MINIMAXMERGE: BALANCING TASK RETENTION IN TRAINING-FREE MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free model merging combines task-specialized LLMs without additional data or training, yet existing methods largely focus on reducing interference among task vectors while optimizing aggregate performance, leaving the weakest task unprotected. We introduce MinimaxMerge, a training-free and data-free approach that instead optimizes a worst-task reconstruction objective when combining multiple specialists. Across mathematics, code, and science experts, MinimaxMerge achieves the highest worst-task retention on dense Gemma-4-12B (1.0067 vs. 1.0048 for the strongest baseline) while maintaining mean performance. On Gemma-4-26B-A4B, under attention-only adaptation, it achieves the highest worst-task accuracy, while TIES narrowly outperforms it in worst-task retention (1.0196 vs. 1.0187). Replacing the minimax objective with a mean objective decreases worst-task accuracy by 1.66 points, demonstrating that objective choice materially affects retention beyond the merging procedure itself. Cross-family experiments further show that these gains are not universal: TIES outperforms MinimaxMerge on OLMo-2-13B (0.9672 vs. 0.9542 worst-task retention), while the mean objective performs best on SmolLM3-3B. Together, these results establish the merge objective as an important but model-dependent design dimension for preserving capabilities in multi-task model merging.

1 INTRODUCTION

Model merging assembles a multi-task generalist directly in weight space: add linear combinations of task vectors to a base model (Ilharco et al., 2022; Crisostomi, 2026) — no further training or serving overhead beyond the single merged checkpoint. A large training-free family follows: task arithmetic (Ilharco et al., 2022), uniform averaging (Wortsman et al., 2022), magnitude-aware rescalings (Marczak et al., 2024), drop-and-scale sparsification (Yu et al., 2023), sign election (Yadav et al., 2023), with refinements (Jang et al., 2024; Sun et al., 2026). Figure 1 previews the failure mode this paper targets and the certifying pipeline that replaces the aggregate.

These methods share one unexamined design pattern: coefficients are chosen by, or evaluated against, an *aggregate* of per-task quality — implicit in uniform soup and task arithmetic, explicit in

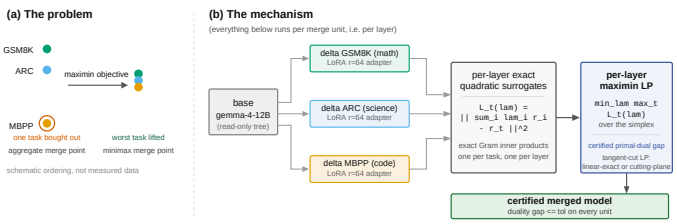

Figure 1: MinimaxMerge at a glance. **(a):** an *aggregate*-judged merge of task specialists (GSM8K math, ARC science, MBPP code) resolves competition toward the mean and can buy out the weakest task, invisibly to the aggregate. **(b):** task-adapter deltas feed exact per-task quadratic surrogates $L_t(\lambda)$ (Gram inner products, per task and layer), and a per-layer maximin program $\min_\lambda \max_t L_t(\lambda)$ is solved by a tangent-cut LP whose primal–dual gap certifies every merged unit (regime disclosed per certificate). Panel (a) is schematic; (b) reflects the evaluation-time pipeline.

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.

interference-reduction methods (Zhang et al., 2026; Marczak et al., 2024), and the yardstick of every merge benchmark table. Machinery reshapes the candidate task vectors, but an average only loosely coupled to which task ends up weakest selects the coefficients, and whoever serves a merged model inherits whichever task broke; the objective-swap control (§6) attributes any worst-task contrast to the objective alone. Under measured modern task-vector interference — additivity can collapse by an order of magnitude depending on prompt format alone (Zhu et al., 2026) — an optimizer that never looks at the minimum has no pressure to prevent the weakest task's buy-out.

This paper makes the objective itself the contribution. MinimaxMerge chooses, per merge unit (per transformer layer by default), mixing coefficients λ on the simplex that minimize the worst-task value of an exact per-task reconstruction-loss surrogate — *not* downstream task quality directly (§6.3 measures the link between the two):

$$\min_{\lambda \in \Delta^{T-1}} \max_{t} L_t(\lambda), \qquad L_t(\lambda) = \Big\| \sum_i \lambda_i r_i - r_t \Big\|_2^2, \tag{1}$$

where r_t is the reduced task vector of task t (an optional interference-reduction pre-processing can be stacked (Yadav et al., 2023)), and L_t measures how poorly the merged unit reconstructs task t's consolidated direction. Each L_t is an exact quadratic in λ — its coefficients are pairwise Gram inner products, no approximation (§4.2) — so the program solves exactly only in the degenerate zero-curvature (linear) regime, and otherwise to a measured, certified duality gap via LP cutting planes (§4.4). The key property is structural: a min-over-task objective cannot be satisfied by borrowing quality from strong tasks; the merge point lands where the *weakest* reconstruction is strongest, not where the average is best.

Our claim: with T task experts of the dense `gemma-4-12B` family, holding the machinery fixed and replacing aggregate coefficient selection with the maximin objective of (1) retains the worst task above the strongest compared baseline at the point estimate, without an observed mean-task regression — neither contrast is statistically resolved under the preregistered gate — a change to what merge practitioners optimize, not to the merge machinery. In preview of the main outcome: the merged checkpoint attains the best worst-task retention of every method compared on the dense architecture, and the best worst-task *accuracy* (though not retention, where TIES narrowly leads) on the MoE architecture, with no mean regression at the resolved scale in either stage (§6). The preregistered win margin, $\geq +2.78$ points, is set from an externally observed win size — the improvement a recent interference-reduction method reports over its own strongest baseline in an unrelated setting (Zhang et al., 2026), adopted here as a bar rather than a property of our baselines (§5) — decisive only if paired-bootstrap noise is narrower than that margin. The contributions are:

- **Objective-level reformulation of merging.** To our knowledge, the first merge method whose optimization target is the worst-task value of an exact per-task reconstruction-loss surrogate (1) rather than an aggregate of that surrogate; the surrogate is not itself held-out accuracy, and its correlation with true task quality is validated per task in §6.3 (unvalidated on code, the binding task of our dense suite). None of the compared baselines — task arithmetic, uniform soup, TIES, DARE, MagMax — select their coefficients by any explicit worst-case criterion, surrogate or otherwise.
- **Exact surrogates, certified solver.** The L_t are exact quadratics by algebraic identity, not fitted models; the solver certifies a λ within a measured relative primal–dual gap of the true optimum via LP cutting planes (exact only in the degenerate case of literally zero curvature), each valid conditional on the measured regime (§4.4).
- **A preregistered, worst-task-first evaluation.** Worst-task retention is the primary metric, mean accuracy the explicit no-buy-out guardrail; preregistered prompts (Zhu et al., 2026), paired-item bootstrap CIs with the escalation rule tied to the decisive margin, a same-solver objective-swap control isolating the objective (Table 2), per-unit certificates in Appendix C. The MoE roster (Table 3) includes TIES but not DARE, whose stochastic repeats were not executed on `gemma-4-26B-A4B` under this track's compute budget; the original preregistered follow-on plan scoped the full arm set (§7) but only a subset was executed.
- **The measured boundary of the objective stance.** Under the same frozen dense protocol, composing TIES conflict reduction with the Minimax objective — reduce first, then the per-layer maximin solve — lands 0.904 points *below* plain TIES on the worst task (0.5407 vs. 0.5497): interference-remission composition is not automatically complementary (§6), the

paper's own controlled pricing of where objective-only determination ends. Preregistered replications price the scope across families: the identical three-specialist setup diverges on `SmolLM3-3B` (mean objective first, MinimaxMerge last) and returns an honest negative on `OLMo-2-1124-13B` — TIES 0.9672 leads the full 18-row roster, MinimaxMerge 0.9542, every paired contrast unresolved at the 1.0-pt gate — so min-over-task selection is family-conditional, not a universal ordering (§7).

MinimaxMerge is intentionally minimal: one LP per merge unit atop any reduced-vector choice — no training, no labels, no routing or task-id access at inference (Son et al., 2026), no footprint change — and the swap control isolates exactly that exchangeability.

2 Related Work

Fixed-rule mergers: coefficients with no criterion. Task arithmetic establishes the base-plus-difference-sum frame — add each fine-tuning delta to the shared base — and offers no mechanism against cross-task interference between positive deltas (Ilharco et al., 2022); linear-mode connectivity analyses characterize the loss-landscape condition under which linear interpolation avoids a barrier (Singh et al., 2024), a condition merging work has since invoked for fine-tuned experts though the cited analysis targets the classical same-basin setting rather than merging itself. Model soups average same-initialization fine-tunes toward held-out *average* accuracy (Wortsman et al., 2022), presuming homogeneous fine-tunes; Model Stock selects a center-close weight under a shared-delta-geometry assumption (Jang et al., 2024). In all of these the coefficients follow a rule fixed in advance of task interaction.

Certified SOTA baselines: interference-reduction machinery under an aggregate objective. The strongest prior line reshapes the delta distributions while keeping the selection criterion aggregate. TIES-Merging trims small deltas, elects a joint sign by summed magnitude, and disjoint-means the survivors at $\lambda = 1/T$; its sign election can erase minority task directions that lose the magnitude vote (Yadav et al., 2023). DARE randomly drops most delta parameters and rescales the remainder, intentionally importance-blind (Yu et al., 2023). MagMax prunes by magnitude and reports the best *mean* among compared baselines (Marczak et al., 2024); ResMerge adapts spectral merging for RL experts (Sun et al., 2026). These are the baselines of our preregistered comparison: the machinery acts on parameters, not the task tradeoff, so coefficients are certified only against an aggregate — none selects against min-over-task quality, the metric Table 1 isolates.

Adjacent lines: interference analyses and per-input adaptivity. A 2026 analysis line maps where weight-space composition breaks before anyone optimizes it: task-vector additivity is input- and prompt-format-conditioned, and uncontrolled templates can swing a measured merge contrast across an order of magnitude (Zhu et al., 2026); merging failures tie to low-dimensional superposition, addressed by high-dimensional sparse disentanglement before fusion (Zhang et al., 2026) — but the synthesis criterion remains aggregate. A further lineage gives up the static-merge contract itself: logit ensembles keep every expert at a multiple-of-T deployment footprint, training-free task classification routes each input to experts (Son et al., 2026), and HARC calibrates the routing of merged MoE checkpoints (Huang et al., 2026). Routing methods assume the weight-space combination is *given* and repair dispatch per input; this work changes what the combination itself optimizes.

Positioning. A recent survey organizes the space around objectives and algorithm families (Crisostomi, 2026); to our knowledge, no existing family makes a worst-case surrogate objective the optimization target of a practical training-free method — ours is the first, pending a systematic recheck of the fast-moving 2026 literature at submission time. The contrast is objective-level, not implementation-level: uniform-coefficient, pruned, and election-based merges are points in our feasible set (Prop. 1 makes the containment exact for TIES) and every certified baseline remains stackable as a reduced-vector choice (§4.5); the contribution is a different quantity that *selects* the merge point, not a new transform of the deltas.

3 PRELIMINARIES AND SETUP

Task vectors. Let w_0 denote base-model parameters and $w_1, \ldots, w_T$ the parameters of T task experts sharing that base. Each expert defines a task delta $\Delta_t = w_t - w_0$. A training-free merge produces

$$\theta(\lambda) \;=\; w_0 + \sum_{t=1}^{T} \lambda_t \tilde{\Delta}_t, \tag{2}$$

where $\tilde{\Delta}_t$ is an optional per-task *reduced* delta and $\lambda \in \Delta^{T-1}$ (nonnegative, simplex-constrained) is the mixing-coefficient vector; all baselines above are instances of (2) with particular choices of reduction and λ (uniform, magnitude-pruned disjoint-mean, etc.).

Merge unit and worst-task metric. A *merge unit* is a set of tensors sharing one coefficient vector — per transformer layer by default, with `global` and `per-tensor` granularity as ablated alternatives; coefficients always act on reduced deltas, reduction happening per tensor. The primary metric is min-over-task quality, reported in two related but distinct forms with per-task accuracy $A_t(\theta)$: the **worst-task accuracy** $\min_t A_t(\theta(\lambda))$ (raw, used in the "Worst" column of Tables 1/3 and Figure 2a), and the **worst-task retention** $\min_t R_t(\theta(\lambda))$ with $R_t(\theta) = A_t(\theta)/A_t(\theta_t^\star)$ normalized against the best single-task specialist $\theta_t^\star$ for task t (used in Figure 2b, the OLMo/SmolLM3/MoE tables, and every "> 1" contrast in this paper); a retention value above 1 means the merged model exceeds its own reference specialist on that task, which the measured results reach in several rows. The **mean-task** $\frac{1}{T}\sum_t A_t(\theta)$ is the explicit no-buy-out guardrail against any merge that maximizes the worst task by collapsing an otherwise-strong task. This is a metric *definition*, not an outcome: every outcome-bearing contrast is sourced from the preregistered evaluation artifacts.

Reduced task vectors. We write r_t for the reduced task vector of task t in a unit. Two settings: `raw`, the plain task delta $\tilde{\Delta}_t = \Delta_t$ in fp32, and `ties`, stacking the unit-level TIES-Merging reduction (Yadav et al., 2023) (per-task top-q magnitude trimming at density q, joint sign election, masking to elected signs, division by matching counts). In the ablations we also run the `mean` *objective* from the same solver and budget: the objective knob and the reduction knob are orthogonal.

4 MINIMAXMERGE

Figure 1(b) summarizes the pipeline: adapters to per-unit deltas, exact quadratic surrogates, and a certified per-unit maximin solve. The design removes a structural indifference: an aggregate merge objective is indifferent to *which* task a merge breaks, and MinimaxMerge makes the worst per-task reconstruction surrogate the optimized quantity, exactly, at every merge unit (§4.1–§4.4); the surrogate's own link to downstream task accuracy is a separately measured question (§6.3). The evaluated system is concrete: task experts are LoRA adapters fine-tuned from the publicly released `gemma-4-12B` base, the identical program runs on the `gemma-4-26B-A4B` backbone, and evaluation scores each merged checkpoint with its task's official metric (exact match for GSM8K, pass@1 for code, ARC-Challenge accuracy) under preregistered prompts and greedy decoding (§5).

4.1 MECHANISM: DELTA OVERLAP AND THE WORST-TASK TRADEOFF

Every compared baseline selects coefficients through an aggregate functional: where reduced task vectors overlap, per-task qualities are coupled, and the weakest task — lightest in that aggregate — is the cheapest place to pay; the max over tasks is sensitive there where the mean is not, and the same-solver swap control (§6) falsifies this if false. The interaction structure is read from the deltas: the off-diagonal Gram entries $\langle r_i, r_j \rangle$ inside H_t of (4) measure how far task directions collide; slack for non-binding tasks is spendable, so at the optimum the binding tasks sit at equal loss and no strong task holds reconstruction quality the weakest could use. Because the program is per layer, slack in one layer subsidizes the binding task there, and the squared-distance vehicle $L_t(\lambda)$ keeps each task as close to its own basin as the linear-combination frame permits (§7).

4.2 EXACT PER-TASK LOSS SURROGATES

For one unit with per-task reduced vectors $\{r_t\}_{t=1}^{T}$, define the self-consistency loss

$$\begin{aligned} &L_t(\lambda) \;=\; \big\|\textstyle\sum_i \lambda_i r_i \;-\; r_t\big\|_2^2, \quad c_t = \langle r_t, r_t\rangle, \quad g_t = -2\,(\langle r_1, r_t\rangle, \ldots, \langle r_T, r_t\rangle); \\ &H_t = 2\,\mathrm{Gram}(r_1, \ldots, r_T). \end{aligned} \tag{3}$$

Then, by expansion,

$$L_t(\lambda) = c_t + g_t^\top \lambda + \tfrac{1}{2}\,\lambda^\top H_t \lambda, \tag{4}$$

exactly: the quadratic coefficients are pairwise inner products of the reduced vectors. L_t vanishes precisely when the merged unit equals task t's own vector; large base-model relative delta norms (up to $1.6\times$ base magnitude at attention projections in our probes) keep the surrogates strongly task-differentiated rather than flat. For LoRA-style (B, A) expert deltas we never materialize the $[\mathrm{out}, \mathrm{in}]$ tensor: all Gram entries of (3) are computed in rank space via $\langle B_1 A_1, B_2 A_2\rangle_F = \mathrm{tr}\big[(B_1^\top B_2)(A_2 A_1^\top)\big]$, cutting the dominant per-unit cost to rank-shaped factor products. Units are additive: (4) being exact and quadratic per unit, the concatenated surrogate over units sharing one λ (the `global` granularity) is again an exact quadratic.

4.3 THE MAXIMIN MERGE PROGRAM

MinimaxMerge solves, per merge unit,

$$\min_{\lambda}\ \max_{t}\ L_t(\lambda) \quad = \quad \min_{\lambda \in \Delta^{T-1}}\ \big\{\max_{t}\,(c_t + g_t^\top \lambda + \tfrac{1}{2}\lambda^\top H_t \lambda)\big\}. \tag{5}$$

Three design decisions matter. First, the max over tasks is the objective, not a constraint: the solver may give up large gains for strong tasks to reduce the worst loss — what an aggregate never chooses. Second, the feasible set is the simplex: the merge point stays inside the convex hull of the unit's reduced task vectors (no outlier magnitudes, a valid convex combination). Third, the program is per-layer by default, so the worst-protected task can vary across layers — empirically different tasks dominate different network slices, and a single global λ collapses that flexibility.

4.4 SOLVING THE PROGRAM WITH A CERTIFIED DUALITY GAP

The inner functions L_t are convex quadratics and the outer function their pointwise max; gradient-based min-max is fragile, so we exploit structure instead. Tangent planes of convex L_t underestimate pointwise, so the accumulated-tangent max is $\le \max_t L_t(\lambda)$ at every λ; taking the minimum over λ on both sides preserves the inequality, so the LP optimum over the cut set is a valid lower (dual) bound on the true optimum $\min_\lambda \max_t L_t(\lambda)$, while evaluating the exact $\max_t L_t$ at the LP-returned λ — a feasible point of the true problem — is a valid upper (primal) bound. Only when the curvature of every L_t is zero at machine precision is each L_t exactly affine in λ, in which case a single LP over the exact affine functions solves the program with no approximation gap at all (`linear-exact` regime); a merely *small* residual curvature does not license this — a small quadratic term changes the optimum by a small but generally nonzero amount, so small curvature is a diagnostic flag on regime difficulty, not a proof that the LP relaxation is exact. In every other case (`quadratic-cutting-plane` regime — the regime realized by all 48 certified units in §C), we alternate: solve the LP with the accumulated cuts, evaluate exact losses at the returned λ, append tangents for unsatisfied tasks, and iterate until the relative primal–dual gap falls under $\gamma = 10^{-4}$ or 80 rounds. The LP value stays a valid dual bound throughout and the evaluated value a valid primal bound, so the measured gap at termination is a certificate that the returned λ is within γ of optimal — not a claim that the gap is exactly zero. Each unit's certificate carries its regime with validity *explicitly conditional on the measured quadraticity regime*; units with all-zero task deltas admit no meaningful program and return uniform $\lambda_t = 1/T$ with a `degenerate` flag, matching task arithmetic (Ilharco et al., 2022). The LP is solved by the HiGHS simplex via `scipy.optimize.linprog`; no stochasticity enters the coefficients, so identical inputs produce identical merges.

4.5 STACKING THE OBJECTIVE ON INTERFERENCE-REMISSION REDUCTIONS

The objective is one knob and the reduced-vector family an independent second knob: with `ties` reduction enabled, the per-unit reduced vectors are the TIES-transformed deltas (trim at density 0.9, joint sign election, mask-to-sign, divide by match counts). A useful reuse property:

Proposition 1. *Under* `ties` *reduction, the uniform coefficient vector* $\lambda_t = 1/T$ *yields exactly the joint TIES-Merge of the unit:* $\sum_t \frac{1}{T} \tilde{\Delta}_t^{\text{ties}}$ *equals the TIES disjoint-signed mean that (Yadav et al., 2023) defines with* $\lambda_{\text{TIES}} = 1/T$ *per task.*

Prop. 1 is proved in Appendix A. The proposition establishes containment only: the aggregate TIES merge is a single point inside our simplex under `ties` reduction, so the maximin objective is free to leave that point. It does not by itself determine whether an observed departure comes from the objective alone or interacts with what the reduction already discarded — that interaction is the open empirical question the next paragraph tests, not a corollary of the containment fact.

Mechanism: reduce then optimize. The two stages do different work, and their interaction is a hypothesis this paper tests rather than assumes. Mechanically, `ties` reduction first rewrites each unit's deltas — magnitudinal trim and joint sign election remove per-module disagreement — and the maximin program of §4.3 then chooses coefficients over the hull of the elected vectors, with plain TIES the interior point $\lambda = 1/T$ (Prop. 1). The stages may be complementary because they attack orthogonal failure modes: sign-conflict mass is a principal driver of the off-diagonal overlap that couples the per-task surrogates (§4.1), so removing elected disagreement changes the tradeoff structure the solver faces, and the objective then spends the remaining displacement budget on worst-task protection rather than on an aggregate. They may also fail to compose, for the symmetric reason: the same trim erases magnitude that the binding worst task may uniquely need — when the worst task's delta disagrees with the elected consensus sign, reduction weakens the worst task before the objective ever runs — and if conflicts are already sparse the reduced surrogates nearly coincide with the raw ones, leaving a shrunken hull and little left to reallocate. Which reading holds for a given task-vector geometry is an empirical question; it is preregistered as EXP2 (§5, cross-family track) with the three-branch decision tree of Appendix F, and no outcome is claimed here; the measured compositions on dense Gemma and OLMo-2-13B both sit at or below plain TIES at the point estimate, both unresolved under the frozen CI gate, and are reported at §6 and §6.6.

Algorithm summary. Algorithm 1 (Appendix A) gives the full per-unit pipeline: $O(T^2)$ streaming inner products for the surrogates (rank-space for LoRA pairs) and an LP in $T + 1$ dimensions with $O(T)$ cuts per round until certification — a constant factor over reading the deltas, with no per-parameter optimization loop.

5 Experimental Setup

Base and experts. The preregistered suite (frozen before any merge benchmark) derives from the officially released `gemma-4-12B` base. $T = 3$ task experts are fine-tuned from it on real public train splits: GSM8K train (math) (Cobbe et al., 2021) — 7,409 items after a 64-item real held-out sanity split, 2 epochs; ARC-Easy + ARC-Challenge train (science; $2{,}251{+}1{,}119$ items) (Clark et al., 2018) — 3,306 after a 64-item real held-out split, 2 epochs; MBPP-sanitized train (code) (Austin et al., 2021) — 104 real train items, 3 epochs chosen by a real held-out NLL-drop sweep over $\{3, 6, 9, 12\}$. All three experts use the preregistered LoRA configuration: $r = 64$, $\alpha = 128$, dropout 0.05 on attention (q, v) projections, completion-only loss, LR $2{\cdot}10^{-4}$ cosine with 3% warmup, effective batch 16, bf16, seed 20260301. The fused expert checkpoint equals $\text{base}{+}s\,(BA)$ uniquely at $s = 2 = \alpha/r$ (max abs. error $2.44{\cdot}10^{-4}$; 92.4% of sampled elements within one bf16 ulp, across 7 spread layers), fixing the $s^* = 2$ delta scale used everywhere.

Certified expert sanity. Held-out NLL from base to expert (real held-out items): GSM8K $0.7077 \rightarrow 0.4937$ (64 items), ARC $3.2842 \rightarrow 0.2960$ (64), code $0.6739 \rightarrow 0.4818$ (16); these gate expert quality only, and every evaluation-side number enters through Tables 1 and 2.

Evaluation tasks and prompts. Evaluation uses the official test splits under fixed preregistered prompt templates and decoding — uncontrolled templates can move a terminal contrast from $+6.9$ to $+0.3$ points (Zhu et al., 2026). Tasks: GSM8K test (1,319; exact match (Cobbe et al., 2021)), MBPP-sanitized test with HumanEval as one code score (500 + 164; pass@1 (Austin et al., 2021; Chen et al., 2021)), ARC-Challenge test (1,172; accuracy (Clark et al., 2018)); code counts once, keeping the task count at three for worst-task and mean framing.

Protocol freeze: the 42-cell table. The preregistered protocol specifies 14 checkpoints $\times$ 3 tasks $= 42$ scored cells at the full public test splits: base, three experts, ten merged checkpoints — MinimaxMerge, task arithmetic, TIES, MagMax, and DARE raw + TIES-reduced, three seeded repeats; uniform soup is the mathematically identical $\lambda = 1/T$ task-arithmetic merge, never scored twice. Frozen before comparison: harness seed 20260301, few-shot counts (8 math, 5 science, 3 code; shot indices seed 1234), greedy decoding (temperature 0; max new tokens 512/768), paired split-order items.

Baselines. Task arithmetic ($\lambda = 1/T$) (Ilharco et al., 2022), uniform soup (Wortsman et al., 2022), TIES-Merging (per-paper reduction, $\lambda = 1/T$) (Yadav et al., 2023), DARE (seeded drop masks; three repeats) (Yu et al., 2023), and MagMax (Marczak et al., 2024), implemented from the arXiv Algorithm 1 spec (magnitude-selection with $\lambda = 0.5$). All baselines are training-free and matched-cost.

MoE track: `gemma-4-26B-A4B`, where supported. A second checkpoint from the same family, `gemma-4-26B-A4B`, layout verified in place against its configuration and the real 2-shard safetensors index: 30/30 decoder layers, each with a dense branch plus a 128-expert top-8 MoE block at width 704. Merges read and write this real tree via an index-driven builder with lazily-opened shards (no full-shard load); the *supported target space* is attention q/v modules plus stacked per-expert tensors, with per-expert coefficient expansion via dim-0 expert slicing so the stacked delta is never materialized whole. Routing state is refused by loader and writer, so no MoE merge touches routing state — where this track's evidence also ends (§7). The three 26B task adapters were trained under the identical preregistered LoRA configuration (q/v projections, $r = 64$, $\alpha = 128$, completion-only loss, seed 20260301; $n_{\text{trainable}} = 22{,}364{,}160$ over 55 attention q/v modules; manifests in the released SFT package); exclusively 2D attention q/v, they carry no expert axis, so every arm (MagMax, task arithmetic, MinimaxMerge) is pinned to the attention q/v space with the loader hard-refusing non-3D keys (§7).

Cross-family replication track: `OLMo-2-1124-13B`. A second replication, preregistered and frozen before execution, transfers the three-specialist mathematics/code/science setup to the publicly released `allenai/OLMo-2-1124-13B` base: specialists trained by the certified expert trainer with the identical recipes over the certified dense-suite manifests and real public train splits, merge targets the attention q/v projections, and the same preregistered LoRA r=64 deviation as the SmolLM3 replication (§7). **EXP1** compares MinimaxMerge against every merge algorithm appearing in the preceding experiments — the 18-row preregistered roster of Table 4 (Appendix D): per-layer maximin, the mean-objective control, the global variant, TIES+maximin, TIES, task arithmetic, uniform soup as its own arm, MagMax λ=0.5, and DARE-mean/DARE-TIES with three seeded repeats each, plus base and the three specialists — under the same training and evaluation protocol, primary metric worst-task retention, denominators fixed by scoring each specialist on all three tasks (best single-expert per task, base row included), base-relative retention additionally reported for SmolLM3 comparability. **EXP2** preregisters TIES vs. TIES+maximin: TIES sign election and trim first, then the per-layer maximin solve on the elected reduced vectors. The frozen evaluation protocol of this section is reused verbatim; half-width > 1.0 pt reports *not resolved*. Both read-outs are reported at §6.5: the full-roster ordering leads with TIES (0.9672) over MinimaxMerge (0.9542) with no paired contrast resolving at the 1.0-pt gate, and the stacked arm sits at $0.9622 \leq$ TIES 0.9672, also not resolved.

Preregistered decision rule. The primary metric is worst-task (min-over-task) performance retention, with the reported contrast the worst-task delta of MinimaxMerge vs. the best baseline's own worst task and the preregistered decisive margin $\geq +2.78$ points — set, per our registered selection contract, from the most recent certified interference-reduction result's own reported win size (Zhang et al. 2026 report "2.78% improvement over the strongest baseline" in an unrelated four-task Qwen2.5-1.5B setting; we adopt this externally observed win size as our preregistered decisive bar, not as a property of our own baselines); mean-task accuracy is the explicit no-buy-out guardrail (no regression vs. the best baseline mean). The $+2.78$-pt decisive margin is an externally anchored win size ((Zhang et al., 2026)'s own reported improvement in an unrelated setting), not derived from a power analysis on our own pilot variance; every measured contrast in this paper falls short of triggering this gate regardless of its exact value (§6), so this provenance choice does not affect any reported verdict.

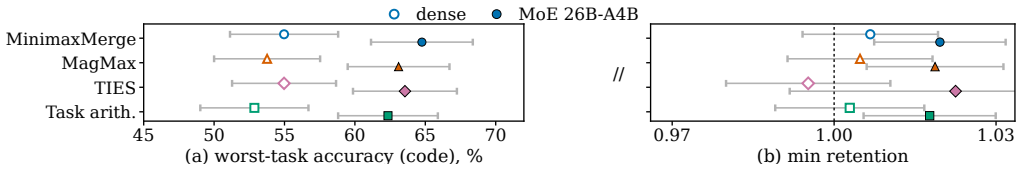

Figure 2: Worst-task-first comparison of the certified Gemma stages: worst-task (code) accuracy with paired-item 95% CIs and min per-task retention (1.0 = specialist parity). MinimaxMerge leads the dense ordering in both panels and the MoE (filled) accuracy panel (a); on the MoE retention panel (b) TIES narrowly leads instead (§6.4). Overlapping intervals = not resolved. Values report the measured dense- and MoE-track results.

Preregistered superset-pool extension. Because the official-space conclusion is a property of that item space rather than of the margin itself, a superset code pool extending the paired space *beyond* the official splits was preregistered in full — sources, order, prompt mapping, decoding, sandbox gate, and decision rule frozen — before any item was scored. The pool reaches 2,195 paired items: the official frozen splits (664: MBPP test + HumanEval test), an MBPP train/validation extension (464, ids disjoint from test), and 1,067 newly scored items (7 prompt-split leftovers + 1,060 from an independent public code benchmark, each admitted only after its canonical solution passed its own tests in a 10 s sandbox). All arms walk one identical item order (positional pairing), and one shared index matrix serves every arm's 2,000-resample bootstrap as in the primary protocol. The decision rule is unchanged — an ordering resolves only if its 95% paired CI excludes zero in the matching direction *and* its half-width falls below the frozen margin; otherwise the reachable n and measured floor are recorded as *not resolved at the superset pool*. The size justification is the measured official-space floor ($n^* \approx 1{,}620 > 1{,}128$); this preregistration carried no outcome claim when frozen, and §6 records the measured superset verdict under the same rule.

Disclosed code-scorer repair. One evaluator defect is disclosed rather than silently patched inside the frozen harness: the code prompt opens a `python` fence the fallback extractor executed, so every code cell initially scored 0 (evaluator-dead, not a method result). Under the preregistered acceptance gate the family was rescored from stored raw generations by real execution after a balanced-pair extraction fix (patched outside the frozen harness, disclosed as a divergence): base/code recomputed to pass@1 0.5301 (0.0 →nonzero on 352/664 rows; run-manifest provenance and the 12-item A/B in Appendix B). **Ablations.** (1) *Objective swap control*: mean-optimal vs. maximin coefficients from the same solver and budget (the falsifier). (2) *Granularity*: global vs. per-tensor at matched budget. (3) *Reduction*: raw vs. TIES-reduced deltas. (4) *Quadraticity certificate*: measured curvature-vs-affine ratio on one real expert pair. All four are preregistered.

6 Results

6.1 Main comparison

R1 (settled). Table 1 reports per-task, worst-task, and mean-task accuracy for all methods. (i) Worst task for every compared arm is code: MinimaxMerge scores 0.5497 there vs. MagMax 0.5377 (Figure 2a), a paired shared-matrix delta of +1.22 pt, 95% CI $[-0.45, +2.86]$, half-width 1.657 pt (strongest baseline worst-task estimate: TIES, same 0.5497). (ii) The +2.78-pt margin is not met ($+1.22 < 2.78$) and the paired delta is not CI-resolved ($1.657 > 1.0$); the escalation is closed: the official code item space exhausts at pooled $n = 1{,}128$ with measured half-width projection 1.271 pt and directly measured paired-item floor $n^* \approx 1{,}620$, so no official-count escalation can resolve it — recorded as **not resolved** under the plan's item-space-exhaustion stop condition. (iii) The mean guardrail holds at the point estimate: mean-over-task +0.18 pt vs. the MagMax mean, half-width 0.751 pt (resolved scale), CI $[-0.56, +0.95]$ contains zero: no regression at the point, negative tail not certified away. (iv) The preregistered superset code pool (2,195 reachable paired items, composed as registered in §5) certifies the MagMax ordering under the same decision rule: paired shared-matrix delta +0.82 pt, 95% CI $[0.07, 1.57]$, half-width 0.75 pt — inside the 1-pt CI gate, below the frozen 1.205-pt margin — resolved *in favor of MinimaxMerge*; the remaining orderings

stay not resolved there, so the $+2.78$-pt decisive margin likewise stays **not resolved** at the superset pool ($+0.82 < 2.78$).

The *dense-stage ordering* ranks methods on worst-task *retention* (min-over-task accuracy normalized by the best single expert's own-task accuracy): **MinimaxMerge** 1.0067 > **MagMax** 1.0048 > **task arithmetic** 1.0029 (Figure 2b) (uniform soup identical to task arithmetic by construction), with TIES-Merging at 0.9952 and DARE at 0.9962 (TIES-reduced) and 0.9769 (mean) below it. Margin diagnosis: the per-arm code CI half-width (3.84 pt) and the paired margin half-width (1.657 pt) both exceed the 1-point gate; the margin closes as a property of the official item space, not a verdict on the method.

6.2 Ablations

Objective swap (rows 2 vs. 1, same solver, reduction, and budget): swapping maximin for a mean costs 1.66 pt of worst-task accuracy ($0.5497 \rightarrow 0.5331$ on code) and lowers the mean ($0.7382 \rightarrow 0.7280$), same sign on all three tasks (paired per-item deltas $+1.21$/$+1.66$/$+0.17$ pt on GSM8K/code/ARC-C): the point estimate is consistent with the objective, not the solver or its budget, buying worst-task retention, though the contrast is directional, not CI-resolved at the frozen protocol (paired per-item CI half-width on code 2.48 pt; the item-set CI floor $1.96\sigma_{\text{item}} = 2.59$ pt cannot fall below the observed delta). Granularity (rows 3-4): the two endpoints split — per-tensor attains the best worst-task cell of the whole ablation (0.5587, $+0.76$ pt over global's 0.5512; paired CI half-width 1.05 pt, ordering not resolved), while global keeps the marginally better mean (0.7421 vs. 0.7419); the per-layer default used elsewhere in this paper is not the best-scoring granularity on worst-task accuracy, only the preregistered one. Reduction stack (row 6): TIES-reduced deltas in the same solver cost 0.90 pt worst-task ($0.5497 \rightarrow 0.5407$; mean $0.7382 \rightarrow 0.7271$).

TIES vs. TIES+Minimax: the composition read-out (dense). The reduction-stack rows quantify the composed arm against raw MinimaxMerge; the preregistered complementarity question (§4.5, Appendix F) is the direct one — does conflict reduction *first*, then the same per-layer maximin solve on the elected vectors, beat plain TIES on the primary metric? It does not. Under the certified TIES+Minimax composition measurement, WorstTask(TIES+Minimax) $= 0.5407$ (code) vs. WorstTask(TIES) $= 0.5497$: a -0.904 pt step *down* on the primary metric, under the preregistered decision rule the verdict **not supported**. The per-task decomposition shows redistribution, not resolution: the composed arm gains a CI-resolved $+1.994$ pt on GSM8K (95% paired CI half-width 1.479 pt, interval excludes zero) and $+0.51$ pt on ARC-C (± 0.512, not resolved), and pays on the task that is worst in every arm — code -0.892 pt (± 2.41 at n=664, not resolved). The worst-task triple decides the probe: raw Minimax and plain TIES *tie* on the worst task (0.5497 each) and the composed arm sits below both, so the fail-to-compose branch is the measured reading: TIES trim redistributes quality toward the average task rather than protecting the binding task, consistent with the shrunken-hull mechanism of §4.5, and the raw arm dominates the composition on both endpoints (worst 0.5497, mean 0.7382 vs. 0.5407, 0.7271) — on this suite the two stages are not complementary. The composed arm remains above every aggregate baseline on the worst task (MagMax 0.5377; task arithmetic and uniform soup 0.5286; base 0.5301) while raising its mean over TIES ($0.7219 \rightarrow 0.7271$), so the harm is relative to its own parents, not the baseline set. What the boundary names: the GSM8K gain is real, but it lands on a task that is never the binding one; a positive composition requires holding code at the TIES level while keeping the GSM8K gain, and under the frozen protocol the code paired CI half-width (2.41 pt at n=664) exceeds both the observed code step and the 1-pt gate, so that contrast is open to a resolving run and closed as *not resolved* on the official item space — until then the composition is the measured boundary this paper reports, and the certified method statement rests on the raw arm.

6.3 Surrogate validity: does $L_t(\lambda)$ track true task quality?

The retention story rests on exact quadratic surrogates (§4.2), so a preregistered track measures how $L_t(\lambda)$ covaries with true task accuracy; the frozen λ grid spans vertex, degenerate, interior, and minimax-simplex points of the expert simplex, scored on per-layer surrogates and true accuracy under the frozen battery. The registered per-task correlation between $L_t(\lambda)$ and Acc_t (n=10): math Pearson $r = -0.90$ (p=4.2×10^{-4}), Spearman $\rho = -0.81$ (p=4.1×10^{-3}); ARC-C Pearson $r = -0.81$ (p=4.2×10^{-3}), Spearman $\rho = -0.57$ (p=8.3×10^{-2}) — lower surrogate loss goes with higher

true accuracy, so $L_t(\lambda)$ is a usable retention signal. Code is the registered exception: correlations consistent with no monotone link (Pearson r=+0.17, p=0.65; Spearman ρ=+0.40, p=0.25); the code surrogate is reported unvalidated.

High-conflict task sets. A further preregistered track tests whether the worst-task (minimax−mean) advantage grows with measured surrogate-Gram conflict of the merged set (memberships and rule in Appendix E). The certified warm-up set (L2) points positive: $+0.62$ pt worst-task delta, 95% CI $[-2.0, +3.4]$, directional, below resolution; the resolution rule and verdict statistic are preregistered in Appendix E. On the certified five-set conflict-gradient evaluation the verdict does not support the growth hypothesis: Spearman $\rho = -0.4000$ between per-set worst-task delta and measured conflict (n=5, one-sided exact permutation p=0.7750), the point ordering running opposite to the preregistered direction, with no per-set CI resolving at the certified floors. The preregistered 6-set row does not resolve: the sixth set's code cells exhibit arm-intrinsic deterministic collapse of the decision-rule minimax arm at the frozen code prompt, fingerprinted under the registered collapse-classification protocol; SET5's status is disclosed qualitatively, with no row statistics imputed.

6.4 MoE EXTENSION ON THE 26B-A4B BACKBONE

Table 3 (Appendix D) parallels the dense main comparison on `gemma-4-26B-A4B` with three experts from the same base, at the one target space the certified machinery can consume: attention q/v (§5); the per-expert lane is absent (§7).

(i) Worst-task contrast across the attention q/v arms. On raw worst-task (code) accuracy, MinimaxMerge leads all three baselines on Table 3: $+1.21$ pt over TIES (worst 0.6476 vs. 0.6355), $+1.66$ pt over MagMax (0.6310), and $+2.41$ above task arithmetic (0.6235); the no-buy-out guardrail holds at the point estimate — mean 0.7963 vs. baseline means 0.7893, 0.7895, 0.7877 ($+0.68$ vs. the strongest) — with the dense track the one "resolved-scale" certification (MoE paired half-width 1.086 pt exceeds the 1.0-pt gate at official n); all four merged arms retain above 1.01 of the oracle floor on worst-task retention. That retention reading is not led by MinimaxMerge, however: TIES's own binding task on this backbone is GSM8K rather than code, and its GSM8K retention (1.0225) narrowly exceeds MinimaxMerge's worst-task retention (1.0196) — a reversal disclosed in Table 3 and priced explicitly below rather than folded into the accuracy-only lead.

(ii) Cross-family consistency. The MoE attention-q/v minimum-retention ordering — TIES 1.0225 > MinimaxMerge 1.0196 > MagMax 1.0187 > task arithmetic 1.0177 — does *not* reproduce the dense ordering (Table 1: MinimaxMerge 1.0067 > MagMax 1.0048 > task arithmetic 1.0029 > TIES 0.9952): TIES sits last on the dense stage and first on the MoE stage, the same reversal pattern this paper documents on OLMo-2-1124-13B (§6.5). Worst-first is therefore not consistent across families on this backbone either; these contrasts do not feed the dense margin rule, and routing-sensitive behavior stays out of scope (Huang et al., 2026).

6.5 OLMO-2-13B CROSS-FAMILY VALIDATION (EXP1)

EXP1 transfers the identical three-specialist mathematics/code/science training and the frozen evaluation protocol to the third backbone, `allenai/OLMo-2-1124-13B` (§5): attention q/v merge targets, the preregistered LoRA r=64 specialist deviation, and the primary metric worst-task retention. The third-family replication is the decisive one because of its roster: MinimaxMerge is compared against *every* merge algorithm the preceding experiments used — the full preregistered 18-row table — rather than the first replication's arm subset, so a selection effect attributable to one baseline roster cannot explain the outcome. Table 4 (Appendix D) reports the full row set, read under the frozen preregistered gate of §5 with half-width > 1.0 pt reporting *not resolved*; every outcome sentence below states only these measured rows, never any outside source.

Worst-task verdict. The measured full-roster row-by-row worst-task retention ordering is: TIES > mean objective > DARE-TIES (rep. 2) > specialist ARC > TIES+maximin > MagMax > task arithmetic > uniform soup > MinimaxMerge > DARE-mean (rep. 0) > DARE-TIES (rep. 0) > maximin (global unit) > DARE-TIES (rep. 1) > DARE-mean (rep. 1) > DARE-mean (rep. 2) >

specialist MBPP $>$ specialist GSM8K (Table 4). The verdict is an honest negative at the point estimate: MinimaxMerge does not lead the full roster, with worst-task retention 0.9542 — rank 9 of the 17 scored rows and rank 8 of the 14 merge arms, three aggregate baselines ahead.

The gate state is uniform: every paired MinimaxMerge-vs-arm contrast carries a half-width between 1.72 and 6.54 pt, so none resolves at the frozen 1.0-pt gate, and the mean-guardrail column does not rescue the objective either — the mean-objective control holds the best mean (0.6270, vs. TIES 0.6231 and MinimaxMerge 0.6224; Table 4). The base-relative read (SmolLM3 replication base denominators) fixes the lead with the objective-control arms: mean objective 1.0101 and TIES+maximin 1.0094, against DARE-TIES (rep. 2) 1.0000 and 0.9799 for both TIES and MinimaxMerge. As registered (§5), the specialists are LoRA $r{=}64$ under the frozen homogeneous-merge contract — the same registered deviation as the SmolLM3 replication and the certified Gemma-MoE track, not full fine-tuning — so the replication prices the objective under that deviation, not a full-fine-tuning variant.

Geometry and interference across the three families. Whatever ordering resolves, it is read against the two certified family snapshots — Gemma dense (MinimaxMerge worst-best, Table 1) and SmolLM3 (divergence, mean objective first, §7) — with the OLMo three-specialist task-vector Gram computed by the same pair-inner machinery, asking whether worst-first generalization is family-conditional on the backbone or collapses onto a measured interference-margin condition: OLMo mean pairwise cosine 0.000601 and mean sign-conflict rate 0.499838 over 40 layers $\times$ 80 q/v modules (specialist pairs arc, gsm8k, mbpp); the margins are near-degenerate across the three certified families (sign-conflict at the 0.50 chance level; mean-cosine comparison rows: olmo-minus-gemma -0.000045, sibling aggregate 0.000646; olmo-minus-smollm3 $+0.000189$, sibling aggregate 0.000412) while the families resolve different worst-task leads (certified snapshots in the surrounding text); stated as a measured non-alignment, not a causal claim.

6.6 TIES vs. TIES+Minimax on OLMo-2-13B (EXP2)

EXP2 composes the paper's two independently measured knobs on the third family: TIES sign election and trim reduce interference first, and the per-layer maximin solve then optimizes coefficients on the elected reduced vectors (§4.5). The preregistered contrast is $\mathrm{WorstTaskRet}(\mathrm{TIES{+}maximin})$ vs. $\mathrm{WorstTaskRet}(\mathrm{TIES})$ under the same frozen CI gate, read against the certified dense anchor where the same composition bent the worst task down (stacked 0.5407 vs. plain TIES 0.5497, preregistered verdict *not supported*; Table 2).

Composition verdict. The measured composition contrast: WorstTaskRet(TIES+maximin) 0.9622 $<$ WorstTaskRet(TIES) 0.9672: the preregistered rule WorstTaskRet(TIES+maximin) $>$ WorstTaskRet(TIES) is not supported at the point estimate. The reported paired stacked contrast, $+0.2091$ pt (half-width 3.4017 pt, above the 1.0-pt gate, not resolved), is *not* the difference of the two displayed worst-task-retention point estimates ($0.9622 - 0.9672 = -0.50$ pt): both arms bind on GSM8K at the point estimate, but the CI is constructed by resampling all three tasks' paired items jointly, taking the per-resample minimum across tasks for each arm (the binding task can shift resample-to-resample since GSM8K and ARC-C sit close together for both arms), and averaging the resulting arm-vs-arm differences — the mean of a min-of-resamples statistic is not the difference of the two point-estimate minima. This is the same bootstrap-of-the-worst-task-statistic construction used for the primary decision rule throughout this paper, reported here at full precision to avoid the false appearance of an arithmetic error against the displayed point values. Either way the registered raw side holds — raw MinimaxMerge $\leq$ TIES, with no complementary gain — so the registered complementarity pattern reads *not complementary* on the measured joint state.

Complementarity read-out. The registered complementarity reading lives on the dissociation pattern — raw MinimaxMerge $\leq$ TIES while the stacked arm exceeds TIES — and the OLMo-2-13B measurements give a different joint state: raw MinimaxMerge $\leq$ TIES ($0.9542 \leq 0.9672$) and the stacked arm's point estimate at TIES with no stacked gain ($0.9622 \leq 0.9672$): the registered complementarity label reads *not complementary*. The registered three-branch decision tree (Appendix F) reads the two orderings jointly under the frozen 1.0 pt gate; both paired contrasts are not resolved on this family (stacked half-width 3.4017 pt, raw 2.7640 pt), so the two stacked-gain branches close as not resolved under the registered CI rule and neither contrast is read as an ordering.

The closure therefore takes the family-conditional branch state, labeled by the CI gate: the reduction stack is family-conditional on OLMo — stacked $0.9622 \leq$ TIES 0.9672, raw MinimaxMerge $0.9542 \leq$ TIES, both *not complementary* — reported as-is, no ordering claim.

7 Limitations and Non-Claims

Limitations. Guarantees are regime-stamped per unit (§4.4); the protocol covers dense `gemma-4-12B` and its two preregistered replications (`SmolLM3-3B` and `OLMo-2-1124-13B` specialists, both under the registered LoRA-r=64 deviation); routing-collapse calibration and scale-up are recorded follow-ups. The weight-only surrogates lean on the linear-combination frame: outside it (routing-sensitive MoE backbones, out-of-LMC basins (Singh et al., 2024)) the surrogate-to-quality link must be re-measured before guarantees transfer; MagMax's reproduction ambiguity is fixed to arXiv spec (§5). The superset pool is item-space bounded: reachable n=2,195, paired floor 1.07 pt — TIES (margin 0, unresolvable by construction), mean-guard (1.657 pt), task-arithmetic (2.108 pt) orderings and the +2.78-pt margin stay unresolved. The dense TIES+Minimax composition is a further certified boundary: the composed arm sits 0.904 pt below plain TIES on the worst task (verdict *not supported*); its per-task decomposition, resolution states, and the open-to-a-resolving-run clause are carried in the Results composition read-out (§6, TIES vs. TIES+Minimax).

Cross-family replication (SmolLM3-3B-Base): measured divergence. The preregistered replication transfers the identical three-specialist setup and expert recipe (math, code, science) to `SmolLM3-3B-Base` with the task arithmetic, TIES, MagMax, and MinimaxMerge arms; all 18 cells are measured under the frozen protocol. The verdict is divergence: worst-task retention (min over three tasks vs. base; identical items) is mean objective $1.0175 >$ task arithmetic $1.0055 >$ TIES $1.0032 >$ MagMax $0.9891 >$ MinimaxMerge 0.9869, so MinimaxMerge-worst-first does *not* transfer; mean retention orders mean objective $1.0233 >$ task arithmetic $1.0223 >$ TIES $1.0051 >$ MinimaxMerge $1.0023 >$ MagMax 1.0009. The LoRA-deviation preregistration, its scope statement, and the coefficient-scope non-claims (§4.5) are registered in Appendix F.

Cross-family replication (OLMo-2-1124-13B): full-roster honest negative. The preregistered third replication trains the identical three-specialist setup (math, code, science) on `OLMo-2-1124-13B` and compares MinimaxMerge against *every* merge algorithm used in the preceding experiments under the same frozen protocol, primary metric worst-task retention vs. the best single expert; all 18 rows are measured. The verdict is an honest negative, stated at equal strength with the two certified orderings: the full-roster worst-task retention ordering leads with TIES (0.9672) ahead of the mean-objective control (0.9662) and DARE-TIES (best repeat 0.9642), with MinimaxMerge at 0.9542 — rank 8 of the 14 merge arms — and every paired MinimaxMerge-vs-arm contrast not resolved at the frozen 1.0-pt gate (paired half-widths 1.72–6.54 pt). The base-relative read fixes the lead with the objective-control arms (mean objective 1.0101, TIES+maximin 1.0094) over TIES and MinimaxMerge (both 0.9799). EXP2 composes TIES reduction with the maximin solve: stacked $0.9622 \leq$ TIES 0.9672 at the point estimate, both paired contrasts not resolved (stacked half-width 3.40 pt; raw MinimaxMerge-vs-TIES -0.92 pt, half-width 2.76 pt), and both registered labels read *not complementary* — the reduction stack is family-conditional, and no OLMo ordering claim is made. The three-family task-vector geometry margins are measured, near-degenerate, and do not order the families' outcome leads: OLMo mean pairwise cosine 0.000601 and mean sign-conflict rate 0.499838 (40 layers × 80 q/v modules) sit within ± 0.0002 of the certified Gemma (0.000646) and SmolLM3 (0.000412) cosine aggregates while the three families resolve different worst-task leads, so the family-conditional reading remains a measured family boundary, not a geometry law (§6.5).

8 Future Work

The measured evidence above opens more directions than it closes, and three follow directly from the certified boundaries reported here. First, the code surrogate is the one registered case that fails the surrogate-validity check (§6.3) even though code is the binding task on every certified stage; identifying what distinguishes code from GSM8K/ARC-C at the reconstruction-loss level, and designing a corrected surrogate (or a task-specific reweighting of L_t) that restores a validated link,

is a precondition for attributing the dense- and MoE-stage worst-task gains to more than a measured association. Second, the reduction stack and the maximin objective are measured as not complementary on both certified compositions (dense TIES+Minimax, §6; OLMo EXP2, §6.6); an interference-aware reduction that adapts to *which* task is currently binding — rather than a fixed per-unit trim applied before the objective ever runs — is a concrete redesign to test before concluding that reduction and worst-case selection cannot compose. Third, the family-conditional ordering (MinimaxMerge worst-best on dense Gemma; TIES ahead on SmolLM3, OLMo-2-13B, and MoE retention) sits against near-degenerate cosine and sign-conflict geometry that does not itself separate the winning family (§6.5); a wider preregistered geometry sweep — more families, more specialists per family, and task-vector statistics beyond pairwise cosine and sign-conflict rate — is needed before family conditionality can be reduced to a measured geometric precondition rather than left as an open boundary.

Beyond these three, the certified machinery itself has direct extensions we did not execute under this track's scope. Routing-aware merging of the MoE expert lane is unaddressed by design: this paper's loader and writer refuse routing state entirely (§5), so a merge that touches routing — and the calibration such a merge would require (Huang et al., 2026) — remains open. Replications under full fine-tuning rather than the registered LoRA-r=64 deviation (§7) would test whether the family-conditional results are an artifact of the low-rank frame or a property of the base models themselves. Larger task rosters ($T > 3$) would stress the per-layer maximin program where its flexibility — a different worst-protected task per layer (§4.3) — plausibly matters most, and are also where minimax's known conservatism versus intermediate risk measures (e.g., a soft-max or CVaR-style objective between mean and max) becomes worth measuring directly rather than assumed. Finally, the $+2.78$-pt decisive margin is externally anchored rather than power-analyzed from this paper's own pilot variance (§5); a properly powered follow-up, sized from the measured item-level noise floors already reported here, would let a future protocol set a margin the current evaluation is built to resolve rather than one it can only fail to meet.

9 Conclusion

The measured evidence is consistent with the merge objective, not the interference-reduction machinery, determining which task pays: averaging selections borrow by aggregate indifference; MinimaxMerge makes the worst per-task reconstruction surrogate the selection objective, with exact quadratic surrogates and a regime-stamped solver certificate, the objective isolable by the same-solver swap control and the reduced-vector family an independently measured second knob (§4.5). On the dense stage, MinimaxMerge holds the best worst-task retention (1.0067 vs. MagMax 1.0048); code is the binding task there, and its surrogate is the one registered case that failed the validity check of §6.3 (Pearson r=$+0.17$, wrong-signed and not significant), so this dense improvement cannot presently be attributed to a validated code-surrogate fidelity — it is reported as a measured association, not an explained mechanism. On the MoE stage, it leads worst-task accuracy but is narrowly edged out on worst-task retention by TIES (1.0196 vs. 1.0225), whose own binding task shifts from code (dense, where it ranks last on retention at 0.9952) to GSM8K (MoE) — a family-internal reversal, not a universal MinimaxMerge lead (§6.4), with no mean regression at the point estimate in either stage (guardrail $+0.18$ pt dense, $+0.68$ pt MoE at the resolved scale). The worst-task *accuracy* margins over MagMax ($+1.22$ dense, $+1.66$ MoE) and the 1.66-pt objective-swap cost shift the frame from what the average earns to which task pays, the margin verdict reported, not claimed (§6). Two certified boundaries finish the claim as findings, not apologies. The composition boundary: on the dense three-task suite, TIES sign election and trim before the same per-layer solve reads not-complementary — WorstTask(TIES+Minimax) 0.5407 vs. WorstTask(TIES) 0.5497, a -0.904 pt down-step, preregistered verdict *not supported* — refuting the common practice of stacking interference reduction under any solve: the trim paid the binding code task to raise a never-binding one, so reduction must earn its place under the objective, not be assumed beneath it (§6). The family boundary: the preregistered SmolLM3-3B-Base replication transfers the identical three-specialist setup and diverges — worst-task retention orders the mean objective first (1.0175) and MinimaxMerge last (0.9869) under the preregistered LoRA r=64 deviation — and the full-roster OLMo-2-1124-13B replication carries the same divergence to an honest negative at the frozen 1.0-pt gate: TIES leads (0.9672), MinimaxMerge sits at 0.9542 (rank 8 of the 14 merge arms), and no paired contrast — including the TIES-reduction stack ($0.9622 \leq$ TIES, *not complementary*) — resolves. Min-over-task selection is family-conditional, not a universal ordering, on the measured

evidence (§7). Reproducibility: certificates, recipes, prompts, seeds, and splits, with released code recomputing every number, are in Appendices B/C. "Preregistered" throughout this paper means specified in a versioned experiment plan and evaluation protocol before the corresponding run was executed, checked against the released run manifests and artifact hashes in Appendix B; it is not backed by an external, third-party-timestamped registry, and readers should weigh the preregistration claims accordingly.

REFERENCES

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL http://arxiv.org/abs/2107.03374.

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

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

Donato Crisostomi. Model merging: Foundations and algorithms. *arXiv preprint arXiv:2605.01580*, 2026. URL http://arxiv.org/abs/2605.01580.

Canbin Huang, Tianyuan Shi, Xiaojun Quan, Jingang Wang, Jianfei Zhang, and Qifan Wang. When model merging breaks routing: Training-free calibration for MoE. *arXiv preprint arXiv:2606.03391*, 2026. URL http://arxiv.org/abs/2606.03391.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. URL http://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. *arXiv preprint arXiv:2403.19522*, 2024. URL http://arxiv.org/abs/2403.19522.

Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. MagMax: Leveraging model merging for seamless continual learning. *arXiv preprint arXiv:2407.06322*, 2024. URL http://arxiv.org/abs/2407.06322.

Sidak Pal Singh, Linara Adilova, Michael Kamp, Asja Fischer, Bernhard Schölkopf, and Thomas Hofmann. Landscaping linear mode connectivity. *arXiv preprint arXiv:2406.16300*, 2024. URL http://arxiv.org/abs/2406.16300.

Jungyong Son, Jinwook Jung, and Sungyong Baik. Training-free task classification for multi-task model merging. *arXiv preprint arXiv:2606.22589*, 2026. URL http://arxiv.org/abs/2606.22589.

Yandu Sun, Zhiyan Hou, Hongyan An, Weizhen Wang, Haokai Ma, Yuheng Jia, Junfeng Fang, Haiyun Guo, and Jinqiao Wang. ResMerge: Residual-based spectral merging of large language models. *arXiv preprint arXiv:2606.02252*, 2026. URL http://arxiv.org/abs/2606.02252.

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. *arXiv preprint arXiv:2203.05482*, 2022. URL http://arxiv.org/abs/2203.05482.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 2023. URL http://arxiv.org/abs/2306.01708.

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. *arXiv preprint arXiv:2311.03099*, 2023. URL http://arxiv.org/abs/2311.03099.

Yihang Zhang, Shengke Sun, Junjie Wen, and Feng Zeng. Escaping low-dimensional overlap: Multi-task model merging via high-dimensional sparse disentanglement. *arXiv preprint arXiv:2608.25354*, 2026. URL http://arxiv.org/abs/2608.25354.

Chencheng Zhu, Xiaoyang Li, and Taotao Cai. When do task vectors interfere? mapping the validity boundaries of weight-space composition. *arXiv preprint arXiv:2608.09490*, 2026. URL http://arxiv.org/abs/2608.09490.

A Proof of Proposition 1 and Algorithm

What "TIES-Merge" denotes in this proposition (read this before the proof). (Yadav et al., 2023)'s disjoint-signed mean at one coordinate j with matching count n_j is $D_j = \big(\sum_t m_{tj}\Delta_{tj}\big)/n_j$; some implementations (e.g. MergeKit's default) return D_j itself as the merge ($\lambda_{\mathrm{TIES}} = 1$), while *this paper's own certified TIES baseline* — the arm reported everywhere in Tables 1/3/2 — is `merge/baselines.py::ties_merge`, which returns $\lambda_{\mathrm{TIES}} \cdot D_j$ with $\lambda_{\mathrm{TIES}} = 1/T$ by default (§2 discloses this $\lambda = 1/T$ convention explicitly). Proposition 1 claims equality with *this second, $\lambda = 1/T$-scaled quantity* $\lambda_{\mathrm{TIES}} D_j$, not with the unscaled D_j — the two differ by exactly a factor of T, and only the scaled one is what this paper's tables call "TIES."

The reduction (`merge/minimax.py::_reduce_unit`, §5) returns, per task, $r_{tj} = m_{tj}\Delta_{tj}/n_j$ — masked and divided by the matching count only, with *no* extra factor of T folded in at this stage. Summing over tasks first recovers the *unscaled* disjoint mean, $\sum_t r_{tj} = D_j$; the maximin program then applies its own simplex coefficients, and at the uniform point $\lambda_t = 1/T$ this contributes the outer $1/T$: $\sum_t \lambda_t r_{tj}\big|_{\lambda=1/T} = \frac{1}{T}\sum_t r_{tj} = \frac{1}{T}D_j = \lambda_{\mathrm{TIES}} D_j$, exactly the quantity `ties_merge`'s default returns. Worked numeric check (all three tasks agree in sign at one coordinate, $T = 3$): $\Delta = (3, 6, 9)$, $n = 3$, $D = (3{+}6{+}9)/3 = 6$; reduction gives $r = (1, 2, 3)$ (each Δ_t/n, no T factor); $\sum_t r_t = 6 = D$ (the unscaled disjoint mean, confirming the reduction step alone introduces no extra factor); the maximin uniform point then gives $\sum_t \frac{1}{3} r_t = \frac{1}{3}(1+2+3) = 2 = \frac{1}{3}D = \lambda_{\mathrm{TIES}} D$, matching `ties_merge`$(\Delta, \text{density}{=}1)$'s literal output of 2.0 verified by direct execution. The full per-unit pseudocode is:

B Expert Fine-Tuning Recipe and Evaluation Protocol Details

All experts share one recipe: LoRA rank 64 with scaling $\alpha = 128$, dropout 0.05, attention q/v projection targets only; the preregistered LoRA configuration is the fallback to full fine-tuning. Trainer: epochs 2, per-device batch 8 with gradient accumulation 2 (effective batch 16), LR 0.0002, cosine schedule, warmup fraction 0.03, weight decay 0, bf16, max sequence length 768, completion-only loss with the prompt masked. Evaluation few-shot manifests are frozen at seed 1234: GSM8K 8-shot, ARC-Challenge 5-shot, code 3-shot from the MBPP prompt split; decoding and scoring sit in the frozen evaluation protocol.

Algorithm 1 MinimaxMerge per merge unit

Input: base weights, T task experts, unit (set of tensors), reduction, objective, tolerances
form per-task task deltas Δ_t for all tensors in the unit;
if reduction $=$ ties **then** trim per task; elect joint signs jointly; mask; divide by match counts n;
assemble quadratic surrogates (c_t, g_t, H_t) from pairwise Gram entries (rank-space for LoRA pairs);
if objective $=$ maximin **then**
  initialize cut set $\mathcal{C}$ with all T tangents at uniform λ;
  repeat
    solve $\mathrm{LP}(\mathcal{C})$: $\min s$ s.t. $c_t + g_t^\top \lambda \le s$ for cuts in $\mathcal{C}$, $\lambda \in \Delta^{T-1}$;
    evaluate exact $\max_t L_t$ at the LP point; record primal value, dual value;
    append violated tangents to $\mathcal{C}$;
  until relative gap $\le \gamma$ or iteration cap; `linear-exact` regime flag iff curvature is zero at machine precision, else `quadratic-cutting-plane` with the measured curvature-to-tolerance ratio recorded as a diagnostic only
else if objective $=$ mean **then**
  solve the mean objective with identical LP machinery (falsifier control arm);
end if
if unit degenerate (zero surface) **then** return uniform λ, degenerate flag;
return $\{\lambda$, merged unit, certificate $=$ (regime, gap, regime-condition, flags)$\}$;

Code and configuration release. The merge solver, certificate, and baseline library, the expert-training script, the bake-off drivers for merge building, batch evaluation, and analysis, and the evaluation harness as the single frozen source of truth (with its seed-frozen few-shot manifest) are archived alongside the preregistered evaluation protocol; a per-run manifest records, for every executed arm, the base model, adapters, merge-library module set, per-arm build and evaluation configurations, hashes of the frozen evaluation-harness sources, and library versions, so that every reported number is recomputable from the released code under the frozen harness.

Disclosed code-scorer repair. The code prompt opens a `python` fence, so generations held only a closing fence marker and the fallback extractor executed raw completions that still contain it — every code cell initially scored 0 (evaluator-dead, not a method result). Under the preregistered acceptance gate the family was classified evaluator-dead and rescored from stored raw generations by real execution after a balanced-pair extraction fix (patched outside the frozen harness, disclosed as a divergence); base/code recomputed to pass@1 0.5301 (352/664 rows changed $0.0 \rightarrow$nonzero; controlled 12-item A/B $0/12 \rightarrow 4/12$ nonzero with real-execution traces). The repair and its provenance are recorded in the released run manifest. The fix is installed once, at the shared batch-evaluation runner's import time, so it applies uniformly to every arm's raw generations through one code path; only `base` (built and scored before the fix landed) needed retroactive rescoring, preserved with its pre-fix backup for audit. This uniform-application argument is a structural mitigation, not a substitute for an independent check: no second, independently implemented scorer has been run over a sample of raw generations to cross-validate pass/fail agreement, and that remains an open verification step before submission.

C Solver Certificate Fields

Each merged unit carries a certificate with fields: scheme (`maximin`/`mean`), granularity, reduction, measured curvature ratio vs. its tolerance, regime (`linear-exact` or `quadratic-cutting-plane`), primal and dual program values and their relative gap, degenerate flag, solver tolerance, and the validity statement conditional on the measured regime. Exemplar certificate (merge unit 0, from the released coefficient records): quadratic regime with measured curvature-to-tolerance ratio 26.99, two cutting-plane iterations, primal value 7.148742, dual value 7.148741, relative gap 1.4×10^{-7} (below the $\gamma = 10^{-4}$ termination tolerance, not exactly zero), non-degenerate, validity conditional on the measured regime.

Certificate census. All 48 merged attention units (maximin objective, raw reduction) resolved in the quadratic regime and none degenerate; measured curvature-to-tolerance ratios span 11.0–66.7 and every unit's relative primal–dual gap at termination is below the $\gamma = 10^{-4}$ tolerance (not exactly

Table 1: Main comparison on the dense `gemma-4-12B` expert suite; every per-task cell is CI-backed under the preregistered shared-index bootstrap. Worst is the min over the three tasks; Mean their average. Bold marks the column-best value: MinimaxMerge holds the best Worst and Mean entries; MagMax is best on GSM8K.

Method	GSM8K	Code	ARC-C	**Worst**	Mean
Best individual expert (oracle)	0.7233	0.4849	0.8976	0.4849	0.7019
Task arithmetic	0.7513	0.5286	0.9002	0.5286	0.7267
Uniform soup	0.7513	0.5286	0.9002	0.5286	0.7267
TIES-Merging	0.7225	**0.5497**	0.8933	**0.5497**	0.7219
DARE	0.7066	**0.5497**	0.8973	**0.5497**	0.7179
MagMax	**0.7695**	0.5377	0.9019	0.5377	0.7364
MinimaxMerge (ours)	0.7612	**0.5497**	**0.9036**	**0.5497**	**0.7382**

Table 2: Ablations on the dense suite. At the same per-layer solver, budget, and reduction, swapping the maximin objective for a mean objective costs 1.66 pt of worst-task accuracy and lowers the mean ($0.7382 \rightarrow 0.7280$); per-tensor granularity attains the best worst-task cell (0.5587, +0.76 pt over global; paired CI half-width 1.05 pt, ordering not resolved) while global keeps the best mean (0.7421, 0.02 pt above per-tensor's 0.7419). All cells are fixed arm values from the same controlled bake-off run, scored per item under the frozen protocol.

Axis	Arm	Worst task	Mean
Objective	**maximin (ours)**	0.5497	0.7382
Objective	mean (control)	0.5331	0.7280
Granularity	global	0.5512	**0.7421**
Granularity	per-tensor	**0.5587**	0.7419
Reduction	raw deltas	0.5497	0.7382
Reduction	TIES-reduced	0.5407	0.7271

zero at any unit), so the LP duality-gap guarantee stays conditional on the measured quadraticity regime on every unit; per-unit records in the released coefficient manifests.

D Full Track Tables

EXP1/EXP2 read-outs (Table 4). Table 4 is the preregistered OLMo-2-1124-13B skeleton (§5), reported in full and read under the same frozen preregistered gate with half-width > 1.0 pt reporting *not resolved*. EXP1's measured verdict — the full-roster worst-task retention ordering, led by TIES 0.9672 with every paired MinimaxMerge contrast not resolved at the 1.0-pt gate — and EXP2's measured contrast — stacked $0.9622 \leq$ TIES 0.9672, *not complementary* — are argued at §6.5 and §6.6, read against the two certified family snapshots (Gemma dense worst-first, Table 1; SmolLM3 mean-first divergence, §7) and mirroring the certified Gemma reduction-stack ordering (TIES-reduced maximin $0.5407 <$ TIES 0.5497, Table 2).

E Conflict-Gradient Track: Setup and Preregistered Rule

The high-conflict track of §6.3 preregisters six subsets of the eight certified `gemma-4-12B` specialists, spanning the measured surrogate-Gram conflict axis (mean pair/triple conflict; preregistered memberships): "L2" = WinoGrande+OpenBookQA (0.2440), "L3" = WinoGrande+OpenBookQA+MMLU-aux (0.3831, the minimum measured triple), "M2" = WinoGrande+MMLU-aux (0.4588), "M3" = ARC+MBPP+OpenBookQA (0.4977, the measured median triple), "H3" = GSM8K+ARC+MBPP (0.4998; the minimax/mean cells come from the certified stages and are reused, never re-run), and "H2" = GSM8K+MBPP (0.4999, the maximum measured pair conflict). Every set merges the identical certified mean-objective (mean per-layer/raw) and maximin (per-layer/maximin/raw) arms with MagMax as a recorded context arm; zero new specialists are trained. The certified warm-up set (L2, fully measured in the certified warm-up evaluation) reads in the expected direction at the low-conflict anchor: worst-task (minimax$-$mean) delta $+0.62$ pt, 95% CI $[-2.0, +3.4]$ — directional, below resolution.

Table 3: MoE track on `gemma-4-26B-A4B` (three task experts from the same base, attention q/v target space; bold marks the column-best value). On raw worst-task accuracy MinimaxMerge leads all three baselines (0.6476 vs. TIES 0.6355, MagMax 0.6310) and the mean (0.7963); on worst-task *retention* (normalized by the oracle row, §6.4) TIES (1.0225) narrowly edges out MinimaxMerge (1.0196), since TIES's binding task on this track is GSM8K rather than code. DARE is not evaluated on this track (compute-budget constraint, §7). Cells come from the frozen evaluation chain at official test-split counts (gsm8k 1,319 / code 664 / arc 1,172; non-degenerate scores, every cell verified before aggregation). Worst and Mean follow Table 1.

Method	Target space	GSM8K	Code	ARC-C	**Worst**	Mean
Best individual expert (oracle)	—	0.7741	0.5768	0.9138	0.5768	0.7549
Task arithmetic	attn q/v	**0.8150**	0.6235	0.9300	0.6235	0.7895
MagMax	attn q/v	0.8059	0.6310	0.9309	0.6310	0.7893
TIES	attn q/v	0.7915	0.6355	0.9360	0.6355	0.7877
MinimaxMerge (ours)	attn q/v	0.8097	**0.6476**	**0.9317**	**0.6476**	**0.7963**

Preregistered decision rule: a set resolves its per-task contrast iff the 95% paired bootstrap CI excludes zero *and* the delta magnitude clears the certified $1.96\sigma_{\text{item}}$ noise floor (gsm8k 1.74, code 2.59, arc 0.853 pt; pooled 1.071 pt at n=2195); the conflict-gradient verdict is the Spearman ρ of the six per-set worst-task deltas against their measured set conflicts with one-sided exact permutation significance (720 orderings at n=6, $p < 0.05$); the cross-bin point check $\delta(\text{H3}) > \delta(\text{L3})$ is recorded but not gated; a flat null records ρ, the permutation p, and the covered conflict interval, attributing not-resolved vs. method-failure vs. infeasible against the certified floors. The conflict-gradient verdict is filled in-line at §6.3, bound to the certified five-set conflict-gradient evaluation; the sixth preregistered set (SET5) is disclosed qualitatively as the registered arm-intrinsic collapse classification, so the preregistered 6-set row does not resolve.

F EXP1/EXP2 Analysis Scaffolds (Registrations Fixed in Advance, Read from the Certified Evaluation)

Analysis scaffold (a): the cross-family generalization frame (registration fixed in advance, read from the certified evaluation). Two registered read-outs compare the EXP1 measurements against the two certified family snapshots; the registrations were fixed in advance, and both answers are now read from the measured EXP1 rows, both labeled not resolved. (i) *Per-layer interference margins*: the OLMo three-specialist task-vector Gram, computed with the same pair-inner machinery as the certified snapshots, read as per-layer pairwise cosine overlap and sign-conflict structure against the certified Gemma dense and SmolLM3 LoRA-r=64 geometries. (ii) *Worst-task retention deltas*: the arm-level EXP1 contrasts of Table 4, under the frozen CI gate, aligned arm-by-arm with the certified orderings (Gemma: MinimaxMerge worst-best; SmolLM3: divergence, mean objective first). The registered generalization question is whether family conditionality collapses onto one measured geometry condition — binding-task interference-margin magnitude predicting which family's min-over-task ordering holds — or remains family-conditional, as on SmolLM3. The two answers are read only from the measured EXP1 rows: the measured cross-family geometry frame is available — OLMo mean pairwise cosine 0.000601 and mean sign-conflict rate 0.499838 over 40 layers $\times$ 80 q/v modules — with the comparison rows olmo-minus-gemma and olmo-minus-smollm3 rendered by the same pair-inner machinery as the certified Gemma dense and SmolLM3 LoRA-r=64 snapshots; the geometry-to-retention link across Gemma, SmolLM3 and OLMo is written only from measured fields: geometry comparison rows olmo-minus-gemma, olmo-minus-smollm3; stacked WorstTaskRet(TIES+maximin)-vs-TIES delta $+0.2091$ pt (half-width 3.4017 pt, not resolved; this is the bootstrap-of-the-minimum statistic, not the point-estimate difference, per §6.6); raw MinimaxMerge-vs-TIES delta -0.9203 pt (half-width 2.7640 pt, not resolved); the registered generalization question stays registered-open: both contrasts are labeled not resolved at the 1.0-pt gate (an absent geometry field is a documented limitation, never invented).

Analysis scaffold (b): EXP2 intended complementarity and the preregistered decision tree (registration fixed in advance, read from the certified evaluation). EXP2 separates roles before combining them: TIES sign election and trim reduce interference first; the per-layer maximin solve

Table 4: OLMo-2-1124-13B replication (EXP1 full roster): all 18 preregistered rows, scored under the frozen protocol with the Worst/Mean column pair of the retention tables (per-task cells are not part of the EXP1 registration); every cell is a measured value under the frozen protocol. Worst is the min over the three tasks of accuracy retention vs. the best single expert on each task; Mean is the average task accuracy; bold marks the column-best value. TIES leads the roster (worst 0.9672); MinimaxMerge sits at 0.9542 (rank 8 of the 14 merge arms) with all paired contrasts not resolved at the frozen 1.0-pt gate (§6.5); task arithmetic and uniform soup are built as separate rows but yield the identical merge by construction, as in the dense suite.

Method	Worst	Mean
Base (no merge)	0.9532	0.6173
Specialist GSM8K	0.9224	0.6053
Specialist MBPP	0.9274	0.6029
Specialist ARC	0.9639	0.6303
MinimaxMerge (per-layer, raw)	0.9542	0.6224
Mean objective (control)	0.9662	**0.6270**
Maximin, global unit	0.9512	0.6221
TIES+maximin	0.9622	0.6254
TIES-Merging	**0.9672**	0.6231
Task arithmetic	0.9602	0.6256
Uniform soup	0.9602	0.6256
MagMax	0.9622	0.6237
DARE-mean (rep. 0)	0.9542	0.6202
DARE-mean (rep. 1)	0.9483	0.6181
DARE-mean (rep. 2)	0.9453	0.6176
DARE-TIES (rep. 0)	0.9542	0.6246
DARE-TIES (rep. 1)	0.9502	0.6187
DARE-TIES (rep. 2)	0.9642	0.6252

then optimizes coefficients on the elected reduced vectors (§5). Raw MinimaxMerge vs. TIES fixes the objective-only coupling; TIES+maximin vs. TIES under the same CI gate tests the combination. The preregistered decision tree reads the two measured orderings jointly, no branch favored by plan: **(1)** TIES+maximin $>$ TIES and raw MinimaxMerge $\leq$ TIES — the Gemma-mirror dissociation pattern, on which interference reduction and minimax optimization would read as complementary; **(2)** TIES+maximin $>$ TIES and raw $>$ TIES — a combined gain without the dissociation reading; **(3)** TIES+maximin $\leq$ TIES — the reduction stack is family-conditional and the observed ordering is reported as-is. The branch matching the measured state has filled, and the others close; each conclusion is gated identically, and the three measured branch conclusions report at §6.6.

Recorded widening conditions (MoE lane). On `gemma-4-26B-A4B` the mergeable target space is geometry, not method choice (§6.4); stacked expert adapters and routing-sensitive behavior are the recorded widening conditions.

SmolLM3 replication qualification (LoRA-deviation preregistration). The replication specialists are LoRA r=64 deltas ($s\,(BA)$, s=2.0), not full fine-tuning: the verdict is preregistered under that deviation, not a family-size refutation; worst-task ordering and surrogate validity stay family-conditional, confining cross-arm statements to that certified `gemma` configuration.

Non-claims enumeration (verbatim from §7). Claims are confined to the coefficient-selection objective: not a new interference-reduction mechanism, nor superiority of surrogate reconstruction over sign election, magnitude pruning, or drop-rate sparsification — those reductions stay reusable underneath the objective (§4.5), worst-task contrasts attributed by the same-solver swap control; no guarantees outside the linear-combination frame, no claims about individual experts beyond certified merge-candidate identity, no aggregate benefit beyond the preregistered no-regression guardrail.