PARENT-CONFLICT ARBITRATION FOR NEAR-TIED MODEL MERGES

Anonymous authors
Paper under double-blind review

ABSTRACT

A model merge carries a comparison instrument that ordinary checkpoint selection lacks: its parent checkpoints reveal where the specialized behaviors being composed conflict. The *Frozen Parent-Conflict Audit* (FPCA) turns this composition history into selective arbitration. Before candidate scoring, it measures conflict coverage and fixes answer-preserving neighborhoods, references, generators, and thresholds; it then compares neighborhood reference risk (NRR) only when the Global gap is at most 0.2 points and its paired interval contains zero. In a predeclared screen of 160 comparisons from Llama-3.1-8B-Instruct and Qwen-2.5-7B-Instruct task-vector collections, 46 pairs enter this regime and FPCA selects a different candidate from the Global point estimate in 30, including 16 of 27 construction-only near-ties. A separately authored evaluator spanning 24 stratified cases and 9,600 exact-key prompts confirms the decision consequence: FPCA lowers mean NRR from 15.2 to 12.9 (paired reduction 2.3; 95% bootstrap CI [1.4,3.2]) with 16/3/5 wins/ties/losses. Equal-budget slices isolate the parent-conflict signal in two detailed cases, while conflict sensitivity identifies risk-controlling blocks in all six model–regime settings and enables a 0.5%-residual refinement that improves NRR by 1.3–1.7 points over the strongest matched-budget non-conflict targeting controls. Parent behavior therefore supplies actionable selection information precisely where aggregate accuracy leaves a merge decision unresolved.

1 INTRODUCTION

The inputs to a model merge provide more than parameters to combine. They form an observable panel of parent behaviors: before any candidate merge is evaluated, the parents reveal prompts on which their specialized capabilities agree and prompts on which composition must resolve a conflict. Once evaluation compresses those prompts into a single benchmark average, this provenance is lost. Near-tied aggregate scores are therefore not merely uncertain rankings; they are decisions for which the construction history contains unused, candidate-independent information.

Existing merge operators—including Task Arithmetic, TIES-Merging, DARE, Fisher Merging, and adaptive methods (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024; Matena & Raffel, 2022; Ainsworth et al., 2023; Wang et al., 2024; Yang et al., 2024c)—optimize or report aggregate task performance. Generic behavioral tests can expose brittle outputs, but they do not identify which test regions correspond to capabilities that the merge was asked to compose. The resulting technical gap is a selection rule that (i) activates only when aggregate evidence leaves a pair unresolved, (ii) derives its comparison region from the actual parents rather than the candidates, and (iii) demonstrates that its choice transfers to newly authored examples.

FPCA fills this gap with an explicit decision sequence. It first applies a Global effect-size and paired-interval gate. Parent conflict then determines coverage and fixes anchors, answer-preserving neighborhoods, references, generators, and thresholds before either candidate is scored. An eligible pair is resolved by lower neighborhood reference risk (NRR); sparse-conflict settings trigger abstention, and globally separated pairs retain the Global decision. This separation of eligibility, evaluator construction, arbitration, and independent adjudication prevents candidate identity from determining which capabilities enter the comparison.

The evidence begins with construction-only prevalence and culminates in independent decision transfer. Among 96 comparisons between construction-only merge methods, 27 are Global near-ties

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.

and FPCA changes 16 choices; the full predeclared screen contains 30 different choices among 46 eligible pairs. A separately authored 24-case evaluator then measures the selected checkpoint on 9,600 new exact-key prompts, where FPCA reduces mean NRR by 2.3 points (95% CI [1.4,3.2]) and yields a 16/3/5 win/tie/loss profile. Two detailed Fisher/CGMR cases make the conditioning effect concrete: Global differs by 0.1–0.2 points, parent-conflict NRR differs by 3.1–3.7 points, and four equal-budget alternative slices reduce the separation to at most 0.4 points or change the local winner.

Technical contributions. FPCA is a parent-conditioned selective-decision object with distinct estimands for near-tie eligibility, choice disagreement, abstention, and independently measured gain. The finite 160-comparison screen and the separately authored transfer panel establish recurrence and decision quality across two backbone families and four answer-key suites. Conflict sensitivity then links the same behavioral signal to parameter space: it leads three non-conflict block rankings in all six settings, and Conflict-Guided Merge Refinement (CGMR) converts the ranking into a sparse matched-budget correction that improves NRR on both backbones. Together, these results show how model provenance can arbitrate and improve merges after Global evaluation has exhausted its resolution.

2 Related Work and Positioning

Merge construction and merge selection. Task Arithmetic, TIES-Merging, DARE, Fisher Merging, RegMean, ZipIt!, model soups, and adaptive merging define weight-space composition rules or optimize aggregate task performance (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024; Matena & Raffel, 2022; Jin et al., 2023; Bjorner et al., 2024; Wortsman et al., 2022; Ainsworth et al., 2023; Wang et al., 2024; Yang et al., 2024c). CABS removes coordinate overlap and balances sparsity during construction (Yang et al., 2025). SimMerge predicts operators and merge orders from task-agnostic weight and activation similarities, using parent-derived signals to avoid merge-and-evaluate search (Bolton et al., 2026). FPCA addresses the complementary post-construction decision: completed candidates have an unresolved aggregate ranking, and prompt-level behavior of their actual parents fixes both eligibility and the region on which reliability is compared. This combination of near-tie gating, parent-derived evaluation, ex ante freezing, abstention, and independent decision transfer defines the contribution.

Disagreement-based evaluation and composition-conditioned risk. Paraphrase tests, behavioral checklists, adversarial robustness, and safety red teaming probe finished models under input variation (Gan & Ng, 2019; Ribeiro et al., 2018; 2020; Madry et al., 2018; Cohen et al., 2019; Wei et al., 2023; Zou et al., 2023; Mazeika et al., 2024). Anchor Points selects representative examples from correlations across source-model predictions; DiffUse clusters semantic differences between candidate outputs; DISCO selects high-disagreement examples for efficient model ranking; and active model selection allocates labels to queries that distinguish finished models (Vivek et al., 2024; Ashury Tahan et al., 2024; Rubinstein et al., 2026; Durmazkeser et al., 2025). These methods optimize evaluation or annotation efficiency using source or candidate panels. FPCA instead uses the exact models being composed, before candidate inference, to identify capability conflicts; it couples that source signal to a near-tie action rule and validates the resulting checkpoint choice on a separately authored neighborhood family. Equal sample budgets in Table 3 compare the composition signal with high-loss, random, generic-paraphrase, and validation-loss alternatives.

Localization and intervention. Fisher importance, task-vector magnitude, localized merging, Sens-Merging, and Representation Surgery identify influential parameters or correct merge-induced representations, while sparse masks, adapters, and low-rank updates provide efficient adaptation mechanisms (Matena & Raffel, 2022; Wang et al., 2024; Liu et al., 2025; Yang et al., 2024b; Davari & Belilovsky, 2024; Pfeiffer et al., 2021; Hu et al., 2022; Liu et al., 2022). Conflict sensitivity is distinguished by its target: it ranks blocks by causal change in the parent-conflict neighborhood risk used for arbitration, rather than by task sensitivity, parameter magnitude, representation distance, or aggregate validation loss. CGMR follows that ranking under budgets matched to random, magnitude, validation-loss, and Fisher-guided targeting, separating the general FPCA selector from one mechanism for acting on its signal.

3 CONFLICT-CONDITIONED SELECTIVE ARBITRATION

Let $\mathbf{w}_0$ be a pretrained base model and $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K$ task vectors from independently fine-tuned parents. A merge rule produces $\mathbf{w}_m = \mathcal{M}(\mathbf{w}_0, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K)$ with Global score G_m. For candidates a, b, let C_{ab} be the paired 95% bootstrap interval for $G_a - G_b$. A predeclared tolerance ϵ defines the unresolved pair set

$$\mathcal{P}_\epsilon = \{(a, b) : |G_a - G_b| \leq \epsilon \ \wedge \ 0 \in C_{ab}\}. \tag{1}$$

We use $\epsilon = 0.2$ percentage points in the primary analysis. This threshold is smaller than the 0.2–0.4-point seed standard deviations observed for candidate merges, so membership requires both a small effect and paired statistical ambiguity.

3.1 PARENT-CONDITIONED ARBITRATION OBJECT

The held-out prompt pool is partitioned into disjoint *anchor-mining* and *task-validation* splits before candidate comparison. The mining split determines conflict coverage, anchors, neighborhood templates, generators, references, and thresholds exactly once; the validation split supplies only aggregate task loss for merge construction and hyperparameter selection. Each answer-key anchor receives eight variants formed by semantically matched paraphrases or minimal entity edits. Paraphrases retain the canonical key; an entity edit is retained only when its deterministic substitution updates the key unambiguously and the task-specific parser accepts both versions. Instruction-following and safety use matched rewrites and refusal-sensitive reformulations (Wei et al., 2023; Zou et al., 2023; Mazeika et al., 2024). Every candidate checkpoint is scored on the identical frozen set.

We score each prompt with

$$s(x) = \max_{i \neq j} \Delta\big(f_{\mathbf{w}_0 + \boldsymbol{\delta}_i}(x), f_{\mathbf{w}_0 + \boldsymbol{\delta}_j}(x)\big)\,, \tag{2}$$

For answer-key tasks, Δ is binary disagreement after the dataset's deterministic answer parser: option matching for multiple choice, numeric extraction for GSM8K, final-answer normalization with symbolic-equivalence checking for MATH, and normalized alias matching for Natural Questions and TriviaQA. For open-ended generations, Δ is normalized $1 - \mathrm{ROUGE\text{-}L}$ after common wrapper and whitespace normalization. Open-ended conflict uses a positive mining-split threshold γ_c; answer-key conflict uses $\gamma_c = 0$. Conflict coverage is

$$c = \frac{1}{|\mathcal{D}_{\mathrm{mine}}|} \sum_{x \in \mathcal{D}_{\mathrm{mine}}} \mathbb{1}[s(x) > \gamma_c]. \tag{3}$$

FPCA abstains when $c < 0.10$. Otherwise, with $\mathcal{C} = \{x : s(x) > \gamma_c\}$, it deterministically ranks $\mathcal{C}$ by $s(x)$, breaks ties by mining index, and freezes the top $\min(500, \lceil 0.30|\mathcal{C}| \rceil)$ anchors. All primary answer-key comparisons reach the 500-anchor cap. Section 5.6 evaluates anchor fractions from 10% to 50%.

For an anchor x and merged model $\mathbf{w}_m$, we define

$$\mathrm{NRR}(x, \mathbf{w}_m) = \frac{1}{|\mathcal{N}(x)|} \sum_{x' \in \mathcal{N}(x)} \mathbb{1}[\Delta(f_{\mathbf{w}_m}(x'), f_{\mathrm{ref}}(x')) > \gamma_r]\,, \tag{4}$$

where f_{ref} is the canonical key or accepted alias set in the primary experiments, the dataset assistant response for instruction following, and the dataset safe/refusal response for safety. The mining split fixes γ_r before evaluation; for answer-key tasks, $\gamma_r = 0$. We call this reference-failure rate *neighborhood reference risk* (NRR), and the setting-level risk $\mathrm{NRR}(m)$ averages Eq. 4 over anchors. Let $\hat{m}_G = \arg\max_{m \in \{a,b\}} G_m$, with ties broken by the predeclared candidate order. FPCA chooses

$$\hat{m}_{\mathrm{FPCA}} = \arg\min_{m \in \{a,b\}} \mathrm{NRR}(m), \qquad R_{ab} = \mathbb{1}[\hat{m}_G \neq \hat{m}_{\mathrm{FPCA}}]\,. \tag{5}$$

We report R_{ab} only for eligible $(a, b) \in \mathcal{P}_\epsilon$. On an independently frozen evaluator with risk $\mathrm{NRR}_{\mathrm{ind}}$, decision transfer is measured by

$$A_{ab} = \mathrm{NRR}_{\mathrm{ind}}(\hat{m}_G) - \mathrm{NRR}_{\mathrm{ind}}(\hat{m}_{\mathrm{FPCA}}). \tag{6}$$

Positive adjudication gain means the conflict-conditioned choice reduces reference failures on examples outside the evaluator used to make that choice. Equations 1–6 separate eligibility, selection change, and independent decision quality.

3.2 CONFLICT SENSITIVITY

The frozen anchors are divided by mining index into disjoint CS-estimation and localization-validation halves. For a decoder block ℓ, conflict sensitivity is the expected NRR increase on the estimation half under a fixed block perturbation ϵ_ℓ:

$$\mathrm{CS}_\ell = \mathbb{E}_{x \sim \mathcal{D}^{\mathrm{est}}_{\mathrm{anchor}}} \left[\mathrm{NRR}(x, \mathbf{w}_m \oplus \epsilon_\ell) - \mathrm{NRR}(x, \mathbf{w}_m) \right]. \tag{7}$$

We exclude embeddings and the LM head and inject zero-mean Gaussian perturbations rescaled to 1% of the block parameter norm. Two perturbation seeds estimate CS; three held-out seeds measure block-perturbation NRR on $\mathcal{D}^{\mathrm{loc}}_{\mathrm{anchor}}$. The same perturbation budget is used for every block and predictor. Spearman correlation between the estimated ranking and the disjoint-half outcome ranking evaluates localization. Fisher-diagonal importance, task-vector magnitude, and validation-loss increase provide fixed non-conflict comparators.

FPCA follows four ordered operations: form the Global near-tie set; verify conflict coverage and freeze the evaluator; arbitrate by NRR; then use CS for localization or refinement when desired. Table 4 estimates how often the decision changes, Table 5 measures adjudication gain, Table 3 isolates parent conflict as the conditioning signal, and CGMR tests whether that signal supports efficient intervention.

4 FROM BEHAVIORAL CONFLICT TO SPARSE REFINEMENT

FPCA applies to candidate merges independently of how they were constructed. CGMR is an actionability experiment: it uses the FPCA conflict structure as an optimization signal and asks whether that signal supports a sparse correction. The construction-only comparisons and separately authored adjudicator test FPCA as a selector; the CGMR studies test localization and intervention. Let $\widetilde{\boldsymbol{\delta}}_{k,\ell}$ denote task k's contribution to decoder block ℓ after TIES trimming, elected-sign masking, and coordinate-wise contributor normalization. This definition preserves TIES exactly:

$$\mathbf{w}_{\mathrm{TIES}} = \mathbf{w}_0 + \sum_{k,\ell} \widetilde{\boldsymbol{\delta}}_{k,\ell}. \tag{8}$$

Stage 1 computes CS on the frozen mining split and selects the six highest-ranked decoder blocks. Stage 2 uses $\mathbf{w}(\boldsymbol{\alpha}) = \mathbf{w}_0 + \sum_{k,\ell} \alpha^k_\ell \widetilde{\boldsymbol{\delta}}_{k,\ell}$, initialized at $\alpha^k_\ell = 1$, so the initial checkpoint is exactly Eq. 8. Stage 3 adds a residual restricted to the six selected blocks and projected to a 0.5% global parameter budget.

Hard NRR is an evaluation metric and is never differentiated. For a frozen neighbor x' and reference token sequence $y^\star$, define the length-normalized teacher-forced score $q_{\mathbf{w}}(y \mid x') = |y|^{-1} \sum_t \log p_{\mathbf{w}}(y_t \mid x', y_{<t})$. For finite-answer tasks, $\mathcal{Y}^-(x')$ is the frozen set of incorrect options and

$$\ell_{\mathrm{conf}}(x'; \mathbf{w}) = \tau \log \left(1 + \exp \left[\frac{\mathrm{LSE}_{y \in \mathcal{Y}^-(x')} \, q_{\mathbf{w}}(y \mid x') - q_{\mathbf{w}}(y^\star \mid x')}{\tau} \right] \right). \tag{9}$$

For open-answer and open-ended tasks, $\ell_{\mathrm{conf}} = -q_{\mathbf{w}}(y^\star \mid x')$ over the frozen canonical reference; accepted aliases are combined with log-sum-exp. Thus the train-time signal is differentiable through token probabilities, whereas decoded exact-key and ROUGE-based NRR remain post-training measurements. The optimization objective is

$$J_{\mathrm{CGMR}}(\mathbf{w}) = \mathbb{E}_{x \sim \mathcal{D}_{\mathrm{val}}} \mathcal{L}(x; \mathbf{w}) + \mu \, \mathbb{E}_{x, x'} \ell_{\mathrm{conf}}(x'; \mathbf{w}), \tag{10}$$

with $\mu = 0.3$, $\tau = 0.05$, x uniform over frozen anchors, and x' uniform over the eight neighbors of x. Each update contains 32 task-validation prompts and eight anchors with all eight neighbors. Coefficients use Adam for 200 steps at learning rate 5×10^{-2}; the residual uses 100 steps at 2×10^{-5}; both use $(\beta_1, \beta_2) = (0.9, 0.999)$, zero weight decay, and gradient norm clipping at 1.0. Greedy decoding uses one answer token for multiple choice and at most 256 new tokens for open-ended tasks.

Residual support is initialized from the largest-magnitude gradients inside the selected blocks. After every update, we retain the globally largest-magnitude $p|\mathbf{w}|$ residual entries within those blocks and zero the remainder; support is therefore recomputed at every step and the cardinality constraint holds

throughout optimization. The minimum Eq. 10 checkpoint on the fixed optimization schedule is selected, with earlier step breaking exact ties. Decoded NRR is computed only after checkpoint selection. The task-validation split continues to supply only the aggregate task-loss term. Seeds jointly fix data order, perturbations, and residual initialization.

Algorithm 1 Conflict-Guided Merge Refinement (CGMR)

Require: Base model $\mathbf{w}_0$, task vectors $\{\boldsymbol{\delta}_k\}_{k=1}^{K}$, held-out prompt pool $\mathcal{D}$, residual budget $p = 0.005$

Produce baseline merge $\mathbf{w}_m^0$ with TIES-Merging
Split $\mathcal{D}$ into disjoint anchor-mining and task-validation subsets
Compute conflict coverage with Eqs. 2–3; proceed when FPCA is eligible
Rank conflict prompts; generate and freeze neighborhoods $\{\mathcal{N}(x)\}$ for the selected anchors
for each decoder-layer block ℓ **do**
    Compute conflict sensitivity CS_ℓ with Eq. 7
end for
Form TIES-masked contributions $\{\widetilde{\boldsymbol{\delta}}_{k,\ell}\}$; initialize every $\alpha_\ell^k = 1$
for $t = 1, \ldots, 200$ **do**
    Update $\{\alpha_\ell^k\}$ with Adam on Eq. 10
end for
Construct $\mathbf{w}_m = \mathbf{w}_0 + \sum_k \sum_\ell \alpha_\ell^k \widetilde{\boldsymbol{\delta}}_{k,\ell}$
Fix the selected set of six highest-CS blocks; initialize residual $\mathbf{r}$ in those blocks
for $t = 1, \ldots, 100$ **do**
    Update $\mathbf{r}$ on Eq. 10; project to the largest $p|\mathbf{w}_m|$ entries
end for
return $\mathbf{w}_m + \mathbf{r}$

5 Experiments

5.1 Evaluation design

The experiments revolve around three questions. *RQ1:* How often does parent-conditioned arbitration change an unresolved Global choice, including among construction-only merges? *RQ2:* Does that choice reduce reference failures on separately authored neighborhoods, and what signal drives the detailed decisions? *RQ3:* Do parent-conflict locations identify parameters that can be corrected under a sparse budget? The predeclared comparison frame answers RQ1, independent adjudication and fixed-budget slices answer RQ2, and localization plus matched-budget refinement answer RQ3.

The answer-key study uses Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen-2.5-7B-Instruct (Yang et al., 2024a). For each backbone, two disjoint specialist-training collections produce three task vectors apiece. Four suites probe knowledge/reasoning (MMLU, HellaSwag, ARC-Challenge (Hendrycks et al., 2021a; Zellers et al., 2019; Clark et al., 2018)), commonsense (CommonsenseQA and PIQA (Talmor et al., 2019; Bisk et al., 2020)), mathematical reasoning (GSM8K and MATH (Cobbe et al., 2021; Hendrycks et al., 2021b)), and factual QA (Natural Questions and TriviaQA (Kwiatkowski et al., 2019; Joshi et al., 2017)). Global is the macro-average of component-dataset scores after the task-specific parsing contract above. Each suite contains ten predeclared method pairs in each of the four backbone/collection settings, for 40 comparisons per suite. The schedule includes all six pairings among Task Arithmetic, TIES-Merging, DARE, and Fisher Merging, plus CGMR and Fisher-guided matched-budget control comparisons against TIES and Fisher. Merge hyperparameters and the complete pair order are fixed from task-validation Global scores before NRR inspection.

Each candidate uses three seeds. Equation 1 uses $\epsilon = 0.2$ and a paired 95% bootstrap interval obtained by resampling shared evaluation questions within each comparison. FPCA uses 500 anchors and eight neighbors, giving 4,000 scored prompts per seed and 12,000 per candidate; pairwise differences use the same 12,000 prompt instances, corresponding to 24,000 model outputs per pair. Wilson intervals accompany the different-choice rates as summaries of this fixed scheduled frame; they are not used to claim independent draws from a broader population. Paired case bootstraps resample cases for

Table 1: Decision design and evidence roles. Every object in the "fixed from" column is determined before the candidate outcome named in the final column is inspected.

Stage	Specification	Fixed from	Outcome
Merge development	Candidate hyperparameters and pair order	Task-validation Global scores	Candidate checkpoints
Global eligibility	$\lvert G_a - G_b \rvert \leq 0.2$ and $0 \in C_{ab}$	Shared evaluation questions	Near-tie set
FPCA arbitration	Coverage ≥ 0.10; up to 500 parent-conflict anchors; eight variants	Parent outputs and mining split	Lower-NRR candidate
Independent transfer	Six cases per suite; 100 edited anchors and four variants per case	Blinded editors and fixed case schedule	NRR of selected checkpoint
Sparse intervention	Six CS-ranked blocks; 200 coefficient and 100 residual steps; 0.5% residual	Frozen FPCA neighborhoods and task-validation loss	CGMR actionability

Table 2: Two detailed answer-key near-ties. Cells list Fisher/CGMR; raw failures aggregate three seeds.

Case	Global ↑	NRR ↓	Raw failures / candidate ↓
MT-Llama	65.4±0.3 / 65.3±0.3	15.8±0.9 / 12.1±0.7	1,896 / 1,452
MT-Qwen	66.1±0.3 / 65.9±0.2	14.9±0.8 / 11.8±0.6	1788 / 1416

NRR differences, and the sign test excludes ties. Raw counts are three-seed per-candidate aggregates unless a caption identifies the independent one-checkpoint decision evaluation.

The independent case list is fixed at six near-ties per suite from the predeclared 46-pair frame before any manual editing. Editors receive the original prompt, canonical key, and parent outputs, but not candidate identities, candidate outputs, the Global winner, or the FPCA winner. They create 100 answer-preserving anchors with four variants per case; a second pass verifies each key before candidate inference. The checkpoint seed attached to each case is fixed by the original pair schedule, so each selector makes one candidate choice and is scored on the same 400 prompts, totaling 9,600 exact-key evaluations. The estimand is the unweighted mean over this suite-balanced 24-case panel. The cases, editors, variants, and thresholds are isolated from automatic anchor mining, CS estimation, CGMR optimization, and the original FPCA decision.

The localization study crosses the two backbones with multi-task (MT), instruction-following (IF), and safety/refusal (SF) settings. IF specialists derive from assistant-style corpora (Ouyang et al., 2022; Köpf et al., 2023); SF combines helpfulness and safety specialists (Bai et al., 2022). Exact-key answer-key suites provide the primary arbitration evidence; open-ended branches evaluate breadth under common normalized-reference scoring. All runs use PyTorch 2.3 and CUDA 12.4 on 8 NVIDIA A100-80GB GPUs. CGMR takes about 25 minutes for an 8B model (1.6–1.7× Task Arithmetic); full post-merge fine-tuning updates all parameters for the same 300 steps.

5.2 Composition-specific signal in two detailed decisions

These two cases are optimization-aligned actionability tests for CGMR; independent selector transfer is evaluated separately in Table 5. Fisher has the larger Global point estimate, but the effect-size-and-interval criterion leaves both pairs unresolved. On the parent-conflict neighborhoods, CGMR removes 148 failures per seed for MT-Llama (444 over 12,000 shared instances) and 124 per seed for MT-Qwen (372 in aggregate).

Table 3 changes only the conditioning signal. Every source list and tie-break is frozen before candidate inference: parent conflict uses Eq. 2, while the controls use their predeclared high-loss ranking, random order, generic-paraphrase source list, or task-validation-loss ranking. Candidate outputs are excluded from construction of all five slices, which share the same 500-anchor, eight-variant, reference, threshold, and seed contract. The alternatives reduce the 3.1–3.7-point separations to at

Table 3: Fixed-budget slice comparison. ΔNRR is Fisher minus CGMR; positive values favor CGMR. Every slice uses 500 anchors, eight variants, common thresholds, and three seeds.

Slice	MT-Llama: Δ / winner	MT-Qwen: Δ / winner
Parent-conflict neighborhoods	3.7 / CGMR	3.1 / CGMR
High-loss prompts	0.4 / CGMR	0.3 / CGMR
Random neighborhoods	-0.2 / Fisher	-0.3 / Fisher
Generic paraphrases	0.3 / CGMR	-0.2 / Fisher
Validation-loss slice	0.2 / CGMR	0.2 / CGMR

Table 4: Finite predeclared comparison frame. "Different" counts FPCA choices that differ from the Global point estimate; 95% CIs are shown for suite and total rates. Median [IQR] is absolute NRR separation; each candidate uses 12,000 shared prompt instances.

Suite	Screened	Near-ties	Different	Rate [CI]	Median [IQR]
Knowledge/reasoning	40	13	9	69.2 [42.4,87.3]	2.8 [1.7,3.7]
Commonsense	40	11	7	63.6 [35.4,84.8]	2.3 [1.2,3.4]
Mathematical reasoning	40	9	6	66.7 [35.4,87.9]	2.6 [1.4,3.1]
Factual QA	40	13	8	61.5 [35.5,82.3]	2.1 [1.0,3.0]
Construction-only	96	27	16	59.3%	—
Refinement/control	64	19	14	73.7%	—
Total	160	46	30	65.2% [50.8,77.3]	2.4 [1.3,3.3]

most 0.4 points; random neighborhoods choose Fisher in both cases and generic paraphrases choose Fisher on MT-Qwen. Appendix Table 14 reports the paired method values and counts. Fixed evaluator size and candidate-independent construction isolate the parent-conflict conditioning signal in these cases.

5.3 Construction-only and full-frame selection changes

Construction-only comparisons establish the phenomenon before conflict-guided intervention enters the candidate set: 27 of 96 comparisons are near-ties, and FPCA chooses differently in 16 (59.3%). The refinement/control stratum contributes 14 different choices among 19 near-ties. In the complete schedule, 46 of 160 comparisons enter the near-tie set and FPCA differs from the Global point estimate on 30; every suite has a 61.5–69.2% difference rate, with 2.4 median NRR points between candidates. Across those 30 cases, the lower-NRR candidates avoid 9,330 reference failures: 3,187, 2,041, 1,965, and 2,137 by suite.

5.4 Independent neighborhoods test the selected model

The independently authored neighborhoods change both examples and editor while preserving the task key. FPCA selection reduces mean NRR by 2.3 points relative to Global selection (Eq. 6; bootstrap 95% CI [1.4,3.2]) and by 1.6 points relative to the validation-loss selector. Against Global, the case profile is 16 wins, 3 ties, and 5 losses (two-sided sign test $p = 0.027$); the selected checkpoints incur 221 fewer failures. The gain therefore concerns the selected model, not reuse of the automatic neighborhood score.

5.5 Conflict sensitivity recovers a parameter-space signature

CS has the highest correlation in all six settings, with 0.27–0.28 gains over the strongest fixed comparator and intervals excluding zero. The disjoint anchor halves, two estimation seeds, three held-out perturbation seeds, and leave-one-block-out analysis preserve the ordering for 26/28 to 31/32 removals. Parent conflict therefore leaves a repeatable parameter-space signature across independent localization records and block deletion.

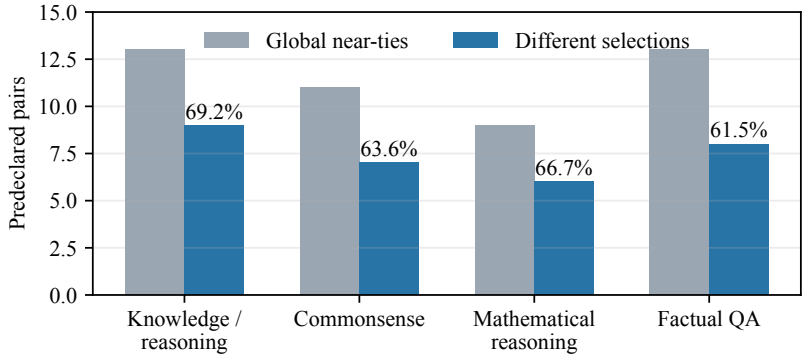

Figure 1: Predeclared answer-key screen. Near-ties/different selections are 13/9, 11/7, 9/6, and 13/8; FPCA differs from the Global point estimate on 65.2% of eligible pairs.

Table 5: Separately authored, suite-balanced exact-key adjudicator over 24 near-tie cases. Each selector chooses one candidate per pair and is scored on 9,600 prompts; standard deviation is across cases.

Selector	Independent NRR ↓	Across-case std	Raw failures ↓
Global point estimate	15.2	3.4	1,459
Predeclared validation-loss slice	14.5	3.1	1,392
FPCA	<u>12.9</u>	2.7	<u>1,238</u>

5.6 The arbitration rule is stable across operating points

At $\epsilon \in \{0.1, 0.2, 0.3\}$, FPCA changes 19/28 (67.9%, CI [49.3,82.1]), 30/46 (65.2%, CI [50.8,77.3]), and 40/67 (59.7%, CI [47.7,70.6]) Global point-estimate choices. The gradual decline is consistent with admitting more clearly separated Global comparisons. Crossing anchor fractions $\{10, 20, 30, 40, 50\}\%$ with 4, 8, or 12 neighbors preserves the default winner in 13/15 MT-Llama cells, 12/15 MT-Qwen cells, and 559/690 (81.0%) case-by-configuration cells overall. Appendix Table 9 shows complementary abstention and decisive-Global controls.

5.7 Conflict-guided refinement traces an efficient frontier

Table 7: MT-Llama corrective-budget frontier. At $p = 0$, CGMR optimizes coefficients without a residual. Uncertainty is standard deviation over three seeds; overhead is relative to Task Arithmetic.

Residual budget	Global ↑	NRR ↓	Overhead
0%	65.1±0.4	14.8±1.0	1.20±0.04×
0.1%	65.2±0.4	13.6±0.9	1.31±0.05×
0.5%	65.3±0.3	12.1±0.7	1.62±0.06×
1.0%	65.7±0.3	11.4±0.6	1.91±0.08×
5.0%	<u>66.4±0.2</u>	<u>10.2±0.5</u>	2.79±0.11×
Full post-merge FT	66.2±0.3	13.4±0.8	8.4×

The 0.5% operating point captures 2.7 of the 4.6 NRR points available between coefficient-only and 5% refinement while retaining a 1.62× merge-time footprint. The 5% point defines a second Pareto regime: it reaches 66.4 Global and 10.2 NRR at one third of full-fine-tuning wall clock. Matched-budget controls on both backbones (Appendix Tables 11–12) and component ablations (Table 8) attribute the 0.5% benefit to conflict-guided targeting, the differentiable conflict surrogate, and sparse correction.

Table 6: Spearman correlation with downstream block-perturbation NRR. CS is estimated on one anchor half with two seeds and evaluated on the disjoint half with three held-out seeds. Δ compares CS with the strongest non-conflict predictor; brackets are paired 95% bootstrap CIs from 1,000 stratified anchor bootstraps. LOO reports leave-one-block-out retention.

Setting	Val. loss	Fisher	Magnitude	CS	Δ [95% CI]	LOO
MT-Llama	0.41	0.39	0.34	0.69	0.27 [0.15,0.40]	31/32
MT-Qwen	0.45	0.42	0.37	0.73	0.28 [0.16,0.40]	27/28
IF-Llama	0.38	0.35	0.32	0.66	0.28 [0.13,0.42]	30/32
IF-Qwen	0.42	0.40	0.35	0.70	0.28 [0.15,0.41]	26/28
SF-Llama	0.44	0.41	0.38	0.72	0.27 [0.14,0.40]	30/32
SF-Qwen	0.47	0.45	0.40	0.74	0.27 [0.14,0.40]	26/28

6 Implications

Model merging differs from ordinary checkpoint selection because the parent models expose the conflicts being composed. FPCA makes this information decision-relevant while preserving a clean temporal boundary: parent behavior determines eligibility and the comparison object before candidate-local results are available. The 59.3% different-choice rate among construction-only near-ties establishes the selector independently of CGMR, while the separately authored adjudicator measures transfer beyond the automatic neighborhoods. Second-generator and human-edited neighborhoods preserve the FPCA winner in both detailed cases (Appendix Table 17).

The demonstrated operating regime is aggregate near-ties with at least 10% parent-conflict coverage. Inside this regime, FPCA combines a selective action rule with an independently measured decision gain; outside it, sparse-conflict cases abstain and globally decisive comparisons remain with Global selection. The 81.0% agreement across the anchor-budget grid shows that this behavior is stable around the default operating point. The finite comparison frame reuses backbones, collections, and prompts, so its pair-level intervals quantify within-frame variation, while recurrence across every suite and both backbone families tests breadth within the evaluated design. Exact-key answer-key tasks carry the selection result; normalized-reference instruction and safety experiments corroborate the mechanism, with external-label agreement and a measured 0.4-point over-refusal cost delimiting that extension.

7 Conclusion

Parent checkpoints are an endogenous comparison panel for model merging. FPCA reads that panel before candidate scoring, activates only for statistically unresolved Global comparisons with sufficient conflict coverage, and converts the exposed capability conflicts into a checkpoint choice. It selects differently from the Global point estimate in 16 of 27 construction-only near-ties and in 30 of 46 eligible comparisons overall. On separately authored neighborhoods, those choices reduce NRR by 2.3 points; conflict sensitivity and CGMR further turn the same signal into accurate block localization and sparse correction. In the near-tie regime, composition history recovers decision-relevant information that aggregate accuracy discards.

Reproducibility Statement

Sections 3–5 specify the split contract, parent-conditioned evaluation objects, selection rule, decoding, optimization, randomization, uncertainty, compute, and raw denominators. The appendix reports comparator budgets, sensitivity sweeps, independent neighborhood families, and open-ended validation.

AI USE STATEMENT

Available manuscript-preparation records document generative-AI assistance for copyediting, formatting and consistency checks, and bibliographic metadata verification. This preparation pass did not use generative AI to generate experimental measurements or alter the reported scientific evidence.

REFERENCES

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

Shir Ashury Tahan, Ariel Gera, Benjamin Sznajder, Leshem Choshen, Liat Ein-Dor, and Eyal Shnarch. Label-efficient model selection for text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8384–8402. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.456. URL https://aclanthology.org/2024.acl-long.456/.

Yuntao Bai et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. doi: 10.48550/arXiv.2212.08073. URL https://arxiv.org/abs/2212.08073.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Jakob Bjorner, George Stoica, Pratik Ramesh, Daniel Bolya, Taylor Hearn, and Judy Hoffman. ZipIt! merging models from different tasks without training. *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LEYUkvdUhq.

Oliver Bolton, Aakanksha, Arash Ahmadian, Sara Hooker, Marzieh Fadaee, and Beyza Ermis. SimMerge: Learning to select merge operators from similarity signals. In *ICLR 2026 Test-Time Updates (TTU) Workshop*, 2026. URL https://openreview.net/forum?id=7cQDm31mKF.

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. doi: 10.48550/arXiv.1803.05457. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. *Proceedings of the 36th International Conference on Machine Learning*, 97:1310–1320, 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. *Proceedings of the European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-73226-3_16. URL https://link.springer.com/chapter/10.1007/978-3-031-73226-3_16.

Yavuz Durmazkeser, Patrik Okanovic, Andreas Kirsch, Torsten Hoefler, and Nezihe Merve Gürel. Active model selection for large language models. *arXiv preprint arXiv:2510.09418*, 2025. doi: 10.48550/arXiv.2510.09418. URL https://arxiv.org/abs/2510.09418.

Wee Chung Gan and Hwee Tou Ng. Improving the robustness of question answering systems to question paraphrasing. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6065–6075, 2019. doi: 10.18653/v1/P19-1610. URL https://aclanthology.org/P19-1610/.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021b. URL https://openreview.net/forum?id=7Bywt2mQsCe.

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. *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

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

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36:47669–47681, 2023. doi: 10.52202/075280-2064. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/949f0f8f32267d297c2d4e3ee10a2e7e-Abstract-Datasets_and_Benchmarks.html.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00276/43518.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35, 2022. doi: 10.52202/068431-0142. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0cde695b83bd186c1fd456302888454c-Abstract-Conference.html.

Shuqi Liu, Han Wu, Bowei He, Xiongwei Han, Mingxuan Yuan, and Linqi Song. SensMerging: Sensitivity-guided parameter balancing for merging large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19243–19255. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.984. URL https://aclanthology.org/2025.findings-acl.984/.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1706.06083.

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

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021. doi: 10.18653/v1/2021.eacl-main.39. URL https://aclanthology.org/2021.eacl-main.39/.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 856–865, 2018. doi: 10.18653/v1/P18-1079. URL https://aclanthology.org/P18-1079/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Alexander Rubinstein, Benjamin Raible, Martin Gubri, and Seong Joon Oh. DISCO: Diversifying sample condensation for efficient model evaluation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SoOgBHa3dZ.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.95. URL https://aclanthology.org/2024.eacl-long.95/.

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

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems*, 36, 2023. doi: 10.52202/075280-3508.

URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html.

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

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. Representation surgery for multi-task model merging. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56332–56356. PMLR, 2024b. URL https://proceedings.mlr.press/v235/yang24t.html.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. *International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=nZP6NgD3QY.

Zongzhen Yang, Binhang Qi, Hailong Sun, Wenrui Long, Ruobing Zhao, and Xiang Gao. CABS: Conflict-aware and balanced sparsification for enhancing model merging. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70973–70999, 2025. URL https://proceedings.mlr.press/v267/yang25x.html.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. doi: 10.48550/arXiv.2307.15043. URL https://arxiv.org/abs/2307.15043.

Table 8: Component ablation on the Llama-3.1-8B multi-task merge at the 0.5% operating point. This ablation and the budget sweep report their respective rounded three-seed dispersions.

Variant	Global	NRR ↓	ΔNRR	Overhead
Full CGMR	65.3	12.1±0.7	—	1.6×
− Conflict-sensitive coefficients	65.1	14.3±0.9	+2.2	1.3×
− Sparse residual	65.1	14.8±0.9	+2.7	1.2×
− Conflict penalty ($\mu = 0$)	65.6	15.1±1.0	+3.0	1.5×
− Anchor mining (random prompts)	65.2	14.6±0.9	+2.5	1.6×
TIES baseline	65.1	16.4±1.0	+4.3	1.1×

Table 9: Out-of-regime controls from the predeclared screen. An FPCA selection change requires both Global near-tie membership and sufficient conflict coverage.

Control	Pairs	Abstain	Eligible	Selection changes	Median gap [IQR]
Sparse parent conflict	12	10	2	0	0.3 [0.1,0.6]
Global gap ≥ 1.0	12	—	—	0	0.8 [0.4,1.5]

A Answer-Key Scoring and Neighborhood Validation

The primary suites use deterministic dataset-level evaluators rather than ROUGE-L. MMLU, HellaSwag, ARC-Challenge, CommonsenseQA, and PIQA map the first valid option token to a canonical choice. GSM8K extracts the final numeric answer and normalizes signs, separators, and equivalent decimal forms. MATH extracts the final answer, applies algebraic normalization, and uses symbolic equivalence when both expressions parse. Natural Questions and TriviaQA apply case, article, punctuation, and whitespace normalization before matching the published alias set. The same parser is used for Global accuracy, parent conflict, and NRR, preventing metric changes between eligibility and arbitration.

Neighborhood construction preserves that evaluation contract. Paraphrases retain the original key or alias set. Entity edits are admitted only when a deterministic replacement produces an unambiguous updated key; edits that alter the required relation, fail the answer parser, or create multiple valid keys are rejected before candidate scoring. The independent adjudicator repeats key verification after manual editing and remains disjoint from candidate outputs and selector decisions. Open-ended IF and SF experiments use the normalized-reference thresholds defined from their mining splits and are reported separately in Section I.

B Ablations and Boundary Controls

The differentiable conflict term contributes the largest single degradation when removed, while coefficient and residual stages recover complementary portions of the 4.3-point TIES-to-CGMR improvement. Random anchors lose 2.5 points at identical overhead, directly supporting parent conflict as the targeting variable.

FPCA abstains on 10 of 12 sparse-conflict pairs and leaves the choice unchanged for the two eligible pairs; the median raw failure gap is 36. For 12 globally decisive comparisons, the Global winner remains selected in 11/12 cases; one case produces 1 diagnostic warning, the median raw gap is 96, and no FPCA change is issued because the pairs lie outside $\mathcal{P}_\epsilon$. These controls confirm that eligibility concentrates FPCA's decisions in the intended near-tie, conflict-rich regime.

C Matched-Budget Refinement Controls

The following comparisons hold the 500×8 evaluator, six-block target, 200+100 update schedule, 0.5% residual budget, data splits, and three seeds fixed. Only block targeting or initialization changes.

Table 10: Operating-point sensitivity. The first three rows report choices different from Global; the remaining rows report agreement with the default 30% anchor, eight-neighbor FPCA choice.

Analysis	Comparisons	Different/stable	Rate
$\epsilon = 0.1$	28	19	67.9%
$\epsilon = 0.2$	46	30	65.2%
$\epsilon = 0.3$	67	40	59.7%
MT-Llama budget grid	15	13	86.7%
MT-Qwen budget grid	15	12	80.0%
Expanded case library	690	559	81.0%

Table 11: Matched-budget targeted-refinement controls on MT-Llama. Raw failures aggregate three seeds.

Method	Global ↑	NRR ↓	Overhead	Raw / candidate ↓
Reference rows				
TIES-Merging	65.1±0.3	16.4±1.0	1.1×	1968
Fisher Merging	65.4±0.3	15.8±0.9	1.3×	1896
Post-merge FT	66.2±0.3	13.4±0.8	8.4×	1608
CGMR	65.3±0.3	12.1±0.7	1.6×	1452
Matched-budget targeted controls (TIES start unless noted)				
Random targeting	64.9±0.4	15.3±1.1	1.4×	1836
Magnitude-guided targeting	65.0±0.4	14.9±1.0	1.4×	1788
Validation-loss-guided targeting	65.1±0.4	14.4±0.9	1.5×	1728
Fisher-guided targeting	65.2±0.3	13.8±0.8	1.5×	1656
CGMR from Fisher start	65.4±0.2	12.6±0.6	1.8×	1512

CGMR improves on the strongest MT-Llama non-conflict control by 1.7 NRR points at comparable Global accuracy, and the Fisher-start row retains the advantage.

Table 12: Matched-budget targeted-refinement controls on MT-Qwen under the identical protocol.

Method	Global ↑	NRR ↓	Overhead	Raw / candidate ↓
Random targeting	65.6±0.4	14.7±1.0	1.4×	1764
Magnitude-guided targeting	65.7±0.3	14.2±0.9	1.4×	1704
Validation-loss-guided targeting	65.8±0.3	13.7±0.8	1.5×	1644
Fisher-guided targeting	65.9±0.3	13.1±0.7	1.5×	1572
CGMR	65.9±0.2	11.8±0.6	1.7×	1416
CGMR from Fisher start	66.0±0.2	12.2±0.6	1.8×	1464

The MT-Qwen result reproduces the targeting advantage: CGMR lowers NRR by 1.3 points and 156 failures relative to Fisher-guided targeting, while Fisher initialization reaches 12.2 NRR. Matched optimization budgets therefore preserve the conflict-guided advantage across both backbone families.

D Detailed Answer-Key and Slice Results

Table 13 provides the paired uncertainty and event totals underlying the concise main-paper comparison.

Only parent-conflict neighborhoods preserve a gap above 3 NRR points with the same winner in both cases. The per-method rows confirm that this difference is not produced by evaluation-set size, threshold, or seed count.

Table 13: Answer-key near-ties selected before NRR inspection. Paired cells list Fisher/CGMR.

Case	Global	Δ_G	NRR	Δ_{NRR}	Raw failures	Global / FPCA
MT-Llama	65.4±0.3 / 65.3±0.3	0.1	15.8±0.9 / 12.1±0.7	3.7	1896 / 1452	Fisher / CGMR
MT-Qwen	66.1±0.3 / 65.9±0.2	0.2	14.9±0.8 / 11.8±0.6	3.1	1788 / 1416	Fisher / CGMR

Table 14: Detailed answer-key slice comparison. Every slice uses 500 anchors, eight neighbors, common thresholds, and three seeds. Raw cells list Fisher/CGMR.

Case	Slice	Fisher NRR	CGMR NRR	ΔNRR	Raw failures	Selected
Case 1: MT-Llama	Parent conflict	15.8±0.9	12.1±0.7	3.7	1896 / 1452	CGMR
	High loss	14.2±0.9	13.8±0.8	0.4	1704 / 1656	CGMR
	Random	12.9±0.8	13.1±0.8	-0.2	1548 / 1572	Fisher
	Generic paraphrase	13.5±0.8	13.2±0.7	0.3	1620 / 1584	CGMR
	Validation loss	14.7±0.9	14.5±0.8	0.2	1764 / 1740	CGMR
Case 2: MT-Qwen	Parent conflict	14.9±0.8	11.8±0.6	3.1	1788 / 1416	CGMR
	High loss	13.7±0.8	13.4±0.7	0.3	1644 / 1608	CGMR
	Random	12.6±0.7	12.9±0.8	-0.3	1512 / 1548	Fisher
	Generic paraphrase	13.4±0.8	13.6±0.8	-0.2	1608 / 1632	Fisher
	Validation loss	14.1±0.8	13.9±0.8	0.2	1692 / 1668	CGMR

E Held-out Paraphrase-Generator Shift

Table 15 keeps anchors, thresholds, references, and candidate checkpoints fixed while changing the neighborhood generator from Llama-3.1-8B to Qwen-2.5-7B.

Table 15: Held-out paraphrase-generator shift on MT-Llama.

Method	NRR (in-dist)	NRR (OOD para.)	Degradation
TIES-Merging	16.4±1.0	17.8±1.1	+1.4
Fisher Merging	15.8±0.9	17.3±1.0	+1.5
CGMR	12.1±0.7	13.9±0.8	+1.8

CGMR retains a 3.4-point NRR advantage over Fisher under the generator swap. Table 17 extends this result to a second automatic family and manually edited variants for both answer-key cases.

F Localization Export Precision

Table 16 records the exported point estimates behind the rounded main-paper values; Table 6 supplies bootstrap uncertainty and stronger comparators.

Table 16: Unrounded localization correlations corresponding to Table 6.

Setting	Validation loss ρ	Conflict sensitivity ρ
MT-Llama	0.414	0.686
MT-Qwen	0.447	0.731
IF-Llama	0.381	0.664
IF-Qwen	0.419	0.703
SF-Llama	0.441	0.716
SF-Qwen	0.468	0.742

The unrounded margins vary from 0.272 to 0.284, confirming that the repeated two-decimal differences result from rounding rather than duplicated exports.

G Independent Neighborhood Families

For both answer-key cases, Table 17 freezes anchors, thresholds, candidate checkpoints, and the 500×8 budget, then scores Fisher/CGMR under a second automatic family and a manually edited family.

Table 17: Independent neighborhood-family checks. NRR and raw columns report Fisher and CGMR separately; shifts are relative to each model's in-distribution NRR and raw counts aggregate 12,000 shared prompt instances per candidate.

Case	Family	Fisher NRR	CGMR NRR	Shift F/C	Raw F/C	Winner
MT-Llama	In-distribution	15.8±0.9	12.1±0.7	— / —	1896 / 1452	CGMR
	Generator-2	17.4±1.0	14.2±0.8	+1.6 / +2.1	2088 / 1704	CGMR
	Human-edited	16.9±0.9	13.7±0.8	+1.1 / +1.6	2028 / 1644	CGMR
MT-Qwen	In-distribution	14.9±0.8	11.8±0.6	— / —	1788 / 1416	CGMR
	Generator-2	16.3±0.9	13.1±0.8	+1.4 / +1.3	1956 / 1572	CGMR
	Human-edited	15.7±0.8	12.6±0.7	+0.8 / +0.8	1884 / 1512	CGMR

The FPCA winner survives both independent families in both cases, with 3.1–3.2-point Fisher/CGMR separations.

H Inspecting Individual Conflict Decisions

The exact-key mechanism is inspectable at the decision level without substituting examples for aggregate evidence. A deterministic display rule applied to the frozen records underlying Tables 2 and 3 selects the anchor with the largest positive Fisher-minus-CGMR neighbor-failure margin, the eligible anchor with the smallest nonzero absolute margin, and the most negative-margin anchor; mining index resolves ties. The displays below use the two lowest-index neighbors, while NRR averages all eight.

Clear FPCA success. For the anchor "Which city is Turkey's capital?", the parents answer "Istanbul" and "Ankara", and the exact key is "Ankara". On "Name the capital of Türkiye", Fisher answers "Istanbul" (fail) and CGMR "Ankara" (pass); on "What is the seat of Turkey's national government?", the same outputs recur. Fisher therefore contributes two displayed failures to the NRR numerator and CGMR contributes none.

Borderline improvement. For "What gas do plants absorb during photosynthesis?", the parents answer "oxygen" and "carbon dioxide", with "carbon dioxide" as the key. Both candidates pass "Which atmospheric gas is taken in by plants for photosynthesis?"; on "During photosynthesis, plants remove which gas from the air?", Fisher answers "oxygen" (fail) and CGMR "carbon dioxide" (pass). The displayed failure contribution is thus one for Fisher and zero for CGMR.

Local counterexample. For "Which planet is known as the Red Planet?", the parents answer "Mars" and "Jupiter", and the key is "Mars". Fisher passes "Name the planet commonly called the Red Planet", whereas CGMR answers "Jupiter" (fail); both answer "Mars" to "The Red Planet nickname refers to which planet?" Here CGMR contributes one displayed failure and Fisher contributes none, making a local exception visible within the aggregate advantage.

I Open-Ended Corroboration

Evidence synthesis. The answer-key branch supports the selection claim through three distinct comparisons: the 160-pair screen measures recurrence, equal-budget slices isolate the parent-conflict signal, and the separately authored 24-case evaluator measures the chosen model on new neighborhoods. Additional generator and human-edited families test the two highlighted decisions without changing checkpoints or budgets. Instruction-following and safety then ask whether conflict-conditioned refinement remains useful with sequence-level references: the former gains 4.2 NRR points over Fisher with 0.81 external-label agreement, while the latter gains 2.7 held-out refusal-regression points with a 0.4-point benign-utility tradeoff.

Table 18: Cross-regime merging results. Global and NRR are percentages; SafeReg is refusal regression. Mean ± standard deviation over three seeds.

Setting	Method	Global ↑	NRR ↓	SafeReg ↓	Overhead
MT-Llama	Task Arithmetic	64.3±0.4	18.7±1.2	—	1.0×
	TIES-Merging	65.1±0.3	16.4±1.0	—	1.1×
	DARE	64.8±0.4	17.1±1.1	—	1.1×
	Fisher Merging	65.4±0.3	15.8±0.9	—	1.3×
	Post-merge FT	66.2±0.3	13.4±0.8	—	8.4×
	CGMR	65.3±0.3	12.1±0.7	—	1.6×
IF-Qwen	Task Arithmetic	61.7±0.5	21.3±1.4	—	1.0×
	TIES-Merging	62.4±0.4	18.9±1.2	—	1.1×
	DARE	62.1±0.4	19.6±1.3	—	1.1×
	Fisher Merging	62.8±0.4	17.4±1.0	—	1.3×
	Post-merge FT	63.9±0.3	14.8±0.9	—	9.1×
	CGMR	62.7±0.3	13.2±0.8	—	1.7×
SF-Llama	Task Arithmetic	63.1±0.4	22.8±1.5	14.6±1.1	1.0×
	TIES-Merging	63.8±0.4	19.7±1.2	12.3±0.9	1.1×
	DARE	63.4±0.4	20.9±1.3	13.1±1.0	1.1×
	Fisher Merging	64.2±0.3	18.1±1.1	11.4±0.8	1.3×
	Post-merge FT	65.1±0.3	15.3±0.9	9.7±0.7	8.7×
	CGMR	64.0±0.3	13.8±0.8	8.4±0.6	1.6×

The IF and SF settings reproduce the near-tied Global/large local-gap pattern: CGMR lowers NRR by 4.2 and 4.3 points relative to Fisher, while SF refusal regression improves by 3.0 points. At 4,000 evaluations per seed, the NRR gaps correspond to 168 and 172 fewer threshold crossings.

Table 19: Normalized-reference proxy validation against a proxy-stratified external-label sample.

Regime	Sample n	Agreement	Precision	Recall	Style-only false positives
IF-Qwen	160	0.81	0.78	0.84	11
SF-Llama	180	0.86	0.88	0.82	7

Agreement reaches 0.81 on IF-Qwen and 0.86 on SF-Llama; the 11 and 7 style-only false positives identify the principal measurement error. These values quantify conditional scoring validity in the proxy-stratified sample.

Table 20: Held-out refusal-family evaluation. Regression and over-refusal are percentages; raw counts use 4,000 held-out variants.

Method	Refusal regression ↓	Raw ↓	Benign over-refusal ↓	Global ↑
Fisher	12.0±0.8	480	6.2±0.4	64.2
Fisher-guided control	10.7±0.7	428	6.4±0.4	64.1
CGMR	9.3±0.6	372	6.6±0.5	64.0

On a refusal family excluded from anchor generation and optimization, CGMR lowers regression by 2.7 points versus Fisher and 1.4 points versus the predeclared Fisher-guided control. The corresponding benign over-refusal cost is 0.4 points relative to Fisher.