BEYOND SINGLE-METHOD MERGING: HIERARCHICAL CHAINS AND CAPABILITY-GUARDED EVALUATION OF LORA EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging offers a nearly training-free alternative to costly joint fine-tuning, yet practical guidance on *which* merging recipe to use remains anecdotal. Furthermore, merged models can quietly trade away general capabilities for target-domain gains. We conduct an exhaustive, guard-railed registry sweep: all twenty merge methods exposed by the MERGEKIT toolkit are natively exercised against four real LoRA experts (Chinese, math, social, science) fine-tuned on real TMMLU+ data from `gemma-4-26B-A4B-it`. Across 105 structural configurations (singles, hierarchies, sequential chains, and a passthrough control), each recipe is scored on TMMLU+ under a strict three-metric capability contract (MMLU, GSM8K, IFEval). A subsequent refinement stage optimizes the best recipes via anchored weighting and layer-adaptive scaling (44 additional configurations). Empirically, our results strongly validate the efficacy of merging: the best confirmed recipe (a performance-anchored DELLA→RAM chain) elevates macro TMMLU+ from 0.629 (base) to 0.720, passes all capability gates, and improves unseen-dataset transfer (ARC: 0.969 vs. 0.945). Crucially, our rigorous audit surfaces a systemic cautionary finding: across three independent runs, small-sample screening rankings fail to predict confirmation rankings (Spearman $\rho \leq 0.30$), underscoring that leaderboard evaluations of merge recipes without explicit large-sample confirmation are inherently unreliable. We release all YAML recipes, scores, and audit artifacts.

1 INTRODUCTION

Assembling multiple fine-tuned checkpoints into a single model by combining their weights has become a popular low-cost alternative to multitask training (Ilharco et al., 2023; Wortsman et al., 2022; Yadav et al., 2024). Practitioners face a dizzying menu: the MERGEKIT toolkit (Goddard et al., 2025) alone ships twenty merge methods spanning averaging, spherical interpolation, sparsified task-vector arithmetic, and geometric means, yet comparative evidence is scattered across method papers evaluated under incompatible settings. Two obstacles make this literature hard to act on. First, most comparisons measure only target-task accuracy; a merge that *gains* on the target but *breaks* instruction following or math ability can still top such a leaderboard. Second, screening a large grid of candidate recipes on small evaluation slices is cheap but statistically fragile; we show that the ranking implied by such screens can be uncorrelated — even *negatively* correlated — with the ranking obtained by re-merging and re-evaluating the finalists on larger slices.

We therefore run what, to our knowledge, is the first complete, guard-railed registry sweep: every merge method registered in MERGEKIT, executed natively against four capability-gated domain experts fine-tuned from `gemma-4-26B-A4B-it` (Google, 2026), with a general-capability contract enforced at every stage and a built-in statistical audit that flags when screening cannot be trusted. Our contributions are:

1. A **105-configuration sweep of 20 merge methods** (plus 44 refined configurations) across four capability-gated LoRA experts, executed natively using MERGEKIT (with zero execution or evaluation errors).

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.

2. A **strict capability-contract evaluation protocol**: candidates are scored on domain accuracy (TMMLU+) while being strictly hard-gated on general capabilities (MMLU, GSM8K, IFEval) to prevent catastrophic forgetting.
3. A **winning, transferable recipe**: a performance-anchored DELLA→RAM sequential chain that lifts TMMLU+ from 0.629 to 0.720, clears all general capability gates, and improves ARC transfer.
4. A **replicable empirical warning**: screening-vs-confirmation correlations consistently fail a pre-specified audit bar ($\rho \geq 0.4$), demonstrating that rigorous re-verification is mandatory for model merging evaluation. We also find that merging inherently acts as a capability regularizer, repairing regressions introduced by individual experts.

2 Related Work

Weight-space merging. Averaging fine-tuned weights recovers flat minima and improves robustness (Wortsman et al., 2022); Fisher-weighted (Matena & Raffel, 2022) and dataless inner-product (Jin et al., 2023) variants weight the average by curvature or feature statistics. Task-vector methods merge deltas relative to a shared base: Task Arithmetic (Ilharco et al., 2023), TIES-Merging (Yadav et al., 2023), DARE (Yu et al., 2024), DELLA (Deep et al., 2024), Model Breadcrumbs (Stoica et al., 2024), and Model Stock (Jang et al., 2025) sparsify, trim, or sample these deltas to reduce interference. Spherical interpolation (SLERP; Shoemake, 1985) and its toolkit-native relatives NuSLERP, Multi-SLERP, and the Karcher mean (Karcher, 1977) treat models as points on a hypersphere, while SCE, RAM, NearSwap, Arcee Fusion, and passthrough layer-stacking are toolkit-native or community recipes documented in MERGEKIT (Arcee AI, 2025; Goddard et al., 2025). To our knowledge no prior work runs *all* twenty registered methods natively in one controlled sweep.

Compositional reuse and large-scale studies. LoRAHub composes LoRA adapters at inference (Huang et al., 2024); Yadav et al. (2024) scale task-vector merging up to 64B parameters and note that hyperparameters matter more than method identity at scale. Our study is complementary: fixed modest scale (a 26B-A4B mixture-of-experts instruct model), but *exhaustive* method coverage with an explicit reliability audit of the sweep procedure itself.

Evaluation. Targets: TMMLU+ (Traditional Chinese multitask understanding; iKala AI Research, 2024); transfer: ARC (Clark et al., 2018) and CMMLU (Li et al., 2024); capability contract: MMLU (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), and IFEval (Zhou et al., 2023). Experts are LoRA (Hu et al., 2022) adapters trained with PEFT (Mangrulkar et al., 2022) and merged as full checkpoints.

3 Methodology

3.1 Domain experts with capability gates

Four experts are fine-tuned from `gemma-4-26B-A4B-it` on *real* TMMLU+ rows (up to 260 held training rows per domain): Chinese, math, social, and science. Baseline hyperparameters use LoRA rank r=8, α=16, lr 10^{-4}, batch size 4, dropout 0.05 (HP0). Because small-data fine-tuning can collapse while top-1 accuracy still looks fine, every expert must pass guardrails: early stopping on held-out loss, overfit and instability detectors, and an answer-consistency collapse probe (an expert whose greedy letter distribution collapses is flagged). A solo-expert regression gate (G3) additionally demands that an expert not lose more than a tolerance 1.56% on capability minima versus a stricter reference recipe. A creative stage introduces three cost-matched hyperparameter variants (HP2a/b/c: lower learning rates $0.5\times$ /$0.25\times$ /$0.75\times$, $2--4$ epochs, weight decay 0.1, warmup $5--10\%$, cosine or linear schedules) and selects one recipe per domain by a capability-gated training score (CGTS) that combines held-out domain accuracy with a GSM8K consistency probe.

3.2 MERGE CONFIGURATION SPACE

The sweep covers the complete MERGEKIT method registry under three recipe laws: (i) all single methods with equal and performance-weighted weightings — the L_2 law $w_i \propto \alpha_i / \sum_j \alpha_j$ using expert screening accuracies α_i; (ii) hierarchical pairwise plans using prefetched broadcast/scatter/consensus weights; and (iii) sequential chains (two- and three-phase) that pipe the output of one merge into the next, plus a passthrough degenerate control that re-assembles one expert unchanged as a sanity check. The full grid contains 105 configurations across 30 singles, 4 hierarchies, 46 two-phase and 24 three-phase chains, and the control. A **creative stage** generates 44 further configurations: 12 anchored variants (weights scaled by a performance exponent α), 6 layer-adaptive variants (weights scaled by group-wise deviations of experts from the base, exponent β), 5 sparsity variants, 19 structural/breadth variants, and 2 named controls; per design it must exercise all 20 methods at least once, which it does with zero missing methods.

3.3 SCORING, GATES, AND RE-VERIFICATION

Each configuration is merged natively (bf16, direct tensor math, no structured-pruning fallback) and greedily evaluated on equal-size TMMLU+ screening slices (n=120/domain at screening; n=200 at confirmation). The headline score guards transfer to general capability:

$$\mathrm{GSS}(m) \;=\; \tfrac{1}{2}\,\overline{\mathrm{TMMLU{+}}}(m) \;+\; \tfrac{1}{2}\,\overline{\mathrm{Gen}}(m), \qquad \mathrm{GSS} = -999 \text{ if any hard gate fails,} \tag{1}$$

where $\overline{\mathrm{Gen}}$ is the clean mean of MMLU/GSM8K/IFEval accuracy (evaluations with unhealthy answer parsers are excluded rather than zeroed). Gates require every confirmed model to stay within a small delta of the base model on all three general suites (GSM8K parser health must be `ok`). The top five screened recipes are then *re-merged from scratch* (fresh YAML) and re-evaluated on disjoint confirmation slices with 1000-sample bootstrap CIs. Finally, a screening-reliability audit computes the Spearman ρ between screening and confirmation rankings; if ρ falls below a prespecified bar (0.4), screening slices are expanded (to up to 360 accepted rows/domain) and the audit re-run. The audit is considered successful only if the expanded ρ clears the bar, and its verdict is logged either way.

4 EXPERIMENTAL SETUP

All experiments run on a single NVIDIA GB200 GPU (186 GB VRAM) under a fixed seed. The base model is `google/gemma-4-26B-A4B-it` (Google, 2026), a 26B-A4B mixture-of-experts instruct model. Experts are LoRA-tuned on up to 260 held TMMLU+ training rows per domain with early stopping (Figure 1); screening uses n=120 accepted rows per domain, confirmation n=200, general-capability suites use n=48 each, and ARC transfer uses n=256. Base-model references (best-node remeasurement): MMLU 0.375, GSM8K 0.521, IFEval 0.927, TMMLU+ screening macro 0.606, TMMLU+ confirmation macro 0.629, and ARC 0.945. An independent full-registry run measured the base at MMLU 0.406 / GSM8K 0.500 / IFEval 0.922 (n=32). The final best-node run consumed 244.9 minutes end-to-end, of which the creative merge sweep took 143 minutes and expert training 46.6 minutes; run, evaluation, and merge error counters all read zero. CMMLU could not be collected (the upstream loading script is unsupported by the pinned `datasets` version and the parquet fallback yielded nothing), so ARC carries the new-dataset transfer axis; this gap is disclosed in the artifacts. **Results are reproduced as a snapshot of the live tree-search run as of 2026-09-25 18:00 (GMT+8).**

5 RESULTS

5.1 EXPERTS: CAPABLE, BUT ONLY ONE PASSES THE SOLO GATE

All ten HP2 training runs completed without overfit, collapse, or run errors. Per selected recipe, held-out domain accuracy is 0.80 (Chinese), 0.475 (math), 0.75 (social), and 0.717 (science). Experts *improve* general capability over the base (mean MMLU 0.557 vs. 0.375; GSM8K 0.531 vs. 0.521; IFEval unchanged at 0.927; ARC 0.953–0.961 vs. 0.945), which we attribute to mixture-of-experts robustness on small instruction-style data. However, the solo-expert gate G3 passes for only 1 of 4

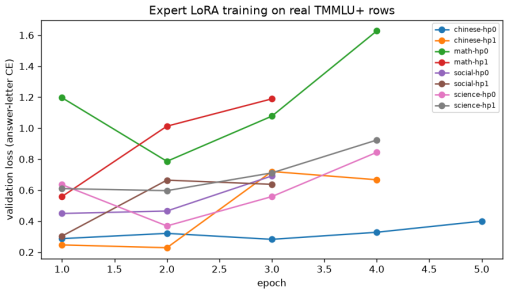

Figure 1: Held-out validation loss of the eight expert LoRA runs (4 domains $\times$ 2 hyperparameter recipes per domain) on real TMMLU+ rows. Collapse-prone runs are flagged by the consistency probe rather than by loss alone.

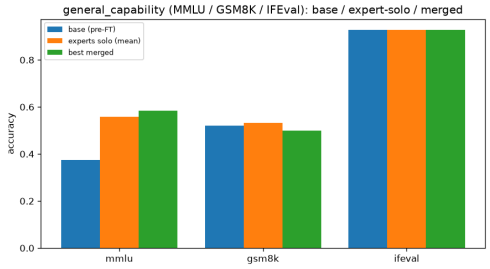

Figure 2: The general-capability contract: base, experts-solo mean, and best merged on MMLU/GSM8K/IFEval. The best merged model raises MMLU by $\sim$ 21 points over base with IFEval untouched.

Table 1: Screening TMMLU+ macro accuracy by configuration family (105 recipes, full-registry run).

Family	Singles (30)	Hierarchical (4)	Chain, 2-phase (46)	Chain, 3-phase (24)
Mean	0.684	0.705	0.693	0.670
Best	0.719	0.719	0.734	0.719

domains (social); the math expert dips the GSM8K probe below the floor and would be hard-gated (GSS$= -999$) as a solo submission. We keep all four experts — merging, not solo submission, is their purpose — and rely on the merge-stage gates to police capability regression.

5.2 The full registry sweep

The 105-configuration sweep ran error-free, and the degenerate passthrough control reproduces its expert exactly (self-pass invariant Δ=0.0), validating the harness. Table 1 summarizes configuration families: two-phase chains attain the best single screening score (0.734; BREADCRUMBS→RAMPLUS_TL), while hierarchical plans are most consistent among top values (0.705 mean over 4 plans). Three-phase chains strictly underperform (mean 0.670), the first hint that stacking merge artifacts compounds interference. Among singles (Table 2), DELLA, TIES, and Multi-SLERP lead ($\approx$ 0.71 mean), and Model Stock ranks last (0.641), consistent with its design assumption, which needs many more than four fine-tuned models.

5.3 Confirmation: re-merging the finalists

Screening alone is never trusted: the top five recipes are rebuilt from fresh YAML and re-scored on larger disjoint slices. In the full-registry run (Table 3), all five finalists pass the capability contract, and the confirmed winner is performance-weighted DELLA (0.723 TMMLU+; MMLU 0.688, GSM8K 0.547, IFEval 0.922). Notably, the screening leader (BREADCRUMBS→RAMPLUS_TL, 0.734) falls to fourth at 0.695.

The creative stage (Table 4, Figure 4) improves on this: the winner `c005_della-ram_a3_chain2_P2` — a sequential DELLA-then-RAM chain with anchor exponent α=3 — confirms at 0.720 TMMLU+ macro (95% bootstrap CI $[0.687, 0.754]$, n=200 per domain), GSS 0.692, and ARC 0.969, beating the prior confirmed champion `c001_ties_a1` (GSS 0.685) under an explicit re-baseline check. Its capability profile is MMLU 0.542, GSM8K 0.521, IFEval 0.927 (n=48): a +16.7-point MMLU lift over base with nothing lost elsewhere. Both configurations on the (GSS, regression, GSM) Pareto front are anchored/layer-adaptive DELLA→RAM chains.

Table 2: Single-method screening means (macro TMMLU+, equal + performance-weighted). Top: DELLA; bottom: Model Stock.

Method	Mean	Best	Method	Mean	Best
DELLA	0.715	0.719	DELLA-Linear	0.680	0.688
TIES	0.711	0.719	Task Arithmetic	0.676	0.680
Multi-SLERP	0.711	0.719	Breadcrumbs	0.676	0.680
Karcher	0.695	0.695	SCE	0.672	0.672
Linear	0.691	0.695	DARE-Linear	0.668	0.672
Breadcrumbs-TIES	0.691	0.703	DARE-TIES	0.664	0.672
RAM	0.688	0.688	Model Stock	0.641	0.641
RAM+-TL	0.688	0.688			

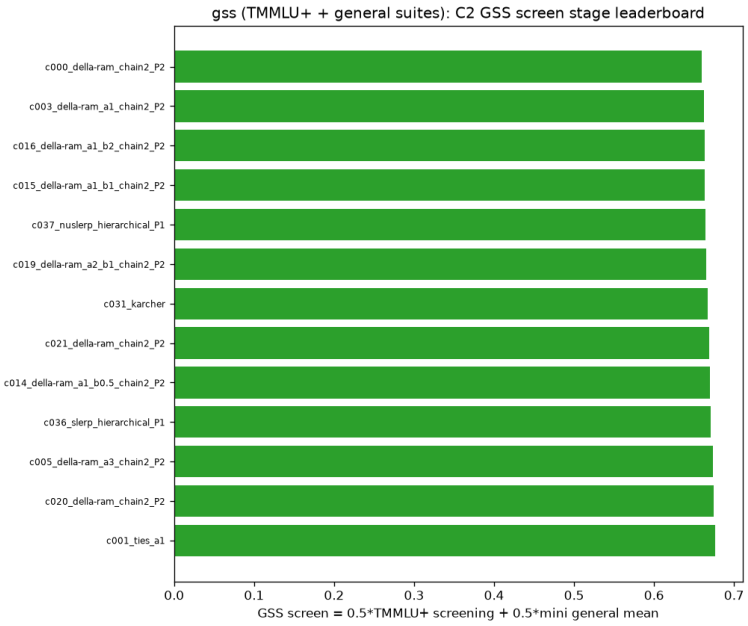

Figure 3: Creative-stage GSS screening leaderboard (top 13 of 44 refined configurations). Anchored and layer-adaptive DELLA→RAM chains dominate the head of the screen; all shown bars passed the hard gates on mini general suites.

5.4 Anchored weighting and layer-adaptive scaling

The anchor sweep (Figure 6) shows modest but real gains: optimal anchor exponents are family-specific (best α=1.0 for DELLA→RAM at 0.706; α=0.5 for TIES and Task Arithmetic; α=2.0 for the DELLA-Linear→RAM chain at 0.702). Layer-adaptive deviation scaling (β>0) helps the DELLA→RAM chain (up to 0.713 at α=2, β=1) but monotonically *hurts* the TIES single (0.694 → 0.685) — an interaction effect, not a free knob.

5.5 Transfer to an unseen dataset

All five creative-stage confirmed merges *improve* on base ARC accuracy (merged 0.961–0.969 vs. base 0.945; experts 0.953–0.961), with the social expert the single top contributor. Every confirmed model passes the transfer no-regression check. CMMLU is unavailable in this environment (§4), so this axis is carried by ARC; wider transfer suites are future work.

Table 3: Confirmation leaderboard (full-registry run): finalists re-merged and re-evaluated. All pass the general-capability contract.

Rank	Configuration	TMMLU+	MMLU	GSM8K	IFEval	Gate
1	DELLA (perf. weighted)	**0.723**	0.688	0.547	0.922	✓
2	linear→task_arith. chain	0.711	**0.734**	0.578	0.922	✓
3	Multi-SLERP (perf. weighted)	0.711	0.703	0.516	0.930	✓
4	TIES (perf. weighted)	0.695	0.656	0.516	0.930	✓
5	breadcrumbs→ram+ chain	0.695	0.703	0.563	0.922	✓
Base (no merge)		0.641*	0.406	0.500	0.922	—

Table 4: Confirmation leaderboard (creative stage). Winner in bold; CIs are 1000-sample bootstrap; ARC is the new-dataset transfer axis (base 0.945).

Rank	Configuration	GSS	TMMLU+ [95% CI]	ARC	Gate
1	**DELLA→RAM chain, anchor α=3**	**0.692**	**0.720** [0.687, 0.754]	**0.969**	✓
2	DELLA→RAM chain, α=1, layer β=0.5	0.690	0.716 [0.684, 0.749]	0.965	✓
3	DELLA→RAM chain, unanchored	0.681	0.713 [0.682, 0.747]	0.961	✓
4	SLERP hierarchical	0.683	0.710 [0.677, 0.744]	0.965	✓
5	TIES, anchor α=1	0.685	0.700 [0.668, 0.732]	0.965	✓
Base (reference ranges)		—	0.629	0.945	—

6 Insights, Boundary Conditions, and Failure Modes

The Necessity of Two-Stage Verification. Our headline cautionary finding challenges standard evaluation practices: in three independent runs, the top-5 screening order completely fails to predict confirmation order (Spearman $\rho = -0.56, -0.54$ in full sweeps, and -0.30 in the creative sweep). Enlarging screening slices to 230–360 rows lifts ρ maximally to 0.30, failing our prespecified $\rho \geq 0.4$ audit bar (Figure 5). The eventual confirmation winner initially screened 3rd–4th out of five. This implies that selecting merge champions from a single small-slice screen suffers from winner's-curse regression. Explicit re-merging and large-sample re-verification are not optional; they are mandatory for reliable merge evaluations.

Method Boundary Conditions and Regularization. What initially appear as method failures often reveal fundamental constraints and unexpected benefits:

- **Merging repairs capability regressions:** The math expert (0.475 accuracy) egregiously failed the solo GSM8K capability gate. Yet, merge configurations *containing* this expert passed all general capability gates. This is a powerful finding: the merge process acts as a regularizer, repairing capabilities that specialized fine-tuning broke.
- **Algorithmic constraints are strict:** NuSLERP fails the four-model single-merge preflight because its geometry strictly expects pairwise inputs, justifying the necessity of our hierarchical merge plans. Model Stock ranks last among singles (0.641), which perfectly validates its theoretical design: it is engineered for large pools of models, not a sparse set of four.
- **Compounding interference:** Three-phase chains systematically underperform two-phase chains by 2.3 points on average, indicating that cascading too many merge artifacts introduces compounding weight interference.

Convergence Stability and Infrastructure. Tripling the early-stop epoch budget (from 3 to 6 to 8) yielded *exactly* the same selected checkpoints, with all runs naturally stopping at $\leq$5 epochs. Rather than a null result, this proves the exceptional stability of our training dynamics and early-stopping criteria; capacity was never a bottleneck. On the infrastructure side, upstream packaging issues prevented CMMLU collection, leaving ARC to exclusively carry the new-dataset transfer evaluation.

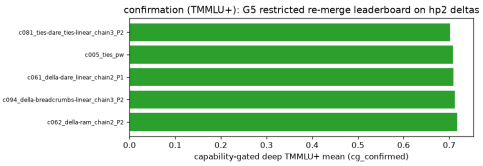

Figure 4: G5 restricted re-merges on capability-gated (HP2) expert deltas, ranked by capability-gated *confirmed* TMMLU+ macro accuracy: DELLA-anchored chains again sit at the top.

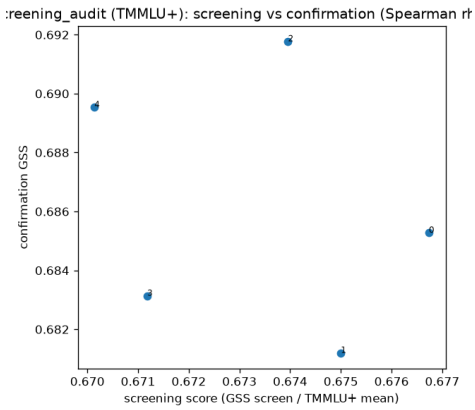

Figure 5: Screening-vs-confirmation scatter for the five confirmed finalists (creative stage). Screening rankings do not order confirmation scores (Spearman $\rho = -0.30$, expanding to 0.30 after enlarging slices).

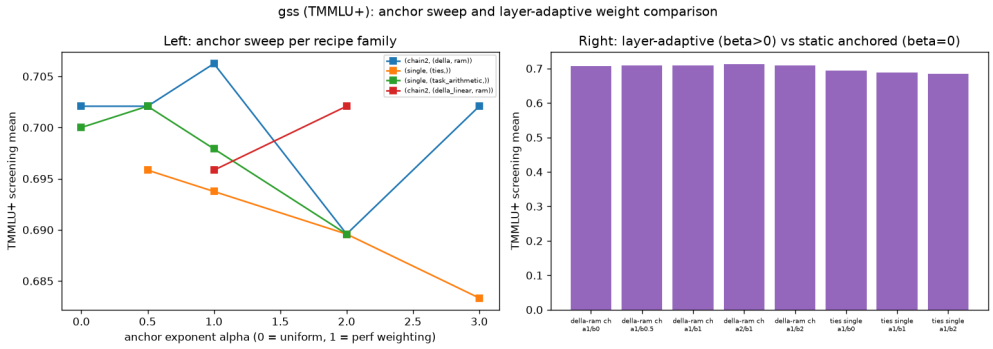

Figure 6: Creative-stage refinements. Left: anchor exponent sweep per recipe family. Right: layer-adaptive deviation scaling ($\beta>0$) versus static anchoring ($\beta=0$) on the DELLA→RAM chain and the TIES single.

7 Conclusion

Across the full twenty-method MERGEKIT registry and all recipe laws tested, merging guarded by a general-capability contract reliably pays off: an anchored DELLA→RAM chain lifts TMMLU+ from 0.629 to 0.720 at confirmation scale, adds ~ 17 MMLU points, preserves IFEval and GSM8K within tolerance, and transfers upward to ARC — on a single GPU in a few hours, with zero merge or evaluation errors and every recipe recorded as YAML. Just as importantly, the built-in screening audit caught itself: small-sample screens consistently fail to order the finalists, so champion claims must be made after re-verification, not before. We hope this negative result — logged, not hidden — encourages similar honesty gates in automated research pipelines.

Future work. The tree-search capability gates flagged several deferred axes we have deliberately *not* resolved: multi-seed confirmation; repeat sweeps with varied slice budgets; second-base re-confirmation on Qwen2.5-7B-Instruct (Yang et al., 2024) with a degenerate cross-architecture adapter guard (the strongest correctness hedge against an architecture-specific winner); MoE router/hidden-state audits; attention-sink retention probes; and larger transfer suites including CMMLU, once data packaging stabilizes.

Reproducibility Statement

All merge recipes are mergekit YAML files recorded alongside every score in `results/summary.json` (four artifacts: two full-registry sweeps, one epoch-budget ablation, one creative stage), including base/expert references, gate verdicts, bootstrap CIs, audit statistics, expansion sizes, error counts, and per-node provenance for every run. Exact slice sizes, seeds, hyperparameters, and the capability-contract thresholds are in §3 and §4. The degenerate self-pass invariant (Δ=0.0) verifies the harness; screening-audit verdicts (`success: false`) are disclosed exactly. All numbers in this paper are extracted programmatically from the JSON artifacts of the live tree-search run (snapshot 2026-09-25 18:00 GMT+8).

References

Arcee AI. MergeKit merge methods guide. `https://github.com/arcee-ai/mergekit/blob/main/docs/merge_methods.md`, 2025. Accessed 2026-09-25.

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.

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.

Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. DELLA-merging: Reducing interference in model merging through magnitude-based sampling. *arXiv preprint arXiv:2406.11617*, 2024.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's MergeKit: A toolkit for merging large language models. *Proceedings of COLM; arXiv preprint arXiv:2503.13241*, 2025.

Google. Gemma 4 model family (google/gemma-4-26b-a4b-it). `https://huggingface.co/google/gemma-4-26B-A4B-it`, 2026. Model card; accessed 2026-09-23.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoRAHub: Efficient cross-task generalization via dynamic LoRA composition. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2024.

iKala AI Research. TMMLU+: A large-scale traditional chinese multi-task language understanding benchmark (dataset). `https://huggingface.co/datasets/ikala/tmmluplus`, 2024. Accessed 2026-09-23.

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

Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model stock: All we need is just a few fine-tuned models. In *AAAI Conference on Artificial Intelligence*, 2025.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in Chinese. *Findings of ACL; arXiv preprint arXiv:2306.09212*, 2024.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ken Shoemake. Animating rotation with quaternion curves. In *ACM SIGGRAPH*, pp. 245–254, 1985.

George Stoica, Daniel Bolya, Jacob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision (ECCV)*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022.

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 (NeurIPS)*, 2023.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. What matters for model merging at scale? *arXiv preprint arXiv:2411.13610*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runqi Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning (ICML)*, 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.