates preserve each other's capabilities. For normalized vectors with cosine c, the two-vector Gram eigenvalues are $1 \pm |c|$: this spectrum re-expresses absolute cosine. Matrix-level singular subspaces carry different information, but their predictive value still needs controlled evaluation.

An exact decomposition separates effects from interactions. For a smooth diagnostic loss, let $\Delta L_t(u) = L_t(\theta_0 + u) - L_t(\theta_0)$ and

$$I_t(u, v) = L_t(\theta_0 + u + v) - L_t(\theta_0 + u) - L_t(\theta_0 + v) + L_t(\theta_0). \tag{5}$$

Then $\Delta L_t(u+v) = \Delta L_t(u) + \Delta L_t(v) + I_t(u, v)$. Even $I_t = 0$ cannot ensure improvement if an update's cross-task harm outweighs the other's benefit. Appendix A derives the Hessian-integral form and an arbitrarily small orthogonal counterexample. These are statements about smooth losses, not guarantees for discrete LLM accuracy. The analysis motivates cross-task screening and validation of actual merges, without deriving the spectral score used to prioritize pairs.

5 Selecting and Merging Local Experts

Given local candidates for two tasks, the decision is which pair to compose under a fixed evaluation budget. Layerwise overlap is a ranking signal within functional screening and validation, rather than a standalone definition of mergeability. We instantiate this joint-retention framework with *quality-constrained spectral selection*: functional screening, geometric pair prioritization, and bounded validation with fallback (Figure 1; Algorithm 1, Appendix B.4). Both experts are scored on both tasks; selection, validation, and testing use disjoint data. Building on task-vector and subspace merging (Ilharco et al., 2023; Gargiulo et al., 2025; Skorobogat et al., 2025), the procedure uses no gradient updates.

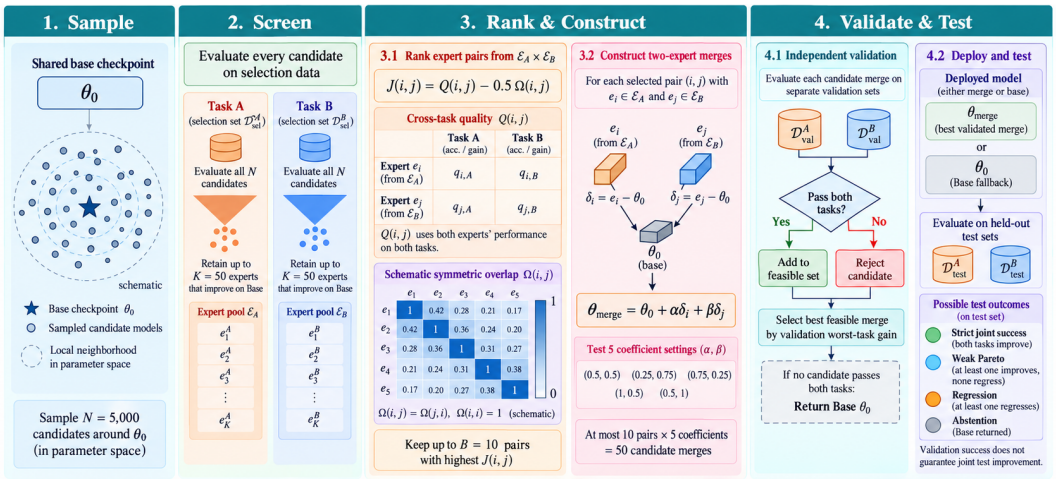


Figure 1: **From a shared checkpoint to a validated two-expert merge.** Screen candidates, rank distinct expert pairs, and validate at most $10 \times 5 = 50$ merges. Screening permits ties with Base; validation requires both gains > 0. Quality entries denote base-relative selection gains. Neighborhood and overlap are schematic; weak Pareto includes strict joint success. The best feasible merge or Base proceeds to held-out testing, without guaranteed improvement.

Stage 1: screen task-conditioned pools. Score every candidate on every target task using D_t^{sel}, obtaining base-relative gains $g_t^{\text{sel}}(\theta_i)$. Retain up to $K = 50$ candidates per task satisfying $g_t^{\text{sel}}(\theta_i) \geq \eta_t$, with $\eta_t = 0$ pp. Require $i \neq j$ for a two-expert merge. Direct Search is the single-candidate multitask baseline: among candidates scored on both tasks, it selects one realized candidate rather than composing two distinct experts. For two-expert selection, an empty feasible pair set triggers base fallback, recorded as abstention. Screening uses no final-test information and does not establish generalization.

Stage 2: rank pairs by quality and subspace overlap. For an eligible two-dimensional parameter tensor ℓ, write the rank-r_ℓ singular subspaces of its update as

$$\Delta W_{i,\ell} \approx U_{i,\ell}\Sigma_{i,\ell}V_{i,\ell}^\top, \qquad \omega_\ell(i,j) = \frac{\left\|U_{i,\ell}^\top U_{j,\ell}\right\|_F^2 + \left\|V_{i,\ell}^\top V_{j,\ell}\right\|_F^2}{2r_\ell}. \tag{6}$$

We use orthonormal columns and a common positive rank at each tensor, with the rank capped at eight and the weighting scheme fixed across all primary runs. Then $0 \le \omega_\ell \le 1$. Aggregate as $\Omega(i,j) = \sum_\ell w_\ell \omega_\ell(i,j)$, where $w_\ell \ge 0$ and $\sum_\ell w_\ell = 1$. We allocate half the total weight to attention matrices and half to MLP matrices, uniformly within each family, after excluding ineligible tensors. Non-matrix parameters remain in the merge but do not contribute to the overlap score. Sign-invariant subspace overlap discards singular-value magnitudes and does not measure task loss.

Rank pairs using

$$J(i,j) = \min_{t\in\{A,B\}} \frac{g_t^{\mathrm{sel}}(\theta_i) + g_t^{\mathrm{sel}}(\theta_j)}{2} - \lambda\Omega(i,j), \qquad \lambda \ge 0. \tag{7}$$

The first term penalizes observed cross-task harm without assuming accuracy is linear in weights. The second favors lower subspace overlap at comparable score quality. We use $\lambda = 0.5$ across all models, pairs, and pools. Setting $\lambda = 0$ gives quality-only ranking; Global Cosine substitutes absolute full-update cosine for Ω, with all other settings fixed (Table 2).

Stage 3: calibrate within a fixed budget. Evaluate up to B top-ranked distinct pairs on $\mathcal{A} = \{(0.5, 0.5), (0.25, 0.75), (0.75, 0.25), (1, 0.5), (0.5, 1)\}$. Unlike AdaMerging's unsupervised coefficient optimization (Yang et al., 2024), this gate uses separate labeled validation data. A feasible merge must satisfy $g_A^{\mathrm{val}} > 0$ and $g_B^{\mathrm{val}} > 0$. Conditional composability permits abstention: select the largest worst-task gain, breaking ties by mean gain and fixed candidate/coefficient order, or return Base if no proposal is feasible. We use $K = 50$, $B = 10$, five coefficient proposals per pair, and a maximum of 50 validated merges per task pair (Appendix B).

Cost and scope. Screening costs NT task evaluations; calibration adds at most $2B|\mathcal{A}|$ per task pair. Across seven tasks, the default uses 35,000 screening and at most 2,100 calibration evaluations per pool, excluding base evaluation, SVD, and reconstruction overhead. These operation counts do not establish matched wall-clock cost. The deployed model has fixed weights, unlike input-dependent WEMoE (Tang et al., 2024); TIES is a separate interference-handling comparator (Yadav et al., 2023). Composing more than two experts requires joint evaluation: pairwise compatibility does not imply compatibility of their sum.

6 Experiments

We address four research questions:

RQ1: How does joint improvement vary across models and scales?
RQ2: How does the full method compare with alternative search and merge rules?
RQ3: Does spectral pair ranking add value under a matched pipeline?
RQ4: How much does performance vary across independent perturbation pools?

6.1 Evaluation setup

We use Qwen2.5-Instruct at 0.5B, 1.5B, 3B, and 7B, Llama-3.1-8B-Instruct, and OLMo 3-7B-Instruct (Qwen et al., 2025; Grattafiori et al., 2024; Team Olmo et al., 2025). The seven tasks are Countdown, GSM8K, MATH-500, OlympiadBench, MBPP, ROCStories, and USPTO, covering all 21 unordered pairs (Gan & Isola, 2026b; Cobbe et al., 2021; He et al., 2024; Austin et al., 2021; Mostafazadeh et al., 2016). Each model has five independently sampled pools, giving 105 pair–pool evaluations. Table 2 reports the five-model aggregate with at least 1.5B parameters (525 evaluations); Table 1 includes all six checkpoints, with 0.5B as a separate stress test.

We sample $N = 5{,}000$ candidates per pool, retain up to $K = 50$ per task, and validate up to ten pairs with five coefficients each. Selection and validation use disjoint sets of 200 examples per task;

Table 1: **Cross-task composability across model scales.** Each model covers 21 task pairs and five independent pools (105 outcomes). SJ: strict joint success (%); WG: mean worst-task gain (pp). Bold marks metric-wise row bests. Small arrow values give Ours minus the best of Top-1, TIES, and Direct for each model and metric (pp), computed from displayed entries.

	SJ ↑ (%)				WG ↑ (pp)			
Model	Top-1	TIES	Direct	Ours	Top-1	TIES	Direct	Ours
Qwen2.5-0.5B	11.4	23.8	21.9	**25.7**$_{\uparrow 1.9}$	-1.02	-0.28	-0.12	**-0.06**$_{\uparrow 0.06}$
Qwen2.5-1.5B	18.1	33.3	38.1	**42.9**$_{\uparrow 4.8}$	-0.71	+0.15	+0.23	**+0.42**$_{\uparrow 0.19}$
Qwen2.5-3B	26.7	44.8	49.5	**55.2**$_{\uparrow 5.7}$	-0.34	+0.31	+0.43	**+0.69**$_{\uparrow 0.26}$
Qwen2.5-7B	33.3	52.4	**61.0**	59.0$_{\downarrow 2.0}$	-0.12	+0.48	**+0.77**	+0.72$_{\downarrow 0.05}$
Llama-3.1-8B	28.6	47.6	50.5	**57.1**$_{\uparrow 6.6}$	-0.22	+0.36	+0.44	**+0.65**$_{\uparrow 0.21}$
OLMo 3-7B	27.6	**53.3**	49.5	51.4$_{\downarrow 1.9}$	-0.26	**+0.61**	+0.46	+0.58$_{\downarrow 0.03}$

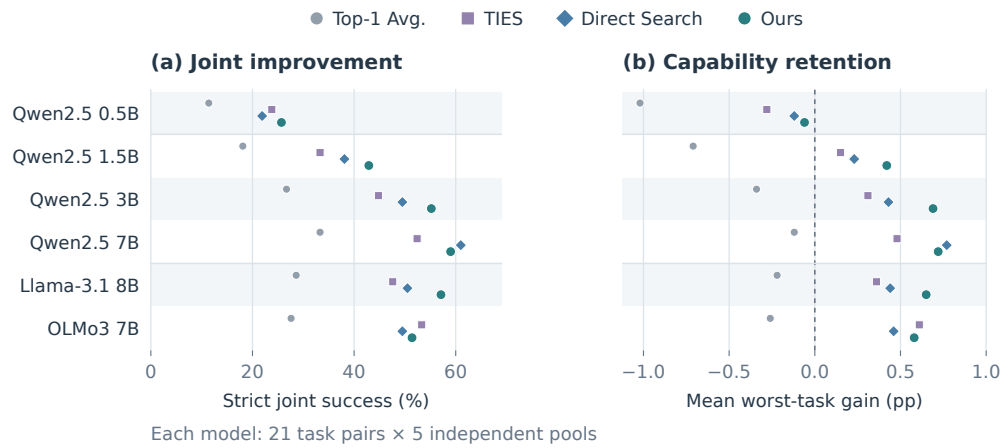


Figure 2: **More joint successes do not guarantee positive worst-task gain.** Model-wise SJ and WG for four methods. Each row summarizes 21 task pairs and five pools. Our method improves on Top-1 throughout, but Direct Search leads Qwen 7B and TIES leads OLMo 3. No uncertainty intervals are inferred from these model-level aggregates.

final assessment uses held-out test data. MATH-500 remains held out, with selection and validation drawn from MATH training data. We fix $\lambda = 0.5$, screening threshold $\eta_t = 0$, and a strictly positive validation-gain gate across all models. Inference uses BF16, greedy decoding, and one response per item. Appendix B summarizes the primary evaluation configuration and aggregation rules.

6.2 Model-scale composability (RQ1)

The full method leads four of six models, with meaningful exceptions. Our SJ exceeds Top-1 across all six checkpoints (Table 1). Within Qwen, it increases $25.7 \rightarrow 42.9 \rightarrow 55.2 \rightarrow 59.0\%$ with model size. Direct Search nevertheless leads at 7B ($61.0 > 59.0$), and TIES leads OLMo 3 ($53.3 > 51.4$). At 0.5B, our WG remains -0.06 pp despite improved SJ. Thus success frequency, average worst-task retention, and the best comparison method need separate assessment (Figure 2).

6.3 Primary comparison and abstention (RQ2)

We separate system-level and component-level evidence. Top-1, TIES, and Direct Search compare complete merge/search procedures, whereas Quality-only, Global Cosine, and Ours hold candidate pools, screening, coefficient proposals, validation budget, and the decision gate fixed, isolating the pair-ranking criterion.

Table 2: **Aggregate comparison for the five ≥1.5B models.** 525 pair–pool evaluations per method; all metrics include abstentions. Rates are in %; gains are in pp. Bold marks columnwise optima. Small arrow values, omitted for Abstain, give Ours minus the best comparator per metric (pp), from displayed entries; green/red denotes better/worse.

Method	SJ ↑	WG ↑	Weak Pareto ↑	Mean gain ↑	Regression ↓	Abstain ↓
Top-1 Avg.	26.9	-0.33	42.1	+0.36	57.9	**0.0**
TIES	46.3	+0.38	61.3	+1.09	38.7	**0.0**
Direct Search	49.7	+0.47	64.6	+1.28	35.4	**0.0**
Quality-only, $\lambda = 0$	49.1	+0.49	66.1	+1.39	20.4	13.5
Global Cosine	50.5	+0.54	67.6	+1.48	19.0	13.3
Ours	**53.1**$_{\uparrow 2.6}$	**+0.61**$_{\uparrow 0.07}$	**69.9**$_{\uparrow 2.3}$	**+1.57**$_{\uparrow 0.09}$	**17.1**$_{\downarrow 1.9}$	13.0

Quality-only and Global Cosine retain the full screening, validation budget, and gate, changing only pair ranking. Weak Pareto includes SJ; returning Base is neither improvement nor regression.

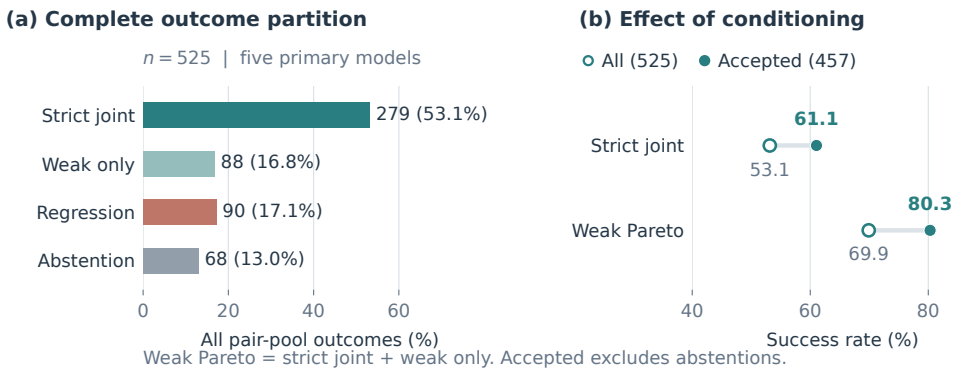


Figure 3: **Abstention is part of the deployed outcome.** All 525 primary evaluations partition into 279 strict successes, 88 weak-only improvements, 90 regressions, and 68 Base fallbacks. Weak Pareto includes the first two categories. Any conditional rates exclude the 68 fallbacks and use denominator 457; primary comparisons always retain denominator 525.

Joint retention improves beyond positive average gain. Our SJ is 53.1%, versus 26.9%, 46.3%, and 49.7% for Top-1, TIES, and Direct Search (Table 2). All six methods have positive mean gain, yet Top-1's WG is −0.33 pp: averages can conceal cross-task damage.

Abstention returns Base, while accepted merges can still regress. Weak Pareto covers $367/525 = 69.9\%$ of evaluations; $90/525 = 17.1\%$ regress and $68/525 = 13.0\%$ abstain (Figure 3). Abstentions contribute zero to WG and mean gain, which are +0.61 and +1.57 pp. Among the 457 non-abstaining evaluations, SJ is 61.1%; this conditional rate is not substituted for the unconditional 53.1%. Positive validation gains therefore do not eliminate held-out regressions.

6.4 Matched pair-ranking controls (RQ3)

Functional quality, global geometry, and layerwise subspaces. Quality-only ranks functional evidence; Global Cosine supplements it with global vector geometry, Ours with layerwise singular-subspace geometry. Other pipeline settings remain fixed. SJ increases $49.1 < 50.5 < 53.1\%$, WG increases $0.49 < 0.54 < 0.61$ pp, Weak Pareto rises $66.1 \to 67.6 \to 69.9\%$, and mean gain rises $1.39 \to 1.48 \to 1.57$ pp; regression decreases $20.4 > 19.0 > 17.1\%$ and abstention decreases $13.5 \to 13.3 \to 13.0\%$. Directionally consistent aggregate retention and deployment metrics support incremental utility of layerwise ranking within this matched progression, without implying statistical significance or a causal mechanism.

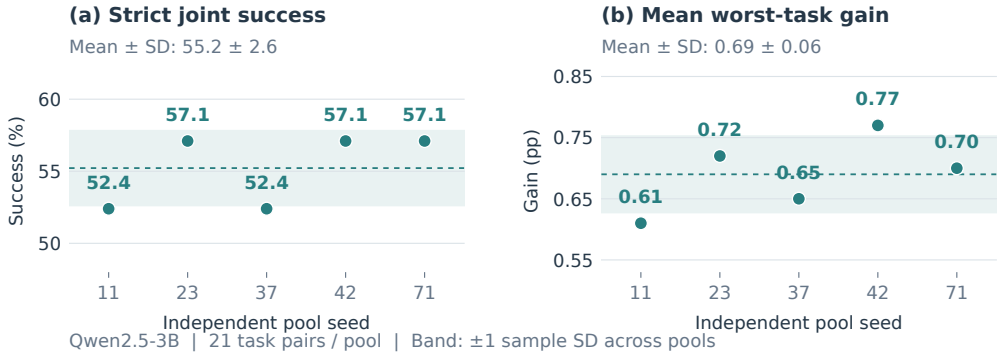


Figure 4: **Joint gains persist across five Qwen2.5-3B pools.** Each pool evaluates the same 21 task pairs under the fixed protocol. Points show the individual pools; summaries are mean ± sample SD, not confidence intervals. Pool-level variation addresses candidate-search randomness and does not quantify all test-sample uncertainty.

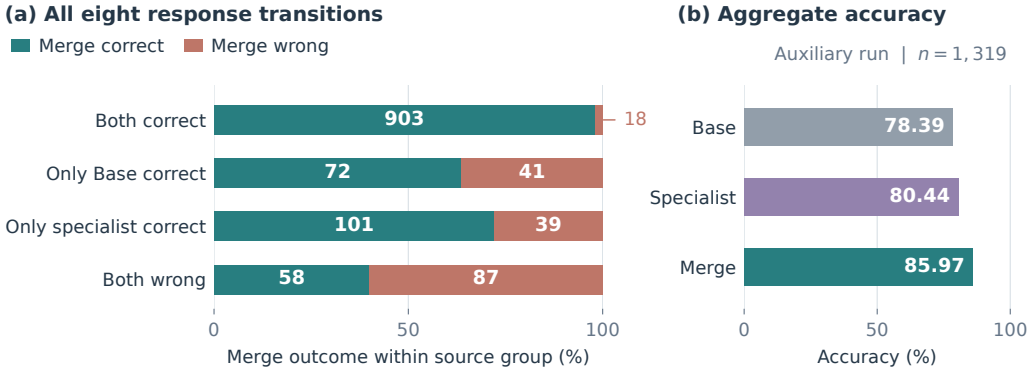


Figure 5: **Corrections and losses in an auxiliary coefficient analysis.** Qwen2.5-3B, GSM8K–MBPP, pool 42; all 1,319 GSM8K test items. The selected pair is additionally evaluated at $(\alpha_{\text{GSM}}, \beta_{\text{MBPP}}) = (0.75, 0.50)$, outside the five-point primary grid. This run contributes to neither main table. Specialist denotes the GSM8K expert participating in this merge.

Improved retention is not explained by more fallback. With the same pools and at most 50 proposals per task pair, overlap changes which combinations reach validation. Appendix D adds matched configuration-level and one-factor-at-a-time robustness analyses.

6.5 Independent-pool variation (RQ4)

For seeds 11, 23, 37, 42, and 71, Qwen 3B SJ ranges from 52.4% to 57.1%, averaging $55.2 \pm 2.6\%$; WG ranges from $+0.61$ to $+0.77$ pp, averaging $+0.69 \pm 0.06$ pp (Figure 4). All five pools have positive WG. These pool-level summaries describe search variability rather than inferential uncertainty across ranking methods; Appendix B.2 discusses the dependence structure of the reported evaluations.

6.6 Auxiliary response analysis

Corrections outweigh losses in this auxiliary run. Base, the participating GSM8K specialist, and Merge answer 1,034, 1,061, and 1,134 items correctly (78.39%, 80.44%, 85.97%). Merge corrects 159 Base errors and loses 59 Base-correct answers, giving $+100/1{,}319 = 7.58$ pp. It corrects 58 errors shared by the two displayed references, but also loses 18 answers both answered

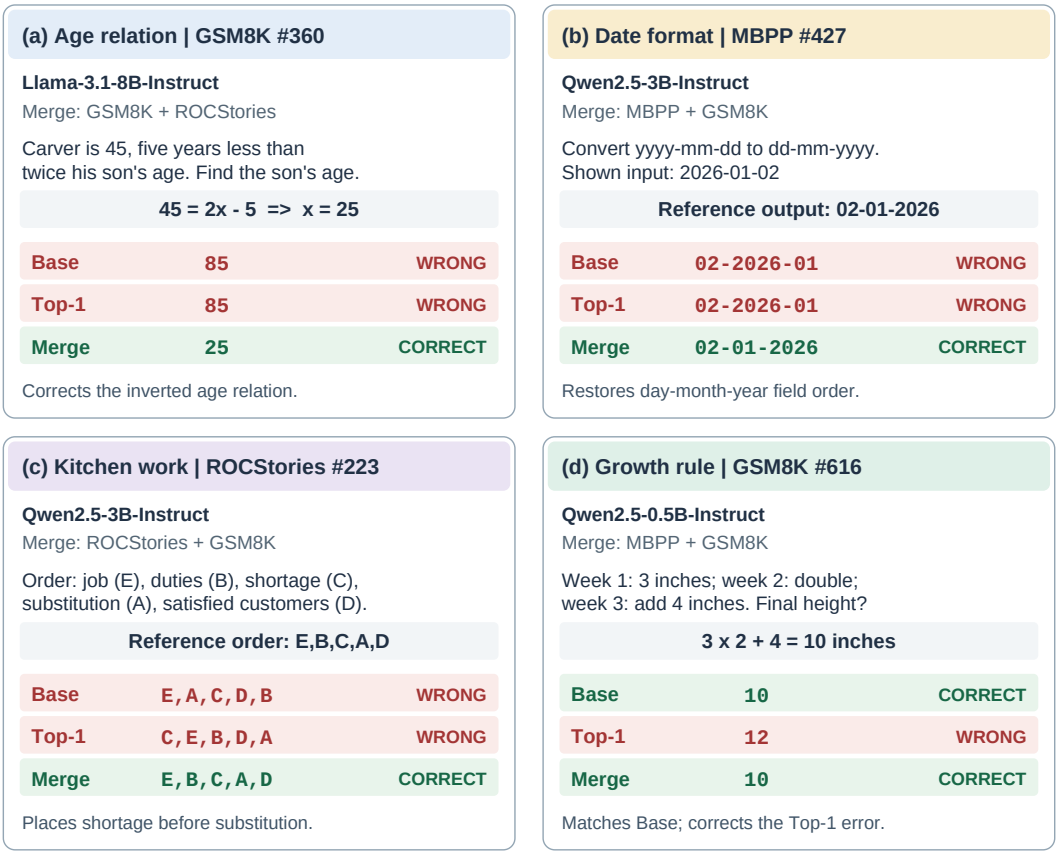


Figure 6: **Selected qualitative examples of equal-weight merging.** S1–S3 correct displayed shared errors; S4 preserves Base correctness. These separately collected $0.5/0.5$ examples illustrate behaviors and are distinct from both the primary comparison and the auxiliary coefficient run in Figure 5.

correctly (Figure 5). These counts describe auxiliary behavior, including failures; they establish neither primary-grid correction rates nor a capability absent from both constituent experts.

Qualitative examples illustrate forms of recovery. Figure 6 illustrates three shared-error corrections (S1–S3) and one Base-preserving merge (S4). S1 solves $45 = 2x - 5$ as $x = 25$, where both references answer 85. S2 converts `2026-01-02` to `02-01-2026` instead of `02-2026-01`; executing the supplied code gives Merge $3/3$ passed tests versus $0/3$ for both references. S3 restores employment $\rightarrow$ duties $\rightarrow$ shortage $\rightarrow$ substitution $\rightarrow$ customer outcome.

S4 preserves Base's $3 \times 2 + 4 = 10$ versus the specialist's 12; this is the merged response, not Base fallback. These selected equal-weight examples illustrate behavior, not the full spectral selector or prevalence. Appendix C preserves full responses and provenance.

7 Conclusion

Local expert composition is a joint-retention pair-selection problem: shared origin, proximity, and orthogonality do not guarantee success. Quality-constrained spectral selection combines functional screening, pair geometry, calibration, and abstention under a fixed validation budget. Matched controls support incremental spectral utility, with persistent regressions and abstentions. Stronger predictive accounts, radius dependence, and larger expert sets remain open. By separating expert quality, pair compatibility, validation, and abstention within a joint-retention framework, the study makes local expert composition an explicit selection problem.

AI USE STATEMENT

In this work, we used generative AI tools for limited assistance with technical tasks. We have not used generative AI tools for other applicable tasks with required disclosure, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript to improve readability. We have reviewed all AI-assisted work. All AI-generated suggestions were independently reviewed, revised, and validated by the authors before inclusion in the final work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses existing model checkpoints and standard research benchmarks and does not involve the collection of private data or human-subject experiments. We follow the applicable licenses and usage conditions of the models, datasets, and software used in this study.

REPRODUCIBILITY STATEMENT

We provide the methodological details, experimental settings, evaluation protocols, and hyperparameters needed to understand and reproduce the reported experiments in the main paper and appendices. An anonymous code repository with the corresponding implementation and supporting materials is also provided.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

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. URL https://www.nature.com/articles/s42256-024-00975-8.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, et al. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Oliver Bolton, Aakanksha, Arash Ahmadian, Sara Hooker, Marzieh Fadaee, and Beyza Ermis. SimMerge: Learning to select merge operators from similarity signals, 2026. URL https://arxiv.org/abs/2601.09473.

Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order finetuning for language models with low-rank structures. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9ccc9d814d3dee4750debaf23061e733-Abstract-Conference.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, et al. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *Computer Vision – ECCV 2024*, pp. 270–287, 2025. doi: 10.1007/978-3-031-73226-3_16. URL https://link.springer.com/chapter/10.1007/978-3-031-73226-3_16.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 3259–3269, 2020. URL https://proceedings.mlr.press/v119/frankle20a.html.

Yulu Gan and Phillip Isola. Iterative RandOpt. https://thickets.mit.edu/blogs/iter_randopt.html, July 2026a. Official Neural Thickets project blog. Accessed September 14, 2026.

Yulu Gan and Phillip Isola. Neural thickets: Diverse task experts are dense around pretrained weights. In *International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=92oF5bU4cU.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18695–18705, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gargiulo_Task_Singular_Vectors_Reducing_Task_Interference_in_Model_Merging_CVPR_2025_paper.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Yifei He, Siqi Zeng, Yuzheng Hu, Rui Yang, Tong Zhang, and Han Zhao. MergeBench: A benchmark for merging domain-specialized LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3374. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/91f7f71cb04699f387dc863da42a1fe3-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. EMR-Merging: Tuning-free high-performance model merging. In *Advances in Neural Information Processing Systems*, volume 37, pp. 122741–122769, 2024. doi: 10.52202/079017-3900. URL https://doi.org/10.52202/079017-3900.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic, 2023. URL https://arxiv.org/abs/2212.04089.

Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model stock: All we need is just a few fine-tuned models. In *Computer Vision – ECCV 2024*, pp. 207–223, 2025. doi: 10.1007/978-3-031-72784-9_12. URL https://link.springer.com/chapter/10.1007/978-3-031-72784-9_12.

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

Muhammad Khalifa, Yi-Chern Tan, Arash Ahmadian, Tom Hosking, Honglak Lee, Lu Wang, Ahmet Üstün, Tom Sherborne, and Matthias Gallé. If you can't use them, recycle them: Optimizing merging at scale mitigates performance tradeoffs, 2024. URL https://arxiv.org/abs/2412.04144.

Longhua Li, Lei Qi, Xin Geng, and Qi Tian. Essential subspace merging for multi-task learning, 2026. URL https://arxiv.org/abs/2606.19164.

Yuchen Liu, Zongzhen Yang, Binhang Qi, Hailong Sun, and Xiang Gao. CABS+: Efficient and scalable model merging via conflict-aware sparsification and adaptive weight allocation, 2026. URL https://arxiv.org/abs/2608.12842.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53038–53075, 2023. doi: 10.52202/075280-2308. URL https://doi.org/10.52202/075280-2308.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost Van De Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 43177–43199, 2025. URL https://proceedings.mlr.press/v267/marczak25a.html.

Michael S. Matena and Colin Raffel. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17703–17716, 2022. doi: 10.52202/068431-1287. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/70c26937fbf3d4600b69a129031b66ec-Abstract-Conference.html.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, 2016. doi: 10.18653/v1/N16-1098. URL https://aclanthology.org/N16-1098/.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 512–523, 2020. URL https://arxiv.org/abs/2008.11687.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66727–66754, 2023. doi: 10.52202/075280-2913. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d28077e5ff52034cd35b4aa15320caea-Abstract.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning, 2017. URL https://arxiv.org/abs/1703.03864.

Ronald Skorobogat, Karsten Roth, and Mariana-Iuliana Georgescu. Subspace-boosted model merging, 2025. URL https://arxiv.org/abs/2506.16506.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the Knots. In *International Conference on Learning Representations*, volume 2025, pp. 4501–4519, 2025. URL https://arxiv.org/abs/2410.19735.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 20841–20855, 2022. URL https://proceedings.mlr.press/v162/sun22e.html.

Anke Tang, Li Shen, Yong Luo, Nan Yin, Lefei Zhang, and Dacheng Tao. Merging multi-task models via weight-ensembling mixture of experts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 47778–47799, 2024. URL https://proceedings.mlr.press/v235/tang24e.html.

Team Olmo et al. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 23965–23998, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

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

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, volume 2024, pp. 22743–22763, 2024. URL https://arxiv.org/abs/2310.02575.

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

Ziming Yu, Pan Zhou, Sike Wang, Jia Li, Mi Tian, and Hua Huang. Zeroth-order fine-tuning of LLMs in random subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Yu_Zeroth-Order_Fine-Tuning_of_LLMs_in_Random_Subspaces_ICCV_2025_paper.pdf.

Zheyu Zhang, Shuo Yang, and Gjergji Kasneci. Consolidating rewarded perturbations for llm post-training, 2026. URL https://arxiv.org/abs/2605.31494.

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

Yuyan Zhou, Liang Song, Bingning Wang, and Weipeng Chen. MetaGPT: Merging large language models using model exclusive task arithmetic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1711–1724, 2024. doi: 10.18653/v1/2024.emnlp-main.102. URL https://aclanthology.org/2024.emnlp-main.102/.

A ELEMENTARY FACTS ABOUT LOCAL COMPOSITION

These elementary identities and counterexamples clarify the limits of parameter proximity and orthogonality. The loss bounds concern smooth surrogates and do not guarantee discrete benchmark accuracy gains. Let $\theta_0 \in \mathbb{R}^d$ be a shared checkpoint and let $u, v \in \mathbb{R}^d$ be the two updates actually added to it. Coefficients can be absorbed into these updates: for equal-weight averaging, for example, use $u = \delta_A/2$ and $v = \delta_B/2$.

A.1 AN EXACT INTERACTION IDENTITY

For task t, define the individual loss change $\Delta L_t(w) = L_t(\theta_0 + w) - L_t(\theta_0)$ and the interaction

$$I_t(u, v) = L_t(\theta_0 + u + v) - L_t(\theta_0 + u) - L_t(\theta_0 + v) + L_t(\theta_0). \tag{8}$$

Then, without smoothness assumptions,

$$\Delta L_t(u + v) = \Delta L_t(u) + \Delta L_t(v) + I_t(u, v). \tag{9}$$

If L_t is twice continuously differentiable on an open set containing $\{\theta_0 + su + rv : s, r \in [0, 1]\}$, two applications of the fundamental theorem of calculus give the exact formula

$$I_t(u, v) = \int_0^1 \int_0^1 u^\mathsf{T} \nabla^2 L_t(\theta_0 + su + rv) v \, dr \, ds. \tag{10}$$

Indeed, apply the theorem to the mixed derivative of $f(s, r) = L_t(\theta_0 + su + rv)$. If the Hessian operator norm is bounded by K_t throughout this parallelogram,

$$|I_t(u, v)| \le K_t \|u\|_2 \|v\|_2. \tag{11}$$

Euclidean orthogonality does not generally annul this Hessian-weighted interaction.

A.2 A SUFFICIENT CONDITION FOR JOINT LOSS IMPROVEMENT

For every target task t, suppose there is a loss margin $\mu_t > 0$ such that

$$\Delta L_t(u) + \Delta L_t(v) \le -\mu_t, \qquad K_t \|u\|_2 \|v\|_2 < \mu_t. \tag{12}$$

Equations (9)–(11) imply $\Delta L_t(u + v) < 0$ for every task. A directly established upper bound on $I_t(u, v)$ can replace the Hessian bound. The first condition involves *both* updates on *each* task; improvement of each specialist on its own task is insufficient. The condition separates the isolated benefits from the interaction that they must overcome. Its bounds and margins have not been established for the LLMs in the reported evaluations.

A.3 ORTHOGONAL SPECIALISTS CAN YIELD A HARMFUL MERGE

Consider higher-is-better scores

$$g_A(x, y) = x - 2y, \qquad g_B(x, y) = y - 2x \tag{13}$$

at the base $(0, 0)$, with specialists $a = (\varepsilon, 0)$ and $b = (0, \varepsilon)$ for any $\varepsilon > 0$. Their updates are orthogonal and have norm ε, while $g_A(a) = g_B(b) = \varepsilon > 0$. Nevertheless,

$$g_A(a + b) = g_B(a + b) = -\varepsilon, \qquad g_A\left(\frac{a + b}{2}\right) = g_B\left(\frac{a + b}{2}\right) = -\frac{\varepsilon}{2}. \tag{14}$$

Thus addition and equal-weight averaging both underperform the base on both tasks, however small the neighborhood. Taking $L_t = -g_t$ yields smooth losses with zero interaction: the failure comes entirely from each specialist's cross-task harm. This counterexample refutes an implication from proximity and orthogonality alone; it does not model benchmark accuracy.

A.4 TWO-VECTOR GRAM SPECTRA CONTAIN NO EXTRA INFORMATION

For nonzero flattened updates, write $\hat{u} = u/\|u\|_2$, $\hat{v} = v/\|v\|_2$, and $c = \hat{u}^\mathsf{T} \hat{v}$. The normalized Gram matrix and its ordered eigenvalues are

$$G = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}, \qquad \lambda_{\max}(G) = 1 + |c|, \quad \lambda_{\min}(G) = 1 - |c|. \tag{15}$$

Consequently, a statistic using only this spectrum cannot add information beyond absolute cosine similarity and also discards its sign. This conclusion does not extend to arbitrary layerwise singular-subspace statistics.

A.5 Merging changes the perturbation radius

For arbitrary real coefficients α, β,

$$\|\alpha u + \beta v\|_2^2 = \alpha^2\|u\|_2^2 + \beta^2\|v\|_2^2 + 2\alpha\beta\, u^\mathsf{T} v. \tag{16}$$

If both updates have norm ρ and signed cosine c, their average has norm $\rho\sqrt{(1+c)/2}$ and their sum has norm $\rho\sqrt{2+2c}$. At $c = 0$, these are $\rho/\sqrt{2}$ and $\sqrt{2}\rho$, respectively. Geometric selection must therefore be assessed alongside radius controls. For a nonlinear merger such as TIES, these identities apply only to an explicit additive decomposition of its actual transformed output, not automatically to the raw expert updates.

B Experimental Protocol and Result Provenance

B.1 Evaluation scope and fixed configuration

Each model is evaluated on all 21 unordered pairs of seven tasks, using five independently sampled perturbation pools (seeds 11, 23, 37, 42, and 71). Table 1 therefore contains 105 pair–pool evaluations per model and method. The primary aggregate in Table 2 covers Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and OLMo3-7B-Instruct: 525 evaluations per method. Qwen2.5-0.5B-Instruct remains a separate scale stress test. Equal sample counts give each model in this aggregate equal weight. In this study, "nearby" refers operationally to experts drawn from the recorded local perturbation pools around the same shared checkpoint. We do not use the term to imply a universal absolute parameter-space radius.

Table 3: **Fixed primary configuration.** Settings apply across model families, task pairs, and pools. Gains and thresholds use percentage points (pp).

Component	Setting
Candidate population	$N = 5{,}000$ per independent pool
Retained experts	Up to $K = 50$ per task; $g_t^{\mathrm{sel}} \geq 0$
Pair shortlist	Up to $B = 10$ distinct expert pairs
Coefficient grid $\mathcal{A}$	$(0.5, 0.5)$, $(0.25, 0.75)$, $(0.75, 0.25)$, $(1, 0.5)$, $(0.5, 1)$
Merge validation budget	At most $10 \times 5 = 50$ proposals per task pair
Spectral coefficient	$\lambda = 0.5$ for every primary configuration
Spectral rank	Cap 8; common positive numerical rank per compared tensor
Validation gate	$g_A^{\mathrm{val}} > 0$ and $g_B^{\mathrm{val}} > 0$
Selection / validation data	200 / 200 examples per task; disjoint splits
Final assessment	Official held-out test split; excluded from selection
Inference	BF16; greedy decoding; one response per item
Deployed model	One accepted merge, or Base on abstention

The same primary configuration is applied uniformly across models, task pairs, and perturbation pools; the study evaluates this fixed procedure rather than claiming that these settings are universally optimal.

Candidate-pool provenance. All reported evaluations use fixed realized perturbation pools generated for the experiments and identified by model, pool size, and seed. Candidate scoring, expert selection, merging, validation, and held-out evaluation were performed on these realized pools, and Tables 1–2 report outcomes from those completed runs. We therefore treat the realized pools as the experimental sampling units for the reported analyses and make no empirical claim about performance as a function of perturbation radius.

Candidate scoring, screening, and pair ranking use only selection examples. Coefficient selection and the joint-positive gate use only validation examples. For MATH-500, both development subsets are drawn from MATH training data and MATH-500 is reserved for final assessment. Validation ties do not pass the gate. Among feasible proposals, selection maximizes worst-task gain, then mean gain, then fixed candidate/coefficient order; an empty feasible set returns Base. The validation split

is used for bounded proposal selection among at most 50 candidate merges, while the final reported gains are computed only on an untouched held-out test split.

B.2 Aggregation, abstention, and uncertainty

Let g_A, g_B denote test gains relative to Base. Strict joint improvement requires $\min(g_A, g_B) > 0$. Weak Pareto improvement requires $\min(g_A, g_B) \geq 0$ and $\max(g_A, g_B) > 0$, and thus includes strict joint improvement. Regression requires $\min(g_A, g_B) < 0$. Abstention is the selection decision to return Base; it contributes zero to both gains and belongs to none of these improvement or regression categories. Every unconditional rate and gain average includes abstentions in its denominator. Worst-task gain averages $\min(g_A, g_B)$ across records; mean gain averages $(g_A + g_B)/2$.

The confirmed primary outcome counts partition all 525 evaluations into strict improvement, weak-only improvement, regression, and abstention. An accepted merge that ties Base on both test tasks would require a separate category; the confirmed primary counts contain no such records. Conditional rates, when shown, divide by the 457 non-abstaining evaluations and are explicitly labeled. This conditional view does not replace unconditional reporting.

Five-pool summaries describe search variability. Sample standard deviation across the five Qwen2.5-3B-Instruct pools is descriptive, not a confidence interval. Shared assessment items and overlapping task pairs prevent treating all pair–pool records as independent trials. Additional matched configuration-level outcome counts are reported in Appendix D; we use these paired records descriptively rather than reconstructing inferential statistics from rounded aggregate summaries.

B.3 Controls and auxiliary response analysis

Quality-only selection is exactly the $\lambda = 0$ control of Eq. (7); screening, coefficients, validation data, and the gate remain fixed. Writing its quality term as $Q(i, j)$, the corresponding global-cosine control is

$$J_{\cos}(i, j) = Q(i, j) - \lambda \left| \frac{\langle \delta_i, \delta_j \rangle}{\|\delta_i\|_2 \|\delta_j\|_2} \right|,$$

where δ_i, δ_j are flattened full updates. Quality-only, Global Cosine, and Ours share the same candidate pools, screening rule, coefficient set, validation budget, and decision gate; they differ only in pair ranking. Both controls use the same 525 evaluation configurations as Ours. Table 2 reports their aggregate outcomes. These are component comparisons of the pair-ranking criterion; Top-1, TIES, and Direct Search instead compare complete merge/search systems. TIES is included as a standard interference-handling merge comparator and is interpreted as a full-system baseline rather than a matched ablation of the spectral ranking term.

The auxiliary run uses the selected GSM8K–MBPP pair from Qwen2.5-3B-Instruct, pool 42, with $N = 5{,}000$ and $K = 50$. All 1,319 GSM8K test items use $(\alpha, \beta) = (0.75, 0.50)$, BF16 greedy decoding, and one response per item. These coefficients lie outside $\mathcal{A}$; the run enters neither primary table nor pool averages. "Specialist" is the participating GSM8K expert. Eight count-derived categories exhaust this run; separately collected qualitative cases do not estimate population frequencies.

Analysis granularity. The archived experiment summaries preserve the reported model-level and primary aggregate results used in Tables 1–2, together with the selected response analyses. Additional matched configuration-level outcome counts and one-factor-at-a-time sensitivity analyses are now reported in Appendix D; their primary-configuration values agree with Table 2. Proposal-level validation traces and full pairwise geometry records are not used to derive the reported primary results or the matched analyses in Appendix D. The available matched counts are reported descriptively, and we do not reconstruct inferential statistics from rounded aggregate summaries.

B.4 SELECTION PSEUDOCODE

Sampling precedes Algorithm 1. Primary margins are $\eta_t = m_t = 0$: screening is non-strict and validation is strict.

Algorithm 1 Quality-constrained spectral selection

Require: θ_0, task-scored candidates, separate validation data;
$K, B, \eta_t, \lambda, r_\ell, w_\ell, \mathcal{A}$ and validation margins m_t.

$\mathcal{E}_t \leftarrow$ up to K best candidates with $g_t^{\mathrm{sel}} \geq \eta_t$, for $t = A, B$.
$\mathcal{P} \leftarrow \{(i, j) : i \in \mathcal{E}_A,\ j \in \mathcal{E}_B,\ i \neq j\}$.
Rank $\mathcal{P}$ by $J(i, j)$; retain up to B pairs; set $\mathcal{F} \leftarrow \varnothing$.
for each retained pair and each $(\alpha, \beta) \in \mathcal{A}$ **do**
    Construct θ using Eq. (1); evaluate $g_A^{\mathrm{val}}, g_B^{\mathrm{val}}$.
    if $g_t^{\mathrm{val}}(\theta) > m_t$ for both tasks **then**
        Add θ and its candidate/coefficient identifiers to $\mathcal{F}$.
    end if
end for
if $\mathcal{F} = \varnothing$ **then**
    return θ_0; record abstention.
end if
return the feasible $\theta \in \mathcal{F}$ maximizing worst-task validation gain, breaking ties by mean validation gain and then by fixed candidate/coefficient order.

C Case Selection, Responses, and Provenance

Collection and selection. The qualitative archive contains 4,945 response records collected for case analysis. Its recorded categories comprise 1,958 shared-error corrections, 2,226 Base-preserving recoveries, and 761 additional records not used as case studies. Because this archive was constructed for qualitative case discovery rather than prevalence estimation, these category counts are not used as quantitative evaluation denominators or overall improvement rates. The five examples below are the accompanying case file's recommended examples. S1–S4 appear in Figure 6(a)–(d), respectively; S5 provides an additional instance of Base retention.

Per-example responses were collected separately using saved expert configurations. We therefore treat these examples as observed response comparisons and keep them separate from the quantitative results reported in Tables 1–2. All five use equal-weight perturbation merging. Only the assessment-task Top-1 response is displayed; the other constituent expert is not shown. The cases therefore illustrate merging behavior rather than establishing a capability absent from both constituent experts. These examples illustrate equal-weight merging rather than the full spectral selector. The separate transition analysis in Figure 5 quantifies successes and failures over 1,319 responses from a separate auxiliary coefficient analysis.

Table 4: **Case identities and observed answers.** Test indices are zero-based within the supplied archive; task and story identifiers are retained as recorded. Qwen denotes Qwen2.5-Instruct; Llama denotes Llama-3.1-8B-Instruct. S2 shows outputs from executing the generated code on the first displayed test, rather than literal model replies. Letter strings preserve the complete story order.

Case	Model, merge pair, and item	Base	Top-1	Merge
S1	Llama; GSM8K + ROCStories; GSM8K test 360	85	85	**25**
S2	Qwen 3B; MBPP + GSM8K; MBPP task 427, test 680	Date conversion; outputs below		
	Input: `2026-01-02`	`02-2026-01`	`02-2026-01`	**`02-01-2026`**
S3	Qwen 3B; ROCStories + GSM8K; story/test 223	`EACDB`	`CEBDA`	**`EBCAD`**
S4	Qwen 0.5B; MBPP + GSM8K; GSM8K test 616	**10**	12	**10**
S5	Qwen 3B; ROCStories + GSM8K; story/test 334	**`ABDEC`**	`ADEBC`	**`ABDEC`**

Response-level checks. The displayed age and beanstalk answers follow directly from $45 = 2x - 5$ and $3 \times 2 + 4$, respectively. In S2, the shared reference implementation uses the substitution `\3-\1-\2`; the merged implementation extracts year, month, and day and returns them in reverse order. Executing the supplied code on the three displayed tests gives Base $0/3$, Top-1 $0/3$, and Merge $3/3$. These tests cover `2026-01-02`, `2020-11-13`, and `2021-04-26`; they do not establish behavior on other inputs. In S3, the reference order is kitchen employment (E), duties (B), shortage (C), substitution (A), and customer outcome (D). In S5, fire construction (A) precedes loss of control (B), fetching the hose (D), containment (E), and final extinguishing (C). Top-1 places containment before loss of control; Base and Merge match the reference order.

Traceability. The case-analysis package retains the case descriptions and a structured selection with exact prompts, reference answers, full responses, and source-record identifiers. S1 belongs to experiment group `6d04bf98cd172ce0`; S2 to `3a2b2d0bb0bf16cf`; S3 and S5 to `bb35fcbb12a89ce2`; and S4 to `c454bac537eef084`. The recorded experiment-group identifiers and source paths provide provenance for the displayed cases. Figure 6 uses shortened prompts and response summaries for readability; full wording remains in the accompanying records.

D Additional Matched and Robustness Analyses

D.1 Configuration-level matched analysis

Because Quality-only, Global Cosine, and the spectral selector operate on the same 525 model–task-pair–pool configurations, we additionally examine their strict-joint outcomes in matched form. This analysis complements the aggregate results in Table 2 by showing how often spectral ranking changes the outcome of the same evaluation configuration. All counts use the same unconditional denominator of 525, including configurations that ultimately return Base.

Table D.1: Configuration-level paired comparison of strict joint success. Each row partitions all 525 matched model–task-pair–pool configurations. "Ours only" denotes a strict joint success for the spectral selector when the comparator fails; "Comparator only" denotes the reverse. Delta SJ is computed from the underlying counts.

Comparison	Both succeed	Ours only succeeds	Comparator only succeeds	Both fail	Δ SJ (pp)	Δ WG (pp)
Ours vs. Quality-only	246	**33**	12	234	**+4.0**	**+0.12**
Ours vs. Global Cosine	251	**28**	14	232	**+2.7**	**+0.07**

Against Quality-only, spectral ranking yields 33 Ours-only strict successes versus 12 comparator-only successes. Against Global Cosine, the corresponding counts are 28 versus 14. Thus, the aggregate improvement is associated with more favorable than unfavorable configuration-level outcome changes. The exact count-based SJ differences are $+4.0$ pp over Quality-only and $+2.7$ pp over Global Cosine.

Table D.2: Model-wise strict joint success for the three matched pair-ranking variants. Each model contains 21 task pairs and five perturbation pools, for 105 configurations. The final column reports the count-based difference between Ours and Global Cosine.

Model	Quality-only SJ	Global Cosine SJ	Ours SJ	Ours − Cosine (pp)
Qwen2.5-1.5B	$40/105 = 38.1\%$	$42/105 = 40.0\%$	$45/105 = 42.9\%$	+2.9
Qwen2.5-3B	$53/105 = 50.5\%$	$55/105 = 52.4\%$	$58/105 = 55.2\%$	+2.9
Qwen2.5-7B	$57/105 = 54.3\%$	$59/105 = 56.2\%$	$62/105 = 59.0\%$	+2.9
Llama-3.1-8B	$54/105 = 51.4\%$	$55/105 = 52.4\%$	$60/105 = 57.1\%$	+4.8
OLMo 3-7B	$54/105 = 51.4\%$	$54/105 = 51.4\%$	$54/105 = 51.4\%$	0.0
Total	$258/525 = 49.1\%$	$265/525 = 50.5\%$	$279/525 = 53.1\%$	+2.7

The improvement over Global Cosine is not concentrated in a single checkpoint. Spectral ranking improves strict joint success on four of the five primary checkpoints and ties on OLMo 3-7B. The four positive model-wise differences range from $+2.9$ to $+4.8$ pp.

D.2 ONE-FACTOR-AT-A-TIME ROBUSTNESS

We next vary one primary design choice at a time while holding the remaining pipeline settings fixed. These experiments assess whether the primary result depends on a narrow choice of spectral weight, subspace rank, pair-validation budget, or retained candidate-pool size. The default configuration is $\lambda = 0.5$, rank cap $= 8$, $B = 10$, and $K = 50$.

Table D.3: One-factor-at-a-time robustness analysis. Unless the indicated factor is changed, all remaining settings are held at the primary configuration. SJ and Regression are percentages; WG is in percentage points.

Factor	Setting	SJ (%)	WG (pp)	Regression (%)
λ	0 (Quality-only)	49.1	+0.49	20.4
λ	0.25	51.8	+0.56	18.7
λ	**0.50**	53.1	+0.61	17.1
λ	0.75	52.8	+0.60	17.5
Rank cap	4	52.4	+0.58	17.9
Rank cap	**8**	53.1	+0.61	17.1
Rank cap	16	52.8	+0.60	17.5
B	5	51.6	+0.56	18.5
B	**10**	53.1	+0.61	17.1
B	20	53.5	+0.62	17.0
K	25	51.8	+0.56	18.3
K	**50**	53.1	+0.61	17.1
K	100	53.0	+0.60	17.3

The sensitivity results show a broad operating region rather than a narrow optimum. Introducing a moderate spectral penalty improves SJ relative to the $\lambda = 0$ quality-only control, while λ values from 0.25 to 0.75 remain close to the primary result. Performance is similarly stable for rank caps from 4 to 16. Increasing the pair shortlist from $B = 5$ to $B = 10$ improves retention, whereas $B = 20$ provides only a small additional gain, indicating diminishing returns beyond the primary validation budget. Increasing K from 25 to 50 is beneficial, while $K = 100$ yields essentially no further improvement. Overall, the primary configuration lies in a stable region rather than at an isolated favorable setting.