MERGE BEFORE YOU DISTILL: A FISHER-GEOMETRIC VIEW OF MULTI-TEACHER ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Consolidating specialized foundation models into a unified generalist is a central goal of post-training. Two approaches are in common use: model merging, which is fast but treated as a heuristic, and multi-teacher on-policy distillation, which matches distributions but demands heavy rollout compute and suffers from cross-task gradient conflict. We show that the two are one objective seen at different orders. Within a shared basin of homologous specialists, the multi-teacher on-policy distillation objective is, to second order, exactly the Fisher-merging objective. Its minimizer is therefore the Fisher merge, a classic model-merging method given by a closed-form combination of the specialists, so *merging is optimal, to second order, as a closed-form stage of distillation*. Building on this identification, we propose *Residual-Governed Branch-and-Merge* (RGBM), a two-phase consolidation procedure. Concretely, a full-parameter Fisher merge is intractable on foundation models, so Phase 1 restricts the merge to the task-vector subspace. Extending tangent-space analyses of merging, we show that the same coefficients that mix task vectors also mix implicit rewards, which lets a residual program on unlabelled prompts solve for them. The program minimizes the worst per-domain residual and reports its energy-normalized values as the *domain residual spectrum*. Phase 2 lets this spectrum govern what may be added: RGBM passivates domains with low normalized residuals, trains isolated branches only for domains above the spectral gap, and writes back the admitted updates with calibrated shrinkage. On two open-source consolidation suites, the Phase 1 merge alone matches or outperforms fully trained multi-teacher distillation, and RGBM improves on it further with zero extra joint-training steps, whereas every un-gated write-back, including continued joint distillation from the merge, falls below it.

1 INTRODUCTION

Post-training of large language models (LLMs) increasingly produces homologous domain specialists—fine-tuned from a common base via reinforcement learning (RL) on distinct reasoning modalities (Agarwal et al., 2024; Shao et al., 2024; DeepSeek-AI, 2025; Team, 2025a; Gao et al., 2026; Liu et al., 2025a). Serving an ensemble of specialists multiplies memory and latency, making consolidation into a single generalist essential.

Two paradigms are in common use, each with its own shortcomings. *Weight-space merging* (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024; Lu et al., 2024; Yang et al., 2026a) adds to the base model a weighted sum of the specialists' weight updates (their *task vectors*) at zero training cost, but the weights are set by hand and, once the merge is formed, nothing tells which domain it failed to recover or whether the chosen weights still combine the specialists faithfully. *Multi-teacher on-policy distillation* (Agarwal et al., 2024; Gao et al., 2026; Sun et al., 2026; Yang et al., 2026b; Chen et al., 2026) (MOPD) instead trains a student to match each specialist (its *teacher*) on the student's own generations, which repairs what merging misses, but at the cost of rollouts on every domain and with no control over which domains its shared-weight gradients overwrite. Neither paradigm can answer the question the other one raises: what a merge has already recovered, and where distillation is still needed without harming what is already correct. Answering it requires relating the two, which is what this paper does.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

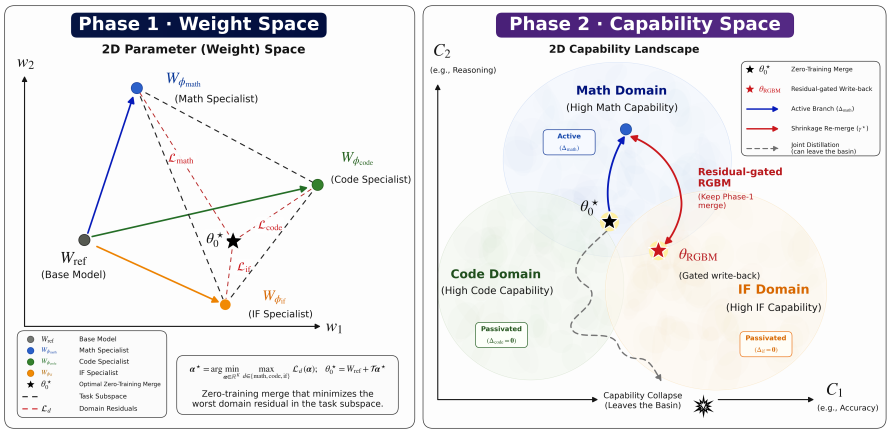


Figure 1: **Merge before you distill (schematic).** (a) Weight space: the dashed triangle is the task subspace spanned by the Math, Code, and IF task vectors. A Minimax projection on unlabelled prompts yields the zero-training merge $\theta_0^\star = W_{\text{ref}} + \mathcal{T}\boldsymbol{\alpha}^\star$, where $\mathcal{T}$ stacks those task vectors, together with the per-domain residuals $\mathcal{L}_d$ (red segments). Normalizing these residuals by domain reward energy gives the domain residual spectrum. (b) Capability space: joint distillation from $\theta_0^\star$ can leave the basin (dashed path). The normalized spectrum passivates Code and IF; RGBM writes back only the admitted Math branch, shrunk by $\gamma^\star$ (the largest factor that still passes the linearity test). See Section 3 for details.

Our starting point is that these paradigms optimize the same objective at different orders of approximation (Section 2). Expanded to second order around the specialists in a shared basin, the distillation objective is exactly the Fisher-merging objective of Matena & Raffel (2022), whose minimizer is a weighted average of the specialists with Fisher-information weights; merging is therefore not an alternative to distillation but is optimal, to second order, as its closed-form stage. A full-parameter Fisher merge is intractable on foundation models, so we restrict the merge to the task-vector subspace (Phase 1, described next). Extending tangent-space analyses of merging, we show that, to first order, the same coefficients that mix the task vectors also mix the specialists' *implicit rewards* (the log-probability ratios between each specialist and the base), which lets us obtain them by minimizing the worst reward-space residual on unlabelled prompts. After solving, reward-energy normalization measures the relative share left unrecovered in each domain. The fidelity of this first-order fit provides an intrinsic, label-free test of whether a checkpoint still lies where the expansion holds.

Grounded in this framework, we propose *Residual-Governed Branch-and-Merge* (RGBM) as shown in Figure 1. In Phase 1, RGBM determines the combination weights by minimizing the worst per-domain residual via a small convex program on unlabelled prompts, yielding the zero-training merged model. It then normalizes the residuals by their domain reward energies to form the *domain residual spectrum*. This spectrum directly dictates what, if anything, should follow the merge. Standard joint distillation tends to degrade the merged model because gradients from domains with low normalized residuals corrupt shared weights. To prevent this interference, Phase 2 leaves those domains untouched and trains isolated branches only for domains above a closed-form spectral gap. Each admitted branch update is then written back, scaled by the largest factor that passes the linearity test.

Empirically, we validate our framework across two multi-teacher consolidation benchmarks: Open-MOPD (Gao et al., 2026) (SmolLM3-3B with Math, Code, and Instruction Following specialists) and G-OPD (Yang et al., 2026b) (Qwen3-4B with Math and Code specialists). On Open-MOPD, the zero-training Phase 1 merge alone achieves a Macro average of 32.52, outperforming the official 600-step distilled student (30.93) without requiring any rollout compute. By passivating domains with low normalized residuals and training an isolated branch solely for Math, Phase 2 further elevates the Macro score to 33.03 with zero joint-training steps. Crucially, un-gated baselines such as TIES (Yadav et al., 2023), DARE (Yu et al., 2024), and continued joint distillation all fall below the merge, confirming that their updates can corrupt shared parameters. The same procedure transfers seamlessly to G-OPD, where RGBM surpasses the strongest trained baseline, demonstrating the generality of residual-governed consolidation.

Contributions. (1) **Theoretical Bridge:** we establish a connection between multi-teacher on-policy distillation and model merging, showing that the MOPD objective coincides to second order with Fisher merging (Matena & Raffel, 2022), and derive an implicit-reward homomorphism that yields a label-free linear basin diagnostic. (2) **Two-Phase RGBM Framework:** we propose *Residual-Governed Branch-and-Merge* (RGBM). Phase 1 computes the zero-training merge from a residual program on unlabelled prompts and forms an energy-normalized spectrum for cross-domain comparison. Phase 2 passivates domains below the spectral gap, trains isolated branches exclusively for domains above it, and writes admitted updates back via calibrated shrinkage. (3) **Empirical Validation and Protection Mechanism:** across two model families (SmolLM3-3B and Qwen3-4B), Phase 1 alone matches or beats fully trained distillation, and Phase 2 advances performance further with zero joint-training steps. Crucially, un-gated write-backs fall below the merge, confirming that RGBM succeeds by shielding domains below the normalized spectral gap from cross-task interference.

2 Merging Is Optimal, to Second Order, as a Closed-Form Stage of Distillation

This section establishes the identification on which the paper rests: to second order, distillation and Fisher merging optimize the same objective (Section 2.2); on the task-vector subspace that objective has a closed-form minimizer (Section 2.3); and, to first order in the tangent space, combining parameters adds implicit rewards, which yields a validity test (Section 2.4). Proofs are in Appendix A.

2.1 Setting

Let $W_{\text{ref}} \in \mathbb{R}^P$ denote the base weights (P parameters) and $\pi_\theta(\cdot \mid x)$ the autoregressive policy of a model with weights $\theta \in \mathbb{R}^P$ on prompt x, with y a completion. We consider K homologous specialists $\{W_{\phi_1}, \dots, W_{\phi_K}\}$, fine-tuned from the same W_{ref} on datasets $\mathcal{D}_d$ via RL on each domain, with policies $\pi_{W_{\phi_d}}$. The task vector of specialist d is $\tau_d := W_{\phi_d} - W_{\text{ref}}$, forming $\mathcal{T} := [\tau_1, \dots, \tau_K] \in \mathbb{R}^{P \times K}$. Multi-teacher on-policy distillation (MOPD) trains a student θ by minimizing the domain-weighted reverse-KL to the specialist policies, with mixture weights p_d:

$$J_{\text{MOPD}}(\theta) := \sum_{d=1}^{K} p_d \, \mathbb{E}_{x \sim \mathcal{D}_d} \Big[\mathrm{KL}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{W_{\phi_d}}(\cdot \mid x)\big) \Big], \quad p_d \geq 0, \quad \sum_{d=1}^{K} p_d = 1. \tag{1}$$

2.2 MOPD Is Fisher Merging to Second Order

Theorem 1 (Second-Order Equivalence of MOPD to Fisher Merging) *Assume the specialists share a basin $\mathcal{B} \subset \mathbb{R}^P$ (a connected neighborhood of W_{ref} containing every W_{ϕ_d}) on which each mapping $\theta \mapsto \mathrm{KL}(\pi_\theta \| \pi_{W_{\phi_d}})$ is three times differentiable with bounded third derivative. For $\theta \in \mathcal{B}$, the second-order Taylor expansion of $J_{\text{MOPD}}(\theta)$ around the specialist checkpoints $\{W_{\phi_d}\}_{d=1}^K$ is:*

$$J_{\text{MOPD}}(\theta) = \frac{1}{2} \sum_{d=1}^{K} p_d \, (\theta - W_{\phi_d})^\top F_d (\theta - W_{\phi_d}) + \mathcal{O}\left(\max_d \|\theta - W_{\phi_d}\|^3 \right), \tag{2}$$

where $F_d := \mathbb{E}_{x \sim \mathcal{D}_d, y \sim \pi_{W_{\phi_d}}} [\nabla_\theta \log \pi_{W_{\phi_d}}(y \mid x) \, \nabla_\theta \log \pi_{W_{\phi_d}}(y \mid x)^\top] \in \mathbb{R}^{P \times P}$ is the empirical Fisher matrix of specialist d. The unconstrained global minimizer of Eq. 2 is the Fisher merge:

$$\theta^\star_{\text{Fisher}} = \left(\sum_{d=1}^{K} p_d F_d \right)^{-1} \sum_{d=1}^{K} p_d F_d W_{\phi_d}. \tag{3}$$

The proof (Appendix A.1) relies only on the fact that at $\theta = W_{\phi_d}$, the KL divergence and its gradient vanish while its Hessian equals F_d. Equation 3 recovers the Fisher-weighted average of Matena & Raffel (2022), originally derived via a Laplace approximation to the posterior. Theorem 1 establishes that this exact point is also the second-order optimum of the multi-teacher reverse-KL objective. The two paradigms therefore differ in approximation order rather than optimization goal: while J_{MOPD} aligns student rollouts to teachers in distribution space, it corresponds in parameter space to the aggregate Fisher distance $\frac{1}{2} \sum_d p_d \|\theta - W_{\phi_d}\|^2_{F_d}$ with $\|v\|^2_{F_d} := v^\top F_d v$. Merging is thus optimal, to second order, as a closed-form stage of distillation: gradient steps taken from the Fisher merge act only on the third-order remainder, so merging should come first. What a linear combination in the task-vector subspace leaves unrecovered is a per-domain error, which the residual program of Section 3.1 measures.

2.3 Subspace Fisher Projection and Task Arithmetic

Eq. 3 requires inverting a $P \times P$ Fisher matrix, which is intractable for modern LLMs. We therefore restrict the student to the K-dimensional task subspace $\theta(\boldsymbol{\alpha}) = W_{\text{ref}} + \mathcal{T}\boldsymbol{\alpha}$, with one coefficient α_d per specialist. Let $\mathbf{e}_d$ be the d-th standard basis vector, so $\theta(\mathbf{e}_d) = W_{\phi_d}$ and $\theta(\boldsymbol{\alpha}) - W_{\phi_d} = \mathcal{T}(\boldsymbol{\alpha} - \mathbf{e}_d)$. That substitution compresses Eq. 2 into a $K \times K$ system:

$$J_{\text{sub}}(\boldsymbol{\alpha}) := \frac{1}{2}\sum_{d=1}^{K} p_d \,(\boldsymbol{\alpha} - \mathbf{e}_d)^\top S_d(\boldsymbol{\alpha} - \mathbf{e}_d), \quad S_d := \mathcal{T}^\top F_d \mathcal{T} \in \mathbb{R}^{K\times K}, \tag{4}$$

where S_d is the projected Fisher of specialist d. Setting $\nabla_{\boldsymbol{\alpha}} J_{\text{sub}} = \mathbf{0}$ gives the closed-form subspace solution:

$$\boldsymbol{\alpha}^\star_{\text{sub}} = \left(\sum_{d=1}^{K} p_d S_d\right)^{-1} \sum_{d=1}^{K} p_d S_d \mathbf{e}_d. \tag{5}$$

Remark 1 *Task Arithmetic (Ilharco et al., 2023) forms $W_{\text{ref}} + \mathcal{T}\boldsymbol{\alpha}$ with hand-set coefficients, typically the isotropic $\boldsymbol{\alpha} = \alpha\mathbf{1}$ with $\alpha = 1$. Eq. 5 identifies the coefficients that make this family the exact metric projection of the quadratic Fisher objective onto the task subspace ($\mathcal{T}^\top \nabla_\theta J_{\text{MOPD}}(\theta(\boldsymbol{\alpha}^\star_{\text{sub}})) = \mathbf{0}$); with these coefficients it is the computable closed-form instance of Fisher merging on LLMs.*

2.4 First-Order Homomorphism and a Validity Diagnostic

The subspace solution still depends on F_d. Forming $S_d = \mathcal{T}^\top F_d \mathcal{T}$ from parameters is impractical on LLMs, so the computations of Section 3 run in the space of implicit rewards. The token-level implicit reward of specialist d relative to W_{ref} is $r_d(x, y) := \log(\pi_{W_{\phi_d}}(y \mid x)/\pi_{W_{\text{ref}}}(y \mid x))$, the log-probability ratio that RL fine-tuning implicitly optimizes (Rafailov et al., 2023; Bai et al., 2026); write $\mathbf{r}(x, y) := [r_1(x, y), \ldots, r_K(x, y)]^\top \in \mathbb{R}^K$ for the stacked reward. Extending tangent-space analyses of merging (Ortiz-Jiménez et al., 2023), a first-order expansion around W_{ref} yields the following homomorphism.

Proposition 2 (First-Order Homomorphism) *Let $f(x; \theta)$ be the logit vector at prompt-prefix x and $Z(x; \theta)$ its softmax partition function. A first-order expansion around W_{ref} gives $r_d(x, y) \approx \nabla_\theta f(x; W_{\text{ref}})[y]^\top \tau_d - \Delta \log Z_d(x)$, where $\Delta \log Z_d(x) := \log Z(x; W_{\phi_d}) - \log Z(x; W_{\text{ref}})$. For any linear combination $\theta(\boldsymbol{\alpha}) = W_{\text{ref}} + \mathcal{T}\boldsymbol{\alpha}$:*

$$\log \frac{\pi_{\theta(\boldsymbol{\alpha})}(y \mid x)}{\pi_{W_{\text{ref}}}(y \mid x)} \approx \sum_{d=1}^{K} \alpha_d r_d(x, y) = \boldsymbol{\alpha}^\top \mathbf{r}(x, y). \tag{6}$$

Proposition 2 provides two operational capabilities. First, the projected score $\mathcal{T}^\top \nabla_\theta \log \pi_{W_{\phi_d}}$ equals $\mathbf{r}$ up to a mean shift (Appendix A.3), so the projected Fisher is an empirical reward covariance, $S_d \approx \text{Cov}_{\pi_{W_{\phi_d}}}(\mathbf{r})$, and the per-domain residual in Section 3.1 can be computed from forward passes alone. Second, the log-ratio of the merged model, $\ell_{\boldsymbol{\alpha}}(x, y) := \log \frac{\pi_{\theta(\boldsymbol{\alpha})}(y|x)}{\pi_{W_{\text{ref}}}(y|x)}$, should scale linearly with the superposed reward $\boldsymbol{\alpha}^\top \mathbf{r}(x, y)$. The unexplained variance when regressing $\ell_{\boldsymbol{\alpha}}$ on $\boldsymbol{\alpha}^\top \mathbf{r}$ over sampled tokens is then a label-free test of whether the first-order superposition of Proposition 2 still holds at that checkpoint:

$$\text{Err}_{\text{lin}}(\theta(\boldsymbol{\alpha})) := 1 - R^2\big(\ell_{\boldsymbol{\alpha}} \sim \boldsymbol{\alpha}^\top \mathbf{r}\big) = \min_{a,b} \frac{\mathbb{E}\big[(\ell_{\boldsymbol{\alpha}} - a\,\boldsymbol{\alpha}^\top \mathbf{r} - b)^2\big]}{\text{Var}[\ell_{\boldsymbol{\alpha}}]}. \tag{7}$$

We treat $\text{Err}_{\text{lin}}(\theta(\boldsymbol{\alpha})) \le \delta_{\text{tol}}$ under a fixed threshold δ_{tol} as the operational criterion for a valid linear basin (we use *linear basin* for this regime throughout). Large Err_{lin} means the first-order fit has broken; Section 3.3 uses the same test to cap how far a Phase 2 write-back may move.

3 RGBM: Residual-Governed Branch-and-Merge

Section 2 establishes that parameter merging should precede distillation, and that whatever the linear combination leaves unrecovered is what a later stage must address. **Residual-Governed Branch-and-Merge** (RGBM; Figure 2; full procedure in Algorithm 1, Appendix C.1) operationalizes this

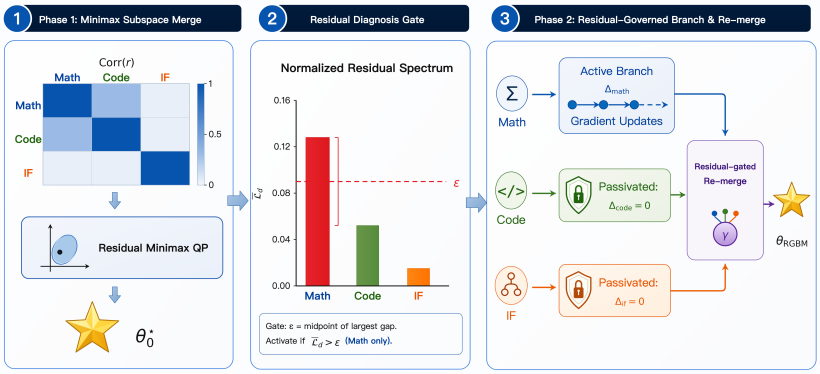


Figure 2: **Two-phase RGBM architecture.** (1) Phase 1 scores implicit rewards on unlabelled prompts, forms per-domain second-moment matrices M_d, and minimizes the worst residual $\mathcal{L}_d$ to obtain $\theta_0^\star$. The heatmap is the reward correlation, not M_d itself. (2) The bars are the energy-normalized Open-MOPD spectrum $\overline{\boldsymbol{\mathcal{L}}}$; the gate cuts at the midpoint of its largest adjacent gap (threshold ϵ; Eq. 10), so only Math is active. (3) Phase 2 writes active updates back with shrinkage γ to obtain θ_{RGBM}; if every domain is passivated, $\theta_{\mathrm{RGBM}} = \theta_0^\star$. See Sections 3.1 and 3.3 for details.

insight into two distinct phases. Phase 1 (Section 3.1) solves a convex subspace merge and then normalizes its per-domain remainders into the *domain residual spectrum*. After analyzing why continuing joint distillation degrades domains with low normalized residuals (Section 3.2), Phase 2 (Section 3.3) trains isolated branches exclusively for domains marked by the spectrum as unrecovered, writing them back via calibrated shrinkage.

3.1 Phase 1: Minimax Subspace Merge and the Domain Residual Spectrum

Per-domain residual from unlabelled prompts. Under the tangent-space homomorphism of Proposition 2 the projected score equals $\mathbf{r}$ up to a mean shift, so $S_d \approx \mathrm{Cov}_{\pi_{W_{\phi_d}}}(\mathbf{r})$ can be estimated without gradients. Let $\mathcal{D}_{\mathrm{calib},d}$ be unlabelled prompts from domain d. Rather than center the rewards, we use the empirical second-moment $M_d := \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{calib},d},\, y\sim\pi_{W_{\phi_d}}}[\mathbf{r}\mathbf{r}^\top] \in \mathbb{R}^{K\times K}$, assembled from forward passes of the base and the K specialists on completions sampled from specialist d. M_d equals that covariance plus a mean-shift outer product, and therefore upper-bounds the projected Fisher residual (Appendix A.3). We define the per-domain residual and its energy-normalized counterpart as

$$\mathcal{L}_d(\boldsymbol{\alpha}) := (\boldsymbol{\alpha} - \mathbf{e}_d)^\top M_d (\boldsymbol{\alpha} - \mathbf{e}_d),$$
$$\overline{\mathcal{L}}_d(\boldsymbol{\alpha}) := \frac{\mathcal{L}_d(\boldsymbol{\alpha})}{\mathbb{E}[r_d^2]}. \tag{8}$$

Phase 1 optimizes $\mathcal{L}_d$. After the merge is fixed, division by $\mathbb{E}[r_d^2]$, the second moment of the specialist's own reward (its *reward energy*), expresses the remainder as a fraction of domain-d energy. Thus $\overline{\mathcal{L}}_d$ is comparable across domains and is used for diagnosis and gating, but not for solving the merge. Because M_d contains the cross-domain products $\mathbb{E}[r_j r_k]$, $\overline{\mathcal{L}}_d(\boldsymbol{\alpha}^\star)$ is the fraction of domain-d reward energy still unexplained after the coefficients have traded the domains off against one another. A value near zero means this joint combination already accounts for domain d. It does not mean an isolated branch on d has nothing left to fit. A large value means a substantial share of that energy remains once the other domains have been accommodated.

Minimax merge. Eq. 5 requires domain weights p_d, which are rarely known and trade one domain against another, so the Fisher stage of Section 2.3 is not directly computable. We remove this degree of freedom by taking the worst case over all admissible weights, the probability simplex Δ_K: since $\max_{\mathbf{p}\in\Delta_K} \sum_d p_d \mathcal{L}_d(\boldsymbol{\alpha}) = \max_d \mathcal{L}_d(\boldsymbol{\alpha})$, the problem becomes the convex Minimax quadratic program (QP)

$$\boldsymbol{\alpha}^\star = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^K} \max_{d\in\{1,\dots,K\}} \mathcal{L}_d(\boldsymbol{\alpha}), \tag{9}$$

i.e., the worst subspace residual under adversarially chosen domain weights. Its solution gives the zero-training checkpoint $\theta_0^\star := W_{\mathrm{ref}} + \mathcal{T}\boldsymbol{\alpha}^\star$. We then report the *Domain Residual Spectrum*

$\overline{\boldsymbol{\mathcal{L}}} := [\overline{\mathcal{L}}_1(\boldsymbol{\alpha}^\star), \ldots, \overline{\mathcal{L}}_K(\boldsymbol{\alpha}^\star)]$ for cross-domain comparison and Phase 2 allocation. The QP is K-dimensional, so its cost is dominated by assembling the M_d (Appendix C.6). For numerical stability, the QP replaces M_d by $\tilde{M}_d = M_d + \lambda I$; the reported spectrum uses the unregularized M_d. We leave $\boldsymbol{\alpha}$ unconstrained, so a coefficient may lie outside $[0, 1]$ if that reduces the worst residual.

3.2 Why Not Simply Keep Distilling from $\theta_0^\star$?

The quadratic model of Theorem 1 is stationary at the Fisher merge, and approximately stationary on the subspace at $\theta(\boldsymbol{\alpha}^\star_{\text{sub}})$. Phase 1's $\theta_0^\star$ is not that point: it minimizes the worst residual, so $\nabla J_{\text{MOPD}}(\theta_0^\star)$ need not vanish, and continued joint distillation is not a refinement of a solved linear stage. Two mechanisms outside the quadratic regime then make it harmful. Each domain divergence $f_d(\theta) := \mathbb{E}_{x\sim\mathcal{D}_d}[\mathrm{KL}(\pi_\theta \,\|\, \pi_{W_{\phi_d}})]$ has $\nabla_\theta f_d(\theta_0^\star) \neq \mathbf{0}$, so mini-batch steps on shared parameters pull in conflicting directions, including components off the task subspace. A near-zero normalized residual $\overline{\mathcal{L}}_d$ adds a second problem, and it is a joint one. An isolated branch can still reduce domain d's own loss, because $\overline{\mathcal{L}}_d$ is the gap left after the cross-domain tradeoff, not a claim that domain d is saturated. Writing that branch into shared weights moves every domain: the step is spent where the joint merge has already closed the gap, so the three-domain loss can worsen even while domain d itself improves. Joint distillation from $\theta_0^\star$ should therefore degrade those low-residual domains on the shared model. A token-level KL penalty to $\pi_{\theta_0^\star}$ cannot prevent this, because the collision is in the weights. Section 4.3 checks both predictions.

3.3 Phase 2: Residual Gate, Isolated Branches, and Shrinkage Re-merge

RGBM therefore never applies a joint gradient to the merge; it partitions domains by the normalized spectrum $\overline{\boldsymbol{\mathcal{L}}}$ and treats the two groups differently.

Gate. Sort $\overline{\mathcal{L}}_{(1)} \geq \cdots \geq \overline{\mathcal{L}}_{(K)}$, append a floor $\overline{\mathcal{L}}_{(K+1)} := 0$, and cut at the midpoint of the largest adjacent gap,

$$k^\star = \arg\max_{i=1,\ldots,K} \big(\overline{\mathcal{L}}_{(i)} - \overline{\mathcal{L}}_{(i+1)}\big), \qquad \epsilon := \tfrac{1}{2}\big(\overline{\mathcal{L}}_{(k^\star)} + \overline{\mathcal{L}}_{(k^\star+1)}\big). \tag{10}$$

Domain d is *active* iff $\overline{\mathcal{L}}_d > \epsilon$, otherwise *passivated*; the floor makes every domain active when normalized residuals cluster, and any ϵ inside the same gap yields the same allocation (Appendix C.4).

Passivated and active domains. Passivated domains ($\overline{\mathcal{L}}_d \leq \epsilon$) receive no shared write-back ($\Delta_d = \mathbf{0}$). The joint merge already accounts for them, which does not preclude an isolated gain on that domain alone. Each active domain trains an isolated branch $\theta_d \leftarrow \theta_0^\star$ on its native data $\mathcal{D}_d$ with no gradient sharing, giving an update $\Delta_d = \theta_d^{(\text{end})} - \theta_0^\star$; isolation guarantees that no gradient computed on domain d touches a passivated domain during training. Whether the update can be written back without damaging them is decided next.

Shrinkage re-merge. The sole Phase 2 write-back is a shrunken sum of the active updates,

$$\theta_{\text{RGBM}} = \theta_0^\star + \sum_{d:\overline{\mathcal{L}}_d > \epsilon} \gamma_d^\star \, \Delta_d, \qquad \gamma_d^\star = \max\big\{\gamma \in \Gamma \mid \mathrm{Err}_{\text{lin}}(\theta_0^\star + \gamma\Delta_d) \leq \delta_{\text{tol}}\big\}, \tag{11}$$

where $\Gamma \subset (0, 1]$ is a fixed discrete grid and $\mathrm{Err}_{\text{lin}}$ is evaluated on the same unlabelled calibration prompts $\mathcal{D}_{\text{calib}} = \bigcup_d \mathcal{D}_{\text{calib},d}$ used for M_d. $\gamma_d^\star$ is the largest grid value at which that single write-back still passes the linearity test of Section 2.4, so the write-back keeps the merged model's log-ratio within δ_{tol} of the Phase 1 superposition $\boldsymbol{\alpha}^{\star\top}\mathbf{r}$, which is held fixed as the predictor; Proposition 4 (Appendix A.4) bounds the first-order remainder at $\theta_0^\star + \gamma\Delta_d$ by a quantity that is non-decreasing in γ. When several branches are active, evaluation uses their shrunken sum. If the gate leaves no active domain, $\theta_{\text{RGBM}} = \theta_0^\star$. The whole procedure has three fixed constants (λ, Γ, δ_{tol}) and no labelled data; their values and the resulting allocations on both suites are given in Section 4.1.

4 Experiments

4.1 Setup

Main suite (Open-MOPD). Our primary testbed is the Open-MOPD specialist suite (Gao et al., 2026): SmolLM3-3B (Bakouch et al., 2025) with $K = 3$ homologous RL specialists (the usual MOPD domain count; Gao et al., 2026; Sun et al., 2026; Chen et al., 2026), fine-tuned from

the mixed-SFT checkpoint MixSFT as W_{ref}. The specialists are Math, Code, and Instruction Following (IF), scored on AIME 2024/2025, LiveCodeBench v5/v6 (Jain et al., 2025), and IFEval/IFBench (Zhou et al., 2023; Pyatkin et al., 2025). We report Macro $= \frac{1}{3}$(Math + Code + IF), where IF is the mean of IFEval and IFBench, and re-evaluate every system under one protocol. The italic RouteRL row is the routing oracle of Gao et al. (2026) that serves all K specialists; Multi-Teacher OPD is their released 600-step student. Every Phase 2 system starts from the same Phase 1 merge $\theta_0^\star$. TIES (Yadav et al., 2023) and DARE (Yu et al., 2024) sparsify and sum the isolated on-policy distillation (OPD) branch updates Δ_d—not the specialist task vectors—and RGBM uses those same branches, differing only in what is written back (Appendix Table 6). Per-benchmark numbers and hyperparameters are in Appendices B.2 and C.5.

Secondary suite (G-OPD). To test transfer across model families we apply the same procedure, unchanged, to the G-OPD suite (Yang et al., 2026b): Qwen3-4B (Team, 2025b) with Math and Code specialists (intermediate checkpoints of Table 2 in Yang et al., 2026b; 500/300 steps) on their 7 official benchmarks. Supervised fine-tuning (SFT), OPD, and the extended recipe ExOPD are from Yang et al. (2026b); merges are evaluated by us under the official protocol. The italic Domain-Matched Teacher row is that specialist pair, an oracle rather than a single generalist. Per-benchmark results and the fully trained pair are in Appendix B.6.

RGBM constants and resulting allocations. We use $\lambda = 10^{-4}$, $\Gamma = \{0.125, 0.25, 0.5, 0.75\}$, and $\delta_{\text{tol}} = 0.10$ on both suites. On Open-MOPD the energy-normalized spectrum is $\overline{\boldsymbol{\mathcal{L}}} = (0.128, 0.052, 0.015)$ for (Math, Code, IF); the largest gap is between Math and Code, so $\epsilon = 0.090$, only Math is active, and Eq. 11 selects $\gamma^\star_{\text{math}} = 0.125$, the largest $\gamma \in \Gamma$ with $\text{Err}_{\text{lin}} \le \delta_{\text{tol}}$ ($\text{Err}_{\text{lin}} = 0.0701$; Appendix Table 8). On G-OPD, $\boldsymbol{\alpha}^\star = (0.675, 0.666)$ and $\overline{\boldsymbol{\mathcal{L}}} = (0.094, 0.090)$ cluster, so the largest gap is down to the floor, $\epsilon = 0.045$, both domains are active, and the same rule gives $\gamma^\star = 0.5$. Hereafter, low- and high-residual domains refer to this normalized spectrum. Assembling the M_d costs < 15 minutes on 8 GPUs (< 2 GPU·h) of unlabelled scoring and the QP solves in < 10 ms.

4.2 Merge Before You Distill: Zero-Training Merge vs. Trained Distillation

The closed-form merge matches or beats trained distillation. On Open-MOPD (Table 1) the Minimax merge $\theta_0^\star$ reaches 32.52 Macro at zero gradient cost, above Multi-Teacher OPD, the released 600-step student (30.93; 144 GPU·h), and above isotropic Task Arithmetic (31.31). On G-OPD (Table 2) the same procedure gives 53.34, above SFT, OPD, and Task Arithmetic and within 0.45 of the trained ExOPD recipe (equal-weight averaging collapses to 36.75: the gain is in the coefficients). The 31.90 row is the subspace Fisher solution of Eq. 5 under uniform weights $p_d = 1/K$, i.e., the minimizer of $\sum_d p_d \mathcal{L}_d(\boldsymbol{\alpha})$ with the Phase 1 second moments M_d in place of S_d. Eq. 5 requires the domain weights p_d, and those weights are not known; the uniform choice is a default, not an estimate. Minimax, which minimizes the worst residual instead, is 0.62 higher and lifts IF from 47.99 to 49.60—when no domain may be sacrificed, the adversarial weighting is the operational choice.

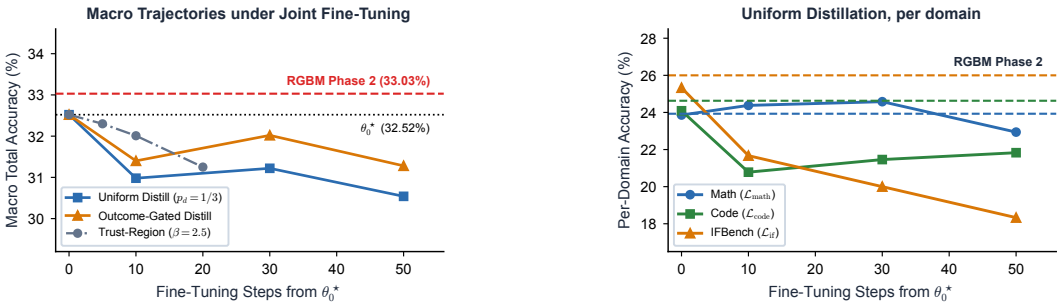
Table 1: **Open-MOPD** (SmolLM3-3B; accuracy). Macro: mean of Math, Code, and IF. Bold: best deployable model. Italic: routing oracle. Gated Math: admitted branch only, at the listed γ.

Method	Math	Code	IF	Macro
RouteRL (Oracle Ensemble)	*27.24*	*26.27*	*53.52*	*35.68*
Multi-Teacher OPD	24.53	21.18	47.12	30.93
Task Arithmetic	24.35	22.81	46.77	31.31
Subspace Fisher Merge	24.11	23.59	47.99	31.90
RGBM Phase 1 (Ours, $\theta_0^\star$)	23.86	24.09	49.60	32.52
Uniform Distillation (from $\theta_0^\star$)	23.97	21.36	47.42	30.91
TIES (all branches)	24.61	21.56	46.72	30.96
DARE (all branches)	24.58	22.40	47.28	31.42
Gated Math ($\gamma = 1.0$)	25.03	22.51	47.87	31.80
Gated Math ($\gamma = 0.5$)	25.42	23.48	48.08	32.32
RGBM Phase 2 (Ours)	23.93	24.63	50.52	33.03

RGBM improves on the merge on both suites. Writing back only the gated Math branch raises Open-MOPD to 33.03 (+0.51 over $\theta_0^\star$) with zero extra joint-training steps; on G-OPD both domains are active, and writing back both isolated branches gives 54.14, above ExOPD.

4.3 Why the Write-Back Must Be Gated: Joint Distillation from the Merge

The obvious second stage is to keep distilling $\theta_0^\star$ jointly on all domains; Section 3.2 argued that this should erode the low-residual domains. Figure 3 tracks three joint objectives initialized from $\theta_0^\star$:



(a) Macro trajectory

(b) Uniform Distillation, per domain

Figure 3: **Joint distillation from $\theta_0^\star$ (Open-MOPD, $\leq$ 50 steps).** (a) Macro under three joint objectives; dashed and dotted lines are RGBM Phase 2 and $\theta_0^\star$. (b) Uniform Distillation: Math stays near the merge before a late drop; Code and IFBench fall. Dashed lines mark RGBM Phase 2 on those axes.

Uniform (equal-weight reverse-KL), *Outcome-Gated* (matching the teacher only on failed rollouts), and *Trust-Region* (outcome-gated reverse-KL with a KL penalty $\beta = 2.5$ toward $\pi_{\theta_0^\star}$). Appendix C.2 extends this analysis to five distinct objectives. (1) **Low-residual domains deteriorate.** At the last listed checkpoint of each recipe trained to step 20 or 50, Code drops by $1.1 \sim 2.9$ points below $\theta_0^\star$, IF falls behind the merge, and Macro stays in $30.5 \sim 32.3$. Joint distillation therefore undoes what Phase 1 had already recovered. (2) **High-residual Math improves only marginally.** Uniform and Outcome-Gated trajectory means are 23.97 and 24.69 (Appendix C.2), against 23.86 at $\theta_0^\star$. Continued joint steps do not convert the Math residual into a gain that offsets the drop on Code and IF. (3) **Interference is parameter-level.** Even when the Code reverse-KL is held below 10^{-3} nat/token, Code still degrades. The student can stay close to the Code teacher in distribution space, so the damage is coming from shared-weight updates, not from a Code-teacher mismatch. (4) **A token-level KL penalty cannot prevent collapse.** The Trust-Region anchor bounds estimated policy drift to $\widehat{\mathrm{KL}}(\pi_\theta \| \pi_{\theta_0^\star}) \approx 1.4 \times 10^{-3}$ nat/token, yet Macro decays to 31.25 by step 20 (Code drops $24.09 \to 21.18$). A scalar KL only restricts isotropic divergence; it does not block interference along the anisotropic directions captured by $\overline{\mathcal{L}}_d$.

4.4 The Residual Spectrum Identifies Which Updates Can Be Written Back Safely

The normalized spectrum $\overline{\boldsymbol{\mathcal{L}}} = (0.128, 0.052, 0.015)$ is computed from the Phase 1 solution before any Phase 2 update, and the gate it induces admits only the Math branch. We test this decision against the alternatives it rejects. Writing back the admitted Math update ($\gamma^\star = 0.125$) raises Macro to 33.03. Writing back the rejected IF update at a mid-grid factor ($\theta_0^\star + 0.5\,\Delta_{\text{if}}$; IF has no $\gamma_{\text{if}}^\star$) lowers Macro to 31.10 and Code to 20.70 (Appendix C.4). Writing back all three, as un-gated TIES and DARE do, gives 30.96 and 31.42, both below the merge. On G-OPD the normalized spectrum clusters, the gate keeps both domains, and writing back both raises the average by 0.80.

Table 2: **G-OPD** (Qwen3-4B). Math / Code: mean of the 4 math / 3 code tasks. Avg: 7-task mean. Italic: teacher oracle.

Method	**Math**	**Code**	**Avg**
Domain-Matched Teacher	*46.00*	*61.17*	*52.50*
SFT	44.33	60.80	51.39
OPD	46.38	60.57	52.46
ExOPD	**47.65**	61.97	53.79
Task Arithmetic	43.05	60.65	50.59
Equal-Weight Average	22.55	55.67	36.75
RGBM Phase 1 (Ours)	46.17	62.89	53.34
RGBM Phase 2 (Ours)	47.55	**62.93**	**54.14**

We stress what this does and does not show. Isolated branches raise their own domain under all three objectives (Appendix Table 9): IF by $+0.48 \sim +1.63$, Code by $+0.43 \sim +1.31$, and Math by $+0.58 \sim +1.83$. A low residual therefore does not mean a domain has nothing left to learn on its own: $\overline{\mathcal{L}}_d$ is the coupled gap at $\boldsymbol{\alpha}^\star$, after all domains have been traded off together. A write-back is one shared update, so it moves domains other than the branch that was trained, and it is scored on Math, Code, and IF together. On that joint score the rejected IF write-back lowers Macro, while the admitted Math write-back raises it only at $\gamma^\star = 0.125$. The spectrum chooses which branch may enter the shared model. The shrinkage grid chooses how far that branch may move it.

4.5 Ablations: Gate, Isolation, and Shrinkage

The lower half of Table 1 isolates each RGBM component (full γ grid in Appendix C.3). Uniform Distillation continues joint training; TIES and DARE write back every isolated branch; Gated Math writes back only the admitted branch, with or without shrinkage. Without the gate, every write-back falls below the merge, whether updates are trained jointly or in isolation, dense or sparsified, because the Code and IF updates damage shared weights. With the gate but $\gamma = 1$, the Math branch alone still costs 0.72 Macro; intermediate γ trades Math against Code and IF (Appendix C.3). The unlabelled rule selects $\gamma^\star = 0.125$, the only setting above the merge.

Where does the gain come from? The $+0.51$ Macro of RGBM is not a Math gain: at $\gamma^\star = 0.125$, Math moves only $+0.07$. It is carried by Code ($+0.54$) and IF ($+0.92$), the two passivated domains, which the small Math write-back moves through the shared weights. Protection explains a different comparison: relative to the merge, the un-gated write-backs lose $1.1 \sim 1.6$ Macro, and the gate removes that loss. Larger γ improves Math ($+1.56$ at $\gamma = 0.5$) only by sacrificing Code and IF. Across all checkpoints of Appendix C.2, the ordering un-gated write-backs $<$ merge $<$ RGBM holds.

4.6 Linearity Diagnostics and the Limits of the Theory

Appendix Figure 4 is a sanity check of Proposition 2 on the isotropic family $\theta(\alpha) = W_{\text{ref}} + \alpha \sum_d \tau_d$ of Remark 1, not on the Minimax vector $\boldsymbol{\alpha}^\star$. Regressing the merged log-ratio on the superposed reward over 38M tokens yields $R^2 = 0.935$ at $\alpha = 0.75$. At $\alpha = 1.50$ the fit breaks ($R^2 = 0.557$), the cross-domain interference ratio $\mathcal{I}_{\text{cross}}$ (the share of merged log-ratio energy not explained by the in-domain reward) rises to 0.775, and Macro falls to 22.48—a failure Err_{lin} flags without labels. Branch updates are not small relative to the merge ($\|\Delta_d\|/\|\mathcal{T}\boldsymbol{\alpha}^\star\| \approx 22\% \sim 29\%$), and the unshrunk Math write-back already breaks the fit ($\text{Err}_{\text{lin}} = 0.72$ at $\gamma = 1$); shrinking to $\gamma^\star = 0.125$ brings it back to $0.070 \leq \delta_{\text{tol}}$ (Appendix Table 8).

5 Related Work

Parameter-space merging. Training-free fusion spans Task Arithmetic (Ilharco et al., 2022; 2023), sparsification (Yadav et al., 2023; Yu et al., 2024; Davari & Belilovsky, 2024), routing (Lu et al., 2024), and second-order schemes such as Fisher averaging (Matena & Raffel, 2022; Roy et al., 2025) and RegMean (Jin et al., 2023). Whereas AdaMerging (Yang et al., 2024) optimizes black-box weights, Phase 1 obtains merge coefficients from a residual program on unlabelled prompts and normalizes the resulting per-domain residuals for diagnosis; Remark 1 identifies the related subspace Fisher projection of the quadratic. Extending tangent-space analyses of merging (Ortiz-Jiménez et al., 2023), Proposition 2 establishes reward-space linearity, and Err_{lin} diagnoses departures from it.

On-policy and multi-teacher distillation. Beyond classical distillation (Hinton et al., 2015; Kim & Rush, 2016; You et al., 2017), on-policy distillation trains on student rollouts (Gu et al., 2024; Agarwal et al., 2024; Gudibande et al., 2023; Zhong et al., 2026; Cheng et al., 2026; Liu et al., 2026; Yan et al., 2026; Liu et al., 2025b); multi-teacher variants (Gao et al., 2026; Sun et al., 2026; Yang et al., 2026b; Chen et al., 2026) consolidate specialists via joint training from base weights. We show that, to second order, this objective has an optimal closed-form stage, and we gate subsequent updates by the normalized residual spectrum.

6 Conclusion

We bridge parameter merging and multi-teacher distillation by showing that to second order, the MOPD objective coincides with Fisher merging. Parameter merging is thus optimal, to second order, as a closed-form stage of distillation. For tractability, we restrict that merge to the task-vector subspace and, because the domain weights are unknown, replace them by their worst case: a residual program on unlabelled prompts yields the task-vector weights, and post-hoc reward-energy normalization yields the domain residual spectrum. Empirically, our zero-training merge beats a distilled student on Open-MOPD and approaches the strongest trained baseline on G-OPD, while residual-gated RGBM advances performance further. Un-gated write-backs degrade below the merge, confirming that post-merge gains are localized, with the normalized spectrum identifying the updates to test for safe write-back. Limitations are discussed in Appendix D.

7 LLM USAGE

In accordance with conference guidelines, we disclose the use of large language models (LLMs) in the preparation of this manuscript. Mainstream LLMs were used exclusively for grammar correction, stylistic polishing, and editorial refinement to enhance the readability and fluency of the text.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Yang Bai, Kaiyuan Liu, Ziyuan Zhuang, Jiahong Zhou, Rongxiang Weng, Xin Chen, Jingang Wang, and Xunliang Cai. Multi-objective and mixed-reward reinforcement learning via reward-decorrelated policy optimization. *arXiv preprint arXiv:2605.13641*, 2026.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Tianlei Chen, Jiao Ou, Ziyuan Liu, Ruiming Tang, Jian Liang, and Han Li. Counteraction-aware multi-teacher on-policy distillation for general capability recovery with domain preservation. *CoRR*, abs/2605.27115, 2026. doi: 10.48550/ARXIV.2605.27115. URL https://doi.org/10.48550/arXiv.2605.27115.

Xueqi Cheng, Xugui Zhou, Tyler Derr, and Yushun Dong. Read: Reinforcement-guided capability distillation for large language models. *CoRR*, abs/2605.11290, 2026. doi: 10.48550/ARXIV.2605.11290. URL https://doi.org/10.48550/arXiv.2605.11290.

MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXV*, volume 15133 of *Lecture Notes in Computer Science*, pp. 270–287. Springer, 2024. doi: 10.1007/978-3-031-73226-3_16. URL https://doi.org/10.1007/978-3-031-73226-3_16.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

Huan-ang Gao, Haohan Chi, Yong Yan, Shiyuan Feng, Hanlin Wu, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Open-mopd: Diagnosing and fixing capability imbalance in multi-teacher on-policy distillation. *CoRR*, abs/2608.19098, 2026. doi: 10.48550/ARXIV.2608.19098. URL https://doi.org/10.48550/arXiv.2608.19098.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. volume abs/2305.15717, 2023. doi: 10.48550/ARXIV.2305.15717. URL https://doi.org/10.48550/arXiv.2305.15717.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/bc6cddcd5d325e1c0f826066c1ad0215-Abstract-Conference.html.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=FCnohuR6AnM.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1317–1327. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1139. URL https://doi.org/10.18653/v1/d16-1139.

Kaiyuan Liu, Chen Shen, Zhanwei Zhang, Junjie Liu, Xiaosong Yuan, and Jieping Ye. Efficient reasoning through suppression of self-affirmation reflections in large reasoning models. *CoRR*, abs/2506.12353, 2025a. doi: 10.48550/ARXIV.2506.12353. URL https://doi.org/10.48550/arXiv.2506.12353.

Kaiyuan Liu, Shaotian Yan, Rui Miao, Bing Wang, Chen Shen, Jun Zhang, and Jieping Ye. Where did this sentence come from? tracing provenance in llm reasoning distillation. *arXiv preprint arXiv:2512.20908*, 2025b.

Kaiyuan Liu, Ziyuan Zhuang, Yang Bai, Bing Wang, Rongxiang Weng, and Jieping Ye. Prefix teach, suffix fade: Local teachability collapse in strong-to-weak on-policy distillation. *arXiv preprint arXiv:2605.13643*, 2026.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. Twin-merging: Dynamic integration of modular expertise in model merging. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8fcd17eb91bae20d9826786d7d6be799-Abstract-Conference.html.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/70c26937fbf3d4600b69a129031b66ec-Abstract-Conference.html.

Guillermo Ortiz-Jiménez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d28077e5ff52034cd35b4aa15320caea-Abstract-Conference.html.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hanna Hajishirzi. Generalizing verifiable instruction following. 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/46499a0622ecf568b72d17b61e45dbd5-Abstract-Datasets_and_Benchmarks_Track.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Aniruddha Roy, Jyoti Patel, Aman Chadha, Vinija Jain, and Amitava Das. Alignmerge - alignment-preserving large language model merging via fisher-guided geometric constraints. *CoRR*, abs/2512.16245, 2025. doi: 10.48550/ARXIV.2512.16245. URL https://doi.org/10.48550/arXiv.2512.16245.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025*, pp. 1279–1297. ACM, 2025. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Zechen Sun, Zhiwei Zhang, Fei Zhao, Juntao Li, Mu Chuan, Huayu Deng, Guojian Zhan, Wenliang Chen, Yao Hu, and Min Zhang. D^3-mopd: Adaptive dynamic domain scheduling for efficient multi-teacher distillation. *CoRR*, abs/2608.24987, 2026. doi: 10.48550/ARXIV.2608.24987. URL https://doi.org/10.48550/arXiv.2608.24987.

Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. *CoRR*, abs/2501.12599, 2025a. doi: 10.48550/ARXIV.2501.12599. URL https://doi.org/10.48550/arXiv.2501.12599.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025b. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html.

Shaotian Yan, Kaiyuan Liu, Chen Shen, Bing Wang, Sinan Fan, Jun Zhang, Yue Wu, and Zheng Wang. Jieping ye. distribution-aligned sequence distillation for superior long-cot reasoning. *arXiv preprint arXiv:2601.09088*, 2026.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=nZP6NgD3QY.

Enneng Yang, Li Shen, Guibing Guo, Xing-Wei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Comput. Surv.*, 58(8):216:1–216:41, 2026a. doi: 10.1145/3787849. URL https://doi.org/10.1145/3787849.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *CoRR*, abs/2602.12125, 2026b. doi: 10.48550/ARXIV.2602.12125. URL https://doi.org/10.48550/arXiv.2602.12125.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pp. 1285–1294. ACM, 2017. doi: 10.1145/3097983.3098135. URL https://doi.org/10.1145/3097983.3098135.

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 Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/yu24p.html.

Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. Weak-to-strong extrapolation expedites alignment. volume abs/2404.16792, 2024. doi: 10.48550/ARXIV.2404.16792. URL https://doi.org/10.48550/arXiv.2404.16792.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. SOD: step-wise on-policy distillation for small language model agents. *CoRR*, abs/2605.07725, 2026. doi: 10.48550/ARXIV.2605.07725. URL https://doi.org/10.48550/arXiv.2605.07725.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911, 2023. doi: 10.48550/ARXIV.2311.07911. URL https://doi.org/10.48550/arXiv.2311.07911.

This appendix is organized as follows. Appendix A proves Theorem 1 and Proposition 2, derives the residual, its scale-normalized spectrum, and Minimax convexity, and bounds the linearization error of shrinkage re-merging. Appendix B reports the six-benchmark Phase 1 breakdown, the empirical reward correlation, the $K = 2$ pairwise merges, generation lengths, and the full G-OPD suite. Appendix C details the joint-distillation recipes, the post-hoc γ write-backs, the ϵ allocation cuts, training infrastructure, the cost of assembling M_d, and the Open-MOPD RGBM Phase 2 specification. Appendix D states the scope of the linear-basin guarantee, the specialist count, and the source of the Open-MOPD margin.

A Mathematical Proofs and Derivations

This section records the proofs deferred from Section 2: the second-order equivalence of multi-teacher distillation to Fisher merging (Appendix A.1), the first-order reward homomorphism (Appendix A.2), the second-moment residual, its scale-normalized spectrum, and uniqueness of the Minimax QP (Appendix A.3), and the shrinkage bound on the linearization error of a re-merged branch (Appendix A.4).

A.1 Proof of Theorem 1 (Second-Order Equivalence to Fisher Merging)

Let $\mathcal{X}_d$ denote the prompt space for domain d, and $\mathcal{Y}$ denote the token output space. For a given prompt $x \sim \mathcal{D}_d$, the reverse Kullback-Leibler (KL) divergence between student policy $\pi_\theta(\cdot \mid x)$ and teacher policy $\pi_{W_{\phi_d}}(\cdot \mid x)$ is:

$$D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{W_{\phi_d}}(\cdot \mid x)) = \sum_{y \in \mathcal{Y}} \pi_\theta(y \mid x) \log \frac{\pi_\theta(y \mid x)}{\pi_{W_{\phi_d}}(y \mid x)}. \tag{12}$$

We define the expected domain divergence function $f_d : \mathbb{R}^P \to \mathbb{R}$ as:

$$f_d(\theta) := \mathbb{E}_{x \sim \mathcal{D}_d} \left[D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{W_{\phi_d}}(\cdot \mid x)) \right]. \tag{13}$$

We perform a multivariate Taylor expansion of $f_d(\theta)$ around the expert weight point $\theta = W_{\phi_d}$:

$$f_d(\theta) = f_d(W_{\phi_d}) + \nabla_\theta f_d(W_{\phi_d})^\top (\theta - W_{\phi_d}) + \frac{1}{2} (\theta - W_{\phi_d})^\top \nabla_\theta^2 f_d(W_{\phi_d}) (\theta - W_{\phi_d}) + \mathcal{O}(\|\theta - W_{\phi_d}\|^3). \tag{14}$$

We evaluate each term:

1. **Zeroth-order term:** At $\theta = W_{\phi_d}$, the distributions are identical:

$$f_d(W_{\phi_d}) = \mathbb{E}_{x \sim \mathcal{D}_d} \left[\sum_y \pi_{W_{\phi_d}}(y \mid x) \log \frac{\pi_{W_{\phi_d}}(y \mid x)}{\pi_{W_{\phi_d}}(y \mid x)} \right] = 0. \tag{15}$$

2. **First-order gradient term:** Taking the gradient of the reverse KL with respect to θ:

$$\begin{aligned} \nabla_\theta D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{W_{\phi_d}}(\cdot \mid x)) &= \nabla_\theta \sum_y \pi_\theta(y \mid x) \log \pi_\theta(y \mid x) \\ &\quad - \nabla_\theta \sum_y \pi_\theta(y \mid x) \log \pi_{W_{\phi_d}}(y \mid x) \\ &= \sum_y \nabla_\theta \pi_\theta(y \mid x) \left(\log \pi_\theta(y \mid x) + 1 - \log \pi_{W_{\phi_d}}(y \mid x) \right) \\ &= \sum_y \nabla_\theta \pi_\theta(y \mid x) \log \frac{\pi_\theta(y \mid x)}{\pi_{W_{\phi_d}}(y \mid x)} + \sum_y \nabla_\theta \pi_\theta(y \mid x). \end{aligned} \tag{16}$$

   Evaluating at $\theta = W_{\phi_d}$: the log ratio $\log(\pi_{W_{\phi_d}} / \pi_{W_{\phi_d}}) = \log 1 = 0$. Furthermore, because $\sum_y \pi_\theta(y \mid x) = 1$ for all θ, the sum of gradients vanishes: $\sum_y \nabla_\theta \pi_\theta(y \mid x) = \nabla_\theta \sum_y \pi_\theta(y \mid x) = \nabla_\theta 1 = \mathbf{0}$. Therefore:

$$\nabla_\theta f_d(W_{\phi_d}) = \mathbf{0}. \tag{17}$$

3. **Second-order Hessian term:** Differentiating the gradient expression again with respect to θ:

$$\nabla^2_\theta D_{\mathrm{KL}}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{W_{\phi_d}}(\cdot \mid x)) = \sum_y \nabla^2_\theta \pi_\theta(y \mid x) \left(\log \frac{\pi_\theta(y \mid x)}{\pi_{W_{\phi_d}}(y \mid x)} + 1 \right) + \sum_y \nabla_\theta \pi_\theta(y \mid x) \nabla_\theta \log \pi_\theta(y \mid x)^\top. \tag{18}$$

Evaluating at $\theta = W_{\phi_d}$: the log ratio vanishes, leaving $\sum_y \nabla^2_\theta \pi_{W_{\phi_d}}(y \mid x) = \nabla^2_\theta (\sum_y \pi_{W_{\phi_d}}(y \mid x)) = \mathbf{0}$. Using the identity $\nabla_\theta \pi_\theta = \pi_\theta \nabla_\theta \log \pi_\theta$, the remaining term becomes:

$$\begin{aligned} \nabla^2_\theta f_d(W_{\phi_d}) &= \mathbb{E}_{x \sim \mathcal{D}_d} \left[\sum_y \pi_{W_{\phi_d}}(y \mid x) \nabla_\theta \log \pi_{W_{\phi_d}}(y \mid x) \nabla_\theta \log \pi_{W_{\phi_d}}(y \mid x)^\top \right] \\ &= \mathbb{E}_{x \sim \mathcal{D}_d, y \sim \pi_{W_{\phi_d}}(\cdot \mid x)} \left[\nabla_\theta \log \pi_{W_{\phi_d}}(y \mid x) \nabla_\theta \log \pi_{W_{\phi_d}}(y \mid x)^\top \right] =: F_d, \end{aligned} \tag{19}$$

which is precisely the empirical Fisher Information Matrix $F_d \in \mathbb{R}^{P \times P}$ of expert policy $\pi_{W_{\phi_d}}$.

Substituting these three results back into the multi-teacher objective $J_{\mathrm{MOPD}}(\theta) = \sum_{d=1}^K p_d f_d(\theta)$ yields:

$$J_{\mathrm{MOPD}}(\theta) = \frac{1}{2} \sum_{d=1}^K p_d (\theta - W_{\phi_d})^\top F_d (\theta - W_{\phi_d}) + \mathcal{O}\left(\max_d \|\theta - W_{\phi_d}\|^3 \right). \tag{20}$$

Taking the gradient with respect to θ and setting it to zero:

$$\nabla_\theta J_{\mathrm{MOPD}}(\theta) = \sum_{d=1}^K p_d F_d (\theta - W_{\phi_d}) = \mathbf{0} \implies \left(\sum_{d=1}^K p_d F_d \right) \theta = \sum_{d=1}^K p_d F_d W_{\phi_d}. \tag{21}$$

Assuming $\sum_{d=1}^K p_d F_d$ is non-singular, we obtain the global unconstrained Fisher merge:

$$\theta^\star_{\mathrm{Fisher}} = \left(\sum_{d=1}^K p_d F_d \right)^{-1} \sum_{d=1}^K p_d F_d W_{\phi_d}. \tag{22}$$

This completes the proof of Theorem 1. ■

A.2 Proof of Proposition 2 (First-Order Homomorphism)

Let $f(x; \theta) \in \mathbb{R}^{|V|}$ denote the logit vector at a generation step under prompt-prefix x. The token distribution is given by the softmax function:

$$\pi_\theta(y \mid x) = \frac{\exp(f(x;\theta)[y])}{\sum_{y' \in V} \exp(f(x;\theta)[y'])} = \frac{\exp(f(x;\theta)[y])}{Z(x;\theta)}. \tag{23}$$

Taking the logarithm gives $\log \pi_\theta(y \mid x) = f(x;\theta)[y] - \log Z(x;\theta)$. Write $\Delta \log Z_d(x) := \log Z(x; W_{\phi_d}) - \log Z(x; W_{\mathrm{ref}})$ and $\Delta \log Z_{\boldsymbol{\alpha}}(x) := \log Z(x; \theta(\boldsymbol{\alpha})) - \log Z(x; W_{\mathrm{ref}})$.

Let $\theta = W_{\mathrm{ref}} + \tau$. A first-order Taylor expansion of $f(x;\theta)[y]$ around W_{ref} yields:

$$f(x; W_{\mathrm{ref}} + \tau)[y] = f(x; W_{\mathrm{ref}})[y] + \nabla_\theta f(x; W_{\mathrm{ref}})[y]^\top \tau + \mathcal{O}(\|\tau\|^2). \tag{24}$$

Applying this to specialist weights $W_{\phi_d} = W_{\mathrm{ref}} + \tau_d$, the implicit reward is:

$$r_d(x, y) = \log \frac{\pi_{W_{\phi_d}}(y \mid x)}{\pi_{W_{\mathrm{ref}}}(y \mid x)} = \nabla_\theta f(x; W_{\mathrm{ref}})[y]^\top \tau_d - \Delta \log Z_d(x) + \mathcal{O}(\|\tau_d\|^2). \tag{25}$$

The log-partition function $\log Z(x;\theta) = \log \sum_{y' \in V} \exp(f(x;\theta)[y'])$ is a smooth function of θ, so the same first-order expansion around W_{ref} gives

$$\log Z(x; W_{\mathrm{ref}} + \tau) - \log Z(x; W_{\mathrm{ref}}) = \nabla_\theta \log Z(x; W_{\mathrm{ref}})^\top \tau + \mathcal{O}(\|\tau\|^2). \tag{26}$$

Taking $\tau = \tau_d$ gives $\Delta \log Z_d(x) = \nabla_\theta \log Z(x; W_{\text{ref}})^\top \tau_d + \mathcal{O}(\|\tau_d\|^2)$, and taking $\tau = \mathcal{T}\boldsymbol{\alpha} = \sum_d \alpha_d \tau_d$ gives, by linearity of the inner product,

$$\begin{aligned}\Delta \log Z_{\boldsymbol{\alpha}}(x) &= \sum_{d=1}^{K} \alpha_d \nabla_\theta \log Z(x; W_{\text{ref}})^\top \tau_d + \mathcal{O}(\|\mathcal{T}\boldsymbol{\alpha}\|^2) \\ &= \sum_{d=1}^{K} \alpha_d \Delta \log Z_d(x) + \mathcal{O}\Big(\|\mathcal{T}\boldsymbol{\alpha}\|^2 + \sum_d |\alpha_d| \|\tau_d\|^2\Big). \end{aligned} \tag{27}$$

Now, for any linear combination $\theta(\boldsymbol{\alpha}) = W_{\text{ref}} + \sum_{d=1}^{K} \alpha_d \tau_d$, by linearity of the gradient operator and Eq. 27:

$$\begin{aligned}\log \frac{\pi_{\theta(\boldsymbol{\alpha})}(y \mid x)}{\pi_{W_{\text{ref}}}(y \mid x)} &= \nabla_\theta f(x; W_{\text{ref}})[y]^\top \left(\sum_{d=1}^{K} \alpha_d \tau_d\right) - \Delta \log Z_{\boldsymbol{\alpha}}(x) + \mathcal{O}\left(\left\|\sum_d \alpha_d \tau_d\right\|^2\right) \\ &= \sum_{d=1}^{K} \alpha_d \Big(\nabla_\theta f(x; W_{\text{ref}})[y]^\top \tau_d - \Delta \log Z_d(x)\Big) + \mathcal{O}\Big(\|\mathcal{T}\boldsymbol{\alpha}\|^2 + \sum_d |\alpha_d| \|\tau_d\|^2\Big) \\ &= \sum_{d=1}^{K} \alpha_d r_d(x, y) + \mathcal{O}\Big(\|\mathcal{T}\boldsymbol{\alpha}\|^2 + \sum_d |\alpha_d| \|\tau_d\|^2\Big) \approx \boldsymbol{\alpha}^\top \mathbf{r}(x, y), \end{aligned} \tag{28}$$

where the last equality substitutes Eq. 25 for each r_d. This proves the First-Order Homomorphism between weight combinations and reward superposition. ■

A.3 Derivation of Second-Moment Residuals and Minimax Convexity

Connection to the Projected Fisher Information Metric. Under the first-order logit expansion (Proposition 2), for any specialist d and prompt-prefix x:

$$\mathcal{T}^\top \nabla_\theta \log \pi_{W_{\phi_d}}(y \mid x) = \mathbf{r}(x, y) - \boldsymbol{\mu}_d(x) + \mathcal{O}(\max_k \|\tau_k\|^2), \tag{29}$$

where $\boldsymbol{\mu}_d(x) := \mathbb{E}_{y' \sim \pi_{W_{\phi_d}}(\cdot|x)}[\mathbf{r}(x, y')]$. The projected Fisher Information Matrix $S_d := \mathcal{T}^\top F_d \mathcal{T} \in \mathbb{R}^{K \times K}$ satisfies:

$$\begin{aligned} S_d &= \mathbb{E}_{x \sim \mathcal{D}_d, y \sim \pi_{W_{\phi_d}}} \left[\big(\mathcal{T}^\top \nabla_\theta \log \pi_{W_{\phi_d}}(y \mid x)\big)\big(\mathcal{T}^\top \nabla_\theta \log \pi_{W_{\phi_d}}(y \mid x)\big)^\top\right] \\ &\approx \mathbb{E}_{x \sim \mathcal{D}_d, y \sim \pi_{W_{\phi_d}}} \left[(\mathbf{r}(x, y) - \boldsymbol{\mu}_d(x))(\mathbf{r}(x, y) - \boldsymbol{\mu}_d(x))^\top\right] = \mathrm{Cov}_{\pi_{W_{\phi_d}}}(\mathbf{r}). \end{aligned} \tag{30}$$

Now, expanding the second-moment matrix $M_d := \mathbb{E}_{x \sim \mathcal{D}_d, y \sim \pi_{W_{\phi_d}}}[\mathbf{r}(x, y)\mathbf{r}(x, y)^\top]$:

$$M_d = \mathrm{Cov}_{\pi_{W_{\phi_d}}}(\mathbf{r}) + \mathbb{E}_{x \sim \mathcal{D}_d}[\boldsymbol{\mu}_d(x)\boldsymbol{\mu}_d(x)^\top] \approx S_d + \mathbb{E}_{x \sim \mathcal{D}_d}[\boldsymbol{\mu}_d(x)\boldsymbol{\mu}_d(x)^\top]. \tag{31}$$

For any weight combination vector $\boldsymbol{\alpha} \in \mathbb{R}^K$, writing $\mathbf{e}_d \in \mathbb{R}^K$ for the d-th standard basis vector, the residual optimized in Phase 1 is

$$\mathcal{L}_d(\boldsymbol{\alpha}) := (\boldsymbol{\alpha} - \mathbf{e}_d)^\top M_d (\boldsymbol{\alpha} - \mathbf{e}_d) \approx (\boldsymbol{\alpha} - \mathbf{e}_d)^\top S_d (\boldsymbol{\alpha} - \mathbf{e}_d) + \mathbb{E}_{x \sim \mathcal{D}_d}\big[\big((\boldsymbol{\alpha} - \mathbf{e}_d)^\top \boldsymbol{\mu}_d(x)\big)^2\big]. \tag{32}$$

$\mathcal{L}_d$ upper-bounds the projected Fisher term $2J_{\text{sub},d}(\boldsymbol{\alpha}) = (\boldsymbol{\alpha} - \mathbf{e}_d)^\top S_d (\boldsymbol{\alpha} - \mathbf{e}_d)$. After solving the Minimax problem, we form the normalized residual $\overline{\mathcal{L}}_d(\boldsymbol{\alpha}) := \mathcal{L}_d(\boldsymbol{\alpha})/\mathbb{E}[r_d^2]$. This post-hoc normalization expresses the remainder as a fraction of domain-d reward energy and makes the spectrum comparable across domains.

Term-wise Decomposition. Writing out the quadratic with $(M_d)_{jk} = \mathbb{E}[r_j r_k]$ and then dividing by $\mathbb{E}[r_d^2]$ yields the normalized diagnostic

$$\overline{\mathcal{L}}_d(\boldsymbol{\alpha}) = (\alpha_d - 1)^2 + \sum_{j \neq d} \alpha_j^2 \frac{\mathbb{E}[r_j^2]}{\mathbb{E}[r_d^2]} + 2(\alpha_d - 1) \sum_{j \neq d} \alpha_j \frac{\mathbb{E}[r_d r_j]}{\mathbb{E}[r_d^2]} + 2 \sum_{\substack{j<k \\ j,k \neq d}} \alpha_j \alpha_k \frac{\mathbb{E}[r_j r_k]}{\mathbb{E}[r_d^2]}. \tag{33}$$

The first term is squared relative under-recovery of the in-domain reward (it vanishes at $\alpha_d = 1$). The remaining terms are cross-domain interference measured in the same units. The correlation structure of Appendix B.3 governs the cross-terms.

Proposition 3 (Convexity of the Minimax QP) *The objective function* $\Phi(\boldsymbol{\alpha}) := \max_{d\in\{1,\dots,K\}} \mathcal{L}_d(\boldsymbol{\alpha})$ *is strictly convex in* $\boldsymbol{\alpha}$ *whenever each* $M_d \succ 0$.

Proof. Each component has Hessian $\nabla^2\mathcal{L}_d(\boldsymbol{\alpha}) = 2M_d$. Since the finite collection $\{M_d\}_{d=1}^K$ is positive definite, all components share a positive strong-convexity lower bound. Their pointwise maximum is therefore strongly convex, so the Minimax problem has a unique global minimizer $\boldsymbol{\alpha}^\star$. In practice, replacing M_d by $M_d + \lambda I$ supplies this condition. ■

A.4 Theoretical Justification of Branch Re-merging within Local Tangent Basins

A natural question is whether the branch updates $\Delta_d = \theta_d - \theta_0^\star$ obtained in Phase 2, once written back with shrinkage, remain within the regime in which the first-order reward superposition of Proposition 2 is accurate. The following statement makes precise the role of the contraction factor γ.

Proposition 4 (Shrinkage Tightens the Linearization Error) *Fix a prompt-prefix* x *and token* y*, and let* $g(\theta) := f(x;\theta)[y]$ *denote the corresponding logit. Assume* g *is twice differentiable with* $\|\nabla^2_\theta g(\theta)\| \le L_H$ *on a convex neighborhood* $\mathcal{B}$ *containing* W_{ref}*,* $\theta_0^\star = W_{\mathrm{ref}} + \mathcal{T}\boldsymbol{\alpha}^\star$*, and the branch endpoint* $\theta_0^\star + \Delta_d$*. Let* $\hat{g}(\theta) := g(W_{\mathrm{ref}}) + \nabla_\theta g(W_{\mathrm{ref}})^\top(\theta - W_{\mathrm{ref}})$ *be the first-order model underlying Proposition 2. Then for every* $\gamma \in [0,1]$*, the re-merged point* $\theta_\gamma := \theta_0^\star + \gamma\Delta_d$ *satisfies*

$$\left|g(\theta_\gamma) - \hat{g}(\theta_\gamma)\right| \;\le\; \frac{L_H}{2}\left\|\mathcal{T}\boldsymbol{\alpha}^\star + \gamma\Delta_d\right\|^2 \;\le\; \frac{L_H}{2}\Big(\|\mathcal{T}\boldsymbol{\alpha}^\star\| + \gamma\|\Delta_d\|\Big)^2, \tag{34}$$

and the right-hand side is non-decreasing in γ*. In particular, the linearization error at the re-merged model is at most that of the un-shrunk branch endpoint (*$\gamma = 1$*) and approaches that of* $\theta_0^\star$ *as* $\gamma \to 0$*; when* $\|\Delta_d\| \ll \|\mathcal{T}\boldsymbol{\alpha}^\star\|$*, the excess error over* $\theta_0^\star$ *is* $L_H\gamma\|\Delta_d\|\|\mathcal{T}\boldsymbol{\alpha}^\star\| + \mathcal{O}(\gamma^2\|\Delta_d\|^2)$*, i.e., first order in* $\gamma\|\Delta_d\|$.

Proof. Since θ_γ is a convex combination of $\theta_0^\star$ and $\theta_0^\star + \Delta_d$, it lies in $\mathcal{B}$. Taylor's theorem with Lagrange remainder along the segment from W_{ref} to θ_γ gives $g(\theta_\gamma) - \hat{g}(\theta_\gamma) = \frac{1}{2}(\theta_\gamma - W_{\mathrm{ref}})^\top\nabla^2_\theta g(\xi)(\theta_\gamma - W_{\mathrm{ref}})$ for some $\xi \in \mathcal{B}$, and $\theta_\gamma - W_{\mathrm{ref}} = \mathcal{T}\boldsymbol{\alpha}^\star + \gamma\Delta_d$. Bounding the Hessian by L_H and applying the triangle inequality yields Eq. 34. Monotonicity in γ follows because $\|\mathcal{T}\boldsymbol{\alpha}^\star\| + \gamma\|\Delta_d\|$ is affine and increasing in γ. Expanding the square gives the excess-error expression. The same argument applies to the log-partition term $\log Z(x;\theta)$, which is a smooth function of the logit vector, so the bound transfers to the log-ratio $\log \pi_{\theta_\gamma}(y \mid x)/\pi_{W_{\mathrm{ref}}}(y \mid x)$ up to a constant factor depending only on the softmax smoothness. ■

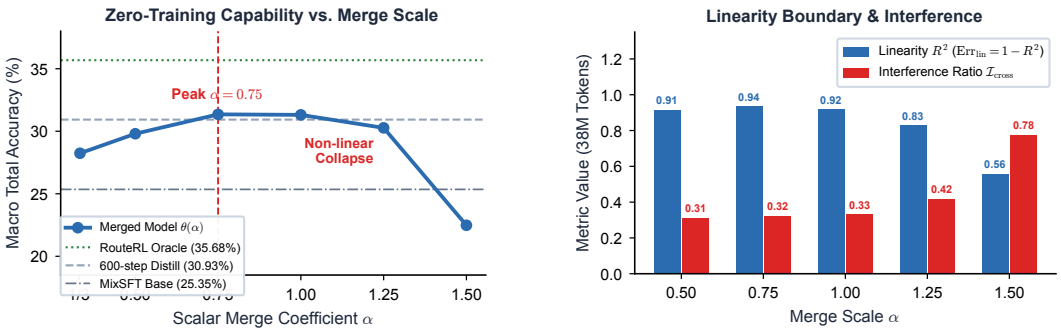
Proposition 4 does not by itself assert that $\|\Delta_d\|$ is small; it states that, whatever displacement branch training produces, shrinkage can only reduce the bound on the linearization error, and that the reduction is first order in γ. On Open-MOPD the residual gate writes back only the Math branch at the unlabelled grid value $\gamma^\star = 0.125$; passivated Code/IF contribute $\Delta_d = \mathbf{0}$.

B Phase 1 Diagnostics and Detailed Benchmark Results

This section supports Sections 2–4 with the linearity diagnostics (Appendix B.1), the six-benchmark Phase 1 breakdown (Appendix B.2), the empirical reward correlation (Appendix B.3), the $K = 2$ pairwise merges (Appendix B.4), generation lengths (Appendix B.5), and the full G-OPD suite (Appendix B.6).

B.1 Linearity Diagnostics on the Isotropic Family

Figure 4 reports the diagnostics discussed in Section 4.6: the isotropic scan $\theta(\alpha) = W_{\mathrm{ref}} + \alpha\sum_d \tau_d$ of Remark 1, the regression R^2 of the merged log-ratio on the superposed reward over 38M tokens, and the cross-domain interference ratio $\mathcal{I}_{\mathrm{cross}} := \sum_d \mathbb{E}_{\mathcal{D}_d}[(\ell_\alpha - \alpha r_d)^2]/\sum_d \mathbb{E}_{\mathcal{D}_d}[\ell_\alpha^2]$, the share of merged log-ratio energy not explained by the in-domain reward. The branch displacements used in Section 4.6 are $\|\mathcal{T}\boldsymbol{\alpha}^\star\| = 1.025$, $\|\Delta_{\mathrm{math}}\| = 0.229$, $\|\Delta_{\mathrm{code}}\| = 0.299$, $\|\Delta_{\mathrm{if}}\| = 0.265$.



(a) Capability vs. isotropic merge scale α

(b) Linearity R^2 and interference $\mathcal{I}_{\mathrm{cross}}$

Figure 4: **Linearity diagnostics on the isotropic family** ($\theta(\alpha) = W_{\mathrm{ref}} + \alpha \sum_d \tau_d$; not the Minimax vector $\boldsymbol{\alpha}^\star$). (a) Macro vs. scalar α. Horizontal lines are RouteRL (35.68), the 600-step student (30.93), and MixSFT (25.35); the peak at $\alpha = 0.75$ is this scan, not $\boldsymbol{\alpha}^\star$. (b) R^2 and $\mathcal{I}_{\mathrm{cross}}$ over 38M tokens.

B.2 Complete Per-Benchmark Phase 1 Results

Table 3 details the individual benchmark evaluations across all six downstream evaluation tasks: AIME 2024, AIME 2025 (Math), LiveCodeBench v5, LiveCodeBench v6 (Code), IFEval, and IFBench (Instruction Following). The Type column is the coefficient family: Scalar is a shared scale $\alpha\mathbf{1}$; 3D Vector is a free $\boldsymbol{\alpha} \in \mathbb{R}^3$. Official Distilled Model is the main-text Multi-Teacher OPD row. Forward-KL Merge uses forward KL in the same isotropic subspace program; Minimax Worst-Domain Merge (Scalar) is Eq. 9 restricted to $\boldsymbol{\alpha} = \alpha\mathbf{1}$. Both Subspace Fisher Merge rows solve Eq. 5 under uniform $p_d = 1/3$ with M_d in place of S_d; the main-text Subspace Fisher Merge is the 3D Vector row.

Table 3: **Comprehensive benchmark evaluation of Phase 1 candidates.** Six-task breakdown. Bold: best zero-training candidate per column (or Phase 2 where it exceeds Phase 1).

Method / Configuration	Type	Math			Code			Instruction Following			Macro
		A24	A25	Avg	v5	v6	Avg	IFEval	IFB	Avg	
MixSFT Base (W_{ref})	–	14.79	18.18	16.49	15.63	18.46	17.05	66.36	18.67	42.52	25.35
RouteRL (Oracle Upper Bound)	–	26.88	27.60	27.24	27.27	25.26	26.27	81.70	25.33	53.52	35.68
Official Distilled Model	Distill	24.38	24.68	24.53	21.05	21.31	21.18	70.24	24.00	47.12	30.93
Equal-Weight Average	Scalar	20.31	21.30	20.81	17.66	20.63	19.15	69.87	19.67	44.77	28.24
Task Arithmetic	Scalar	23.91	24.79	24.35	22.99	22.63	22.81	69.87	23.67	46.77	31.31
Subspace Fisher Merge	Scalar	**25.47**	24.48	**24.97**	23.65	22.91	23.28	73.94	21.33	47.64	31.96
Subspace Fisher Merge	3D Vector	23.85	24.38	24.11	24.49	22.69	23.59	72.64	23.33	47.99	31.90
Forward KL (FKL) Merge	Scalar	24.58	23.28	23.93	23.59	22.34	22.97	71.53	24.33	47.93	31.61
Minimax Worst-Domain Merge	Scalar	25.10	24.84	**24.97**	24.07	22.91	23.49	69.87	20.33	45.10	31.19
Minimax Merge ($\theta_0^\star$)	**3D Vector**	23.91	23.80	23.86	24.27	23.91	24.09	73.87	**25.33**	**49.60**	**32.52**
RGBM Phase 2	Residual-Gated	23.80	24.06	23.93	**25.15**	**24.11**	**24.63**	**75.05**	**26.00**	**50.52**	**33.03**

B.3 Empirical Reward Correlation Geometry

We sample 6,144 rollout sequences across domain training prompts and extract implicit rewards $\mathbf{r} = [r_{\mathrm{math}}, r_{\mathrm{code}}, r_{\mathrm{if}}]^\top$. The empirical correlation matrix is:

$$\mathrm{Corr}(\mathbf{r}) = \begin{bmatrix} 1.000 & \mathbf{0.313} & 0.043 \\ \mathbf{0.313} & 1.000 & 0.035 \\ 0.043 & 0.035 & 1.000 \end{bmatrix}. \tag{35}$$

The off-diagonal correlation between Math and Code is moderately positive (0.313), indicating that mathematical problem-solving and programming share underlying algorithmic reasoning representations. In contrast, the Instruction-Following specialist is strictly orthogonal to both reasoning domains (< 0.045). This geometric structure directly explains why the Minimax solver assigns $\alpha_{\mathrm{if}}^\star \approx 0.800$ (as the IF domain experiences near-zero destructive interference) while compressing $\alpha_{\mathrm{math}}^\star \approx 0.597$ and $\alpha_{\mathrm{code}}^\star \approx 0.620$ to minimize cross-talk.

B.4 Two-Domain ($K = 2$) Sub-Consolidations and Residual Tracking

To test whether the Fisher-geometric theory and the predictive validity of the residual spectrum hold across different expert subsets, we evaluate all three pairwise $K = 2$ specialist consolidations formed from $\{W_{\phi_{\text{math}}}, W_{\phi_{\text{code}}}, W_{\phi_{\text{if}}}\}$:

1. **Math + Code ($K = 2$):** The implicit rewards share positive covariance ($\rho = 0.313$). The Minimax solver allocates balanced weights $\boldsymbol{\alpha}^\star = [0.635, 0.652]^\top$, yielding zero-training performance of Math 24.38, Code 24.81 (Macro 24.60, higher than the 3-expert merge on Code due to zero interference from IF). The normalized spectrum ranks $\overline{\mathcal{L}}_{\text{math}} > \overline{\mathcal{L}}_{\text{code}}$, correctly identifying Math as the remaining active bottleneck.
2. **Math + IF ($K = 2$):** Because the two specialists are nearly orthogonal ($\rho = 0.043$), cross-domain interference is minimal. The Minimax weights are $\boldsymbol{\alpha}^\star = [0.680, 0.885]^\top$, achieving Math 24.82 and IF 48.95 (Macro 36.89). $\overline{\mathcal{L}}_{\text{if}}$ is near the floor (almost complete recovery), while $\overline{\mathcal{L}}_{\text{math}}$ marks Math for post-merge branch training.
3. **Code + IF ($K = 2$):** With near-zero cross-covariance ($\rho = 0.035$), linear superposition incurs minimal interference. The Minimax weights $\boldsymbol{\alpha}^\star = [0.720, 0.890]^\top$ achieve Code 25.02 and IF 49.10 (Macro 37.06). Both normalized residuals sit near the floor, predicting that both domains are passivated and require zero post-merge branch compute.

Across all $K = 2$ sub-settings, the normalized residual spectrum $\overline{\boldsymbol{\mathcal{L}}}$ correctly ranks domain bottlenecks and tracks pairwise recovery.

B.5 Token Generation Length Distributions

Table 4 summarizes the average completion lengths across evaluation benchmarks.

Table 4: **Average generation length across models.** Completion tokens on one representative task per domain (AIME 2024, LiveCodeBench v5, IFBench).

Model	**AIME 2024** (Math)	**LiveCodeBench v5** (Code)	**IFBench** (IF)
MixSFT Base (W_{ref})	3,420	485	310
Math Specialist ($W_{\phi_{\text{math}}}$)	14,850	490	280
Code Specialist ($W_{\phi_{\text{code}}}$)	3,890	720	295
IF Specialist ($W_{\phi_{\text{if}}}$)	2,950	420	315
Equal-Weight Task Arithmetic	11,240	560	410
Official Distilled Model (600 steps)	11,850	580	390
Zero-Training Merge ($\theta_0^\star$)	**12,316**	**615**	**461**

The zero-training merge $\theta_0^\star$ preserves extended reasoning lengths ($\approx$ 12.3k tokens on AIME), confirming that parameter merging successfully preserves reasoning traces without requiring expensive post-hoc length penalties.

B.6 Cross-Suite Generalization on G-OPD Benchmarks (Qwen3-4B)

To further establish that the Fisher-geometric projection generalizes across specialist regimes and model families, we evaluate RGBM on the official G-OPD benchmarks (Yang et al., 2026b) using Qwen3-4B in two specialist regimes: the intermediate checkpoints of Table 2 in Yang et al. (2026b) ($W_{\phi_{\text{math}}}$ at 500 steps, $W_{\phi_{\text{code}}}$ at 300 steps), which are the main-text G-OPD setting, and the fully trained pair of their Table 8 (both at 1200 steps). Table 5 reports the complete per-task benchmark results across all 7 evaluation datasets (AIME 2024, AIME 2025, HMMT Feb, HMMT Nov, HumanEval+ [HE+], MBPP+, and LiveCodeBench [LCB v6]) following the standardized evaluation protocol.

Phase 1 coefficients and Phase 2 allocation. On the intermediate pair the Minimax weights are $\boldsymbol{\alpha}^\star = (0.675, 0.666)$ and the normalized spectrum $\overline{\boldsymbol{\mathcal{L}}} = (0.094, 0.090)$ has its largest gap down to 0, so Eq. 10 gives $\epsilon = 0.045$ and both Math and Code stay active. Isolated distillation branches re-merged at the unlabelled grid value $\gamma = 0.5$ reach 54.14 (Table 5), $+0.80$ over Phase 1 and $+0.35$

Table 5: **Per-benchmark results on the G-OPD suite (Qwen3-4B).** Intermediate and converged specialist regimes under the official 7-task protocol. ExPO is a published G-OPD baseline. Single-Teacher distills from one specialist; Multi-Teacher uses both. Bold: best consolidated model per column (teacher excluded).

Regime	Method / Configuration	AIME24	AIME25	HMMT-Feb	HMMT-Nov	HE+	MBPP+	LCB	7-Task Avg
Intermediate (Math-500 + Code-300)	Teacher ($W_{\phi_{\text{math}}}, W_{\phi_{\text{code}}}$)	58.00	54.60	32.50	38.90	86.00	70.20	27.30	52.50
	Student Base (W_{ref})	21.50	21.90	10.00	8.00	74.70	64.70	17.90	31.24
	SFT	58.50	53.30	30.70	34.80	86.40	69.60	26.40	51.39
	ExPO (Zheng et al., 2024)	57.50	54.50	31.70	36.30	86.70	**72.00**	29.00	52.53
	On-Policy Distillation (OPD)	60.60	54.10	32.50	38.30	84.60	69.50	27.60	52.46
	ExOPD (Yang et al., 2026b)	**61.00**	56.00	**34.40**	39.20	86.30	70.60	29.00	53.79
	Task Arithmetic	56.98	50.42	28.02	36.77	84.30	68.52	29.14	50.59
	Equal-Weight Average	33.75	25.00	15.52	15.94	79.88	67.13	20.00	36.75
	RGBM Phase 1 (Ours)	59.17	55.31	32.71	37.50	**87.96**	71.43	29.29	53.34
	RGBM Phase 2 (Ours)	60.42	**56.56**	33.96	**39.27**	86.91	71.30	**30.57**	**54.14**
Converged (Math-1200 + Code-1200)	Teacher ($W_{\phi_{\text{math}}}, W_{\phi_{\text{code}}}$)	68.20	59.30	37.30	42.90	88.90	72.50	28.00	56.73
	Student Base (W_{ref})	21.50	21.90	10.00	8.00	74.70	64.70	17.90	31.24
	OPD (Single-Teacher)	68.30	58.70	**38.70**	41.20	89.30	71.30	28.00	56.50
	ExOPD (Single-Teacher)	68.40	59.20	38.20	42.80	**89.90**	73.70	29.30	57.36
	OPD (Multi-Teacher)	68.20	60.20	38.50	40.80	86.40	72.10	27.60	56.26
	ExOPD (Multi-Teacher)	**70.10**	59.60	37.50	42.70	89.50	73.90	29.70	**57.57**
	Task Arithmetic	64.79	59.48	35.10	41.56	87.50	74.74	30.86	56.29
	Equal-Weight Average	40.94	35.62	20.10	24.17	80.79	68.85	21.57	41.72
	RGBM Phase 1 (Ours)	67.50	**62.92**	36.46	**43.33**	85.37	**74.87**	**31.00**	57.35

over trained ExOPD. On the converged pair we report only Phase 1 ($\boldsymbol{\alpha}^\star = (0.511, 0.505)$, average 57.35); there is no Phase 2 run.

C Additional Experimental Details and Parameter Sensitivity

This section collects the full RGBM procedure and the systems compared (Appendix C.1), the five joint-distillation recipes (Appendix C.2), the post-hoc γ write-backs (Appendix C.3), the ϵ allocation cuts (Appendix C.4), training infrastructure (Appendix C.5), the cost and robustness of assembling M_d (Appendix C.6), and the Open-MOPD RGBM Phase 2 specification (Appendix C.7).

C.1 RGBM Procedure and Systems Compared

Algorithm 1 lists the complete RGBM procedure of Section 3; Table 6 lists how every Open-MOPD system in Table 1 is built.

Algorithm 1 Residual-Governed Branch-and-Merge (RGBM)

Require: Base weights W_{ref}, specialist weights $\{W_{\phi_d}\}_{d=1}^K$, unlabelled calibration prompts $\mathcal{D}_{\text{calib}} = \{\mathcal{D}_{\text{calib},d}\}_{d=1}^K$, Tikhonov constant λ, discrete shrinkage grid $\Gamma = \{0.125, 0.25, 0.5, 0.75\}$, linearity tolerance δ_{tol}.

1: For each d: draw N_d pairs $x_i \sim \mathcal{D}_{\text{calib},d}$, $y_i \sim \pi_{W_{\phi_d}}(\cdot \mid x_i)$; score $\mathbf{r}(x_i, y_i) = [\log \frac{\pi_{W_{\phi_k}}(y_i|x_i)}{\pi_{W_{\text{ref}}}(y_i|x_i)}]_{k=1}^K$; $M_d \leftarrow \frac{1}{N_d}\sum_i \mathbf{r}\mathbf{r}^\top$.
2: Solve the residual Minimax QP (Eq. 9) with $\tilde{M}_d = M_d + \lambda I$ for $\boldsymbol{\alpha}^\star$; set $\theta_0^\star \leftarrow W_{\text{ref}} + \mathcal{T}\boldsymbol{\alpha}^\star$; using M_d, set $\mathcal{L}_d \leftarrow (\boldsymbol{\alpha}^\star - \mathbf{e}_d)^\top M_d(\boldsymbol{\alpha}^\star - \mathbf{e}_d)$ and $\overline{\mathcal{L}}_d \leftarrow \mathcal{L}_d / \mathbb{E}[r_d^2]$ (Eq. 8).
3: Sort $\overline{\mathcal{L}}_{(1)} \geq \cdots \geq \overline{\mathcal{L}}_{(K)} \geq \overline{\mathcal{L}}_{(K+1)} := 0$; $k^\star \leftarrow \arg\max_i(\overline{\mathcal{L}}_{(i)} - \overline{\mathcal{L}}_{(i+1)})$; $\epsilon \leftarrow \frac{1}{2}(\overline{\mathcal{L}}_{(k^\star)} + \overline{\mathcal{L}}_{(k^\star+1)})$ (Eq. 10).
4: **for** $d = 1, \ldots, K$ **do**
5: **if** $\overline{\mathcal{L}}_d \leq \epsilon$ **then**
6: **Passivate:** $\Delta_d \leftarrow \mathbf{0}$
7: **else**
8: **Active branch:** $\theta_d \leftarrow \theta_0^\star$; train on $\mathcal{D}_d$ only; $\Delta_d \leftarrow \theta_d^{(\text{end})} - \theta_0^\star$.
9: Compute unlabelled shrinkage $\gamma_d^\star \leftarrow \max\{\gamma \in \Gamma \mid \text{Err}_{\text{lin}}(\theta_0^\star + \gamma\Delta_d) \leq \delta_{\text{tol}}\}$ (Eq. 11).
10: **end if**
11: **end for**
12: **Re-merge:** $\theta_{\text{RGBM}} \leftarrow \theta_0^\star + \sum_{d:\overline{\mathcal{L}}_d > \epsilon} \gamma_d^\star \Delta_d$ (Eq. 11).
13: **return** θ_{RGBM}

When evaluating $\mathrm{Err}_{\mathrm{lin}}(\theta_0^\star + \gamma\Delta_d)$ in Algorithm 1, the response is $\ell_{\theta_\gamma} = \log(\pi_{\theta_\gamma}/\pi_{W_{\mathrm{ref}}})$ and the predictor is the Phase 1 superposition $\boldsymbol{\alpha}^{\star\top}\mathbf{r}$, held fixed. Both are scored token-wise on the same Phase 1 pairs (x, y) used to assemble M_d, pooled over $\bigcup_d \mathcal{D}_{\mathrm{calib},d}$. A single OLS with intercept then yields R^2, and $\mathrm{Err}_{\mathrm{lin}} = 1 - R^2$ as in Eq. 7.

Table 6: **Consolidation systems compared on Open-MOPD.** How each row of Table 1 is built: initialization, training, and what is written back.

System	Init	Training	Written back
RouteRL (Gao et al., 2026)	K specialists	none	Routing oracle; serves K models
Multi-Teacher OPD (Gao et al., 2026)	W_{ref}	joint OPD, all domains	Full student checkpoint
Task Arithmetic (Ilharco et al., 2023)	W_{ref}	none	$\sum_d \tau_d$ ($\alpha = 1$)
Subspace Fisher Merge	W_{ref}	none	$\mathcal{T}\boldsymbol{\alpha}^\star_{\mathrm{sub}}$ (Eq. 5, $p_d = 1/3$, M_d for S_d)
RGBM Phase 1 ($\theta_0^\star$)	W_{ref}	none	$\mathcal{T}\boldsymbol{\alpha}^\star$
Uniform Distillation	$\theta_0^\star$	joint OPD, $p_d = 1/3$	Shared weights
TIES (Yadav et al., 2023) / DARE (Yu et al., 2024)	$\theta_0^\star$	isolated branch per domain	All $\{\Delta_d\}_{d=1}^3$, sparsified
Gated Math ($\gamma = 1.0$)	$\theta_0^\star$	isolated Math branch only	Δ_{math} (no shrinkage)
Gated Math ($\gamma = 0.5$)	$\theta_0^\star$	isolated Math branch only	$0.5\,\Delta_{\mathrm{math}}$
RGBM Phase 2 (Ours)	$\theta_0^\star$	the same isolated Math branch	$\gamma^\star\Delta_{\mathrm{math}}$

C.2 Joint Multi-Task Distillation: Exploration of 5 Recipes

To thoroughly investigate whether any joint fine-tuning objective could maintain or improve upon $\theta_0^\star$, we implemented and evaluated five distinct joint training recipes:

1. **Standard Multi-Teacher Distillation (Uniform MOPD):** Multi-teacher reverse-KL on student rollouts with equal domain sampling ($p_d = 1/3$).
2. **Outcome-Gated Distillation:** Outcome-gated on-policy distillation, where teacher distribution matching is active only when student reasoning fails verification.
3. **Outcome-Veto Regularized Distillation:** Regularized distillation applying explicit negative KL penalties to suppress drift on non-target domains.
4. **Trust-Region Regularized Distillation:** Outcome-gated distillation with explicit KL anchoring to $\theta_0^\star$ ($\beta = 2.5$).
5. **Hybrid Teacher-Anchor Distillation:** Convex combination of specialist teacher KL and student-to-$\theta_0^\star$ anchor KL.

Empirical Findings across Checkpoints. On the five recipes in Table 7, Macro Total stays in the $30.5\% \sim 32.3\%$ range; every listed checkpoint lies below residual-gated RGBM Phase 2 (33.03%). At the last listed checkpoint of each recipe, Code is $1.1 \sim 2.9$ points below $\theta_0^\star$, and IF stays below the Phase 1 value 49.60. Within this testbed, no joint objective we tried preserved the low-residual domains while training on the shared weights, which is the behaviour anticipated by the argument of Section 3.3.

C.3 Write-Back Does Not Raise Macro above Residual-Gated RGBM

The main-text Phase 2 scheme is residual-gated RGBM Phase 2 with Math active and Code/IF passivated. Shrinkage is chosen by the unlabelled rule (Eq. 11) on $\Gamma = \{0.125, 0.25, 0.5, 0.75\}$, which selects $\gamma^\star = 0.125$ (Macro 33.03), above the Phase 1 merge (32.52). Table 8 is a post-hoc diagnostic: the remaining grid points, a $\gamma = 1$ (no-shrinkage) foil, and equal-mass alternatives are *not* used to choose γ; none of them exceed 33.03.

Readout. Eq. 11 selects $\gamma^\star = 0.125$ from Γ using only $\mathrm{Err}_{\mathrm{lin}}$ on $\mathcal{D}_{\mathrm{calib}}$. In Table 8 that value is $0.0701 \le \delta_{\mathrm{tol}}$, while $\gamma = 0.25$ already reaches 0.1749. Post-hoc, every larger grid point and every equal-mass alternative in Table 8 has Macro ≤ 32.44. On G-OPD, where both domains require a separate OPD branch, the same unlabelled rule yields $\gamma^\star = 0.5$.

C.4 Residual Threshold Sensitivity and Passivation Robustness

In Phase 2 the threshold ϵ only matters through which domains it activates. Eq. 10 is a closed-form choice: sort the normalized spectrum $\overline{\boldsymbol{\mathcal{L}}} = (0.128, 0.052, 0.015)$, append the floor 0, take the

Table 7: **Trajectory evaluation of joint multi-teacher distillation from $\theta_0^\star$.** Five joint objectives. Trajectory Mean is the average of the listed checkpoints. Trust-Region is scored at steps 5/10/20 because that run ends at step 20.

Distillation Framework	Step	Math	Code	IF	Macro Total	Δ vs. $\theta_0^\star$
RGBM Phase 1 ($\theta_0^\star$)	0	23.86	24.09	49.60	32.52	–
RGBM Phase 2 (Ours)	0	23.93	24.63	50.52	**33.03**	**+0.51**
Uniform Distillation (Gao et al., 2026)	Step 10	24.38	20.78	47.79	30.98	-1.54
	Step 30	24.58	21.46	47.62	31.22	-1.30
	Step 50	22.94	21.83	46.84	30.54	-1.98
	Trajectory Mean	23.97	21.36	47.42	**30.91**	**-1.61**
Outcome-Gated Distillation	Step 10	24.45	21.62	48.12	31.40	-1.12
	Step 30	24.97	22.18	48.91	32.02	-0.50
	Step 50	24.64	22.25	46.97	31.28	-1.24
	Trajectory Mean	24.69	22.02	48.00	**31.57**	**-0.95**
Outcome-Veto Distillation	Step 10	25.23	21.93	47.10	31.42	-1.10
	Step 30	24.24	21.59	48.99	31.61	-0.91
	Step 50	25.08	21.20	47.73	31.33	-1.19
	Trajectory Mean	24.85	21.57	47.94	**31.45**	**-1.07**
Trust-Region Regularization ($\beta = 2.5$)	Step 5	23.57	23.33	49.99	32.30	-0.22
	Step 10	24.84	23.66	47.52	32.01	-0.51
	Step 20	24.95	21.18	47.62	31.25	-1.27
	Trajectory Mean	24.45	22.72	48.38	**31.85**	**-0.67**
Hybrid Teacher-Anchor	Step 10	24.40	21.15	47.73	31.09	-1.43
	Step 30	23.65	22.73	48.69	31.69	-0.83
	Step 50	24.69	22.99	46.38	31.35	-1.17
	Trajectory Mean	24.25	22.29	47.60	**31.38**	**-1.14**

Table 8: **Ablation of Math-branch write-back configurations.** Part A: the grid Γ plus a $\gamma = 1$ no-shrinkage foil. $\mathrm{Err}_{\mathrm{lin}}$ is evaluated on $\mathcal{D}_{\mathrm{calib}}$ and is the quantity in Eq. 11; $\delta_{\mathrm{tol}} = 0.10$. Part B: $\theta^{(1)}_{\mathrm{math}}$ and $\theta^{(2)}_{\mathrm{math}}$ are two intermediate Math-branch checkpoints; the equal-mass rows match their total displacement rather than using $\mathrm{Err}_{\mathrm{lin}}$.

Configuration	Formula / Shrinkage	Math	Code	IF	Macro Total	$\mathrm{Err}_{\mathrm{lin}}$
Part A: Grid $\Gamma = \{0.125, 0.25, 0.5, 0.75\}$ *and* $\gamma = 1$ *foil*						
$\gamma = 0.125$ (Ours, $\gamma^\star$)	$\theta_0^\star + 0.125\,\Delta_{\mathrm{math}}$	23.93	24.63	50.52	**33.03**	0.0701
$\gamma = 0.25$	$\theta_0^\star + 0.25\,\Delta_{\mathrm{math}}$	24.19	24.02	47.82	32.01	0.1749
$\gamma = 0.50$	$\theta_0^\star + 0.50\,\Delta_{\mathrm{math}}$	25.42	23.48	48.08	32.32	0.2594
$\gamma = 0.75$	$\theta_0^\star + 0.75\,\Delta_{\mathrm{math}}$	25.08	22.87	45.51	31.15	0.6442
$\gamma = 1.00$	$\theta_0^\star + 1.00\,\Delta_{\mathrm{math}}$	25.03	22.51	47.87	31.80	0.7231
Part B: Equal Total Weight Mass Alternatives						
Single Point $\theta^{(2)}_{\mathrm{math}}$	$\theta_0^\star + \frac{2}{3}\,\Delta^{(2)}_{\mathrm{math}}$	25.52	23.33	47.57	32.14	–
Single Point $\theta^{(1)}_{\mathrm{math}}$	$\theta_0^\star + \frac{2}{3}\,\Delta^{(1)}_{\mathrm{math}}$	24.64	23.84	48.18	32.22	–
Two-Point Average	$\frac{1}{3}(\theta_0^\star + \theta^{(1)}_{\mathrm{math}} + \theta^{(2)}_{\mathrm{math}})$	24.90	23.50	48.91	32.44	–

midpoint of the largest adjacent gap, and obtain $\epsilon = 0.090$ (Math active). Any other ϵ in the same gap $[0.052, 0.128)$ yields the same allocation; the four possible cuts are:

- $\epsilon \geq 0.128$: no domain is active; the write-back is empty and the checkpoint remains $\theta_0^\star$ (Macro 32.52).
- $0.052 \leq \epsilon < 0.128$ (**Eq. 10 on Open-MOPD**, $\epsilon = 0.090$): only Math is active. RGBM Phase 2 under this cut writes the unlabelled grid value $\gamma^\star = 0.125\,\Delta_{\mathrm{math}}$ (Eq. 11) and is Macro 33.03. Post-hoc write-backs at the remaining grid points and the equal-mass alternatives yield Macro $31.15 \sim 32.44$ (Table 8), all below 33.03.
- $0.015 \leq \epsilon < 0.052$: Code is additionally activated. Isolated Code branches improve their own domain ($+0.43 \sim +1.31$; Table 9). Writing every branch back, as un-gated TIES and DARE do, drops Code from 24.09 at $\theta_0^\star$ to 21.56 and 22.40 (Table 1).

Table 9: **Normalized residual spectrum vs. realized in-domain branch gains.** $\overline{\mathcal{L}}_d(\boldsymbol{\alpha}^\star)$ vs. peak in-domain accuracy minus $\theta_0^\star$ for an isolated branch trained by on-policy distillation, outcome-gated distillation, or RL. A large normalized residual does not imply a large stand-alone gain.

Domain	$\overline{\mathcal{L}}_d(\boldsymbol{\alpha}^\star)$	Rank	$\theta_0^\star$	Distill. Branch	Outcome-Gated	RL Branch
Math	0.128	1	23.86	+1.10	+1.83	+0.58
Code	0.052	2	24.09	+0.43	+1.31	+0.51
IF	0.015	3	49.60	+0.48	+1.63	+0.57

- $\epsilon < 0.015$: IF is additionally activated. Writing an IF branch back ($\theta_0^\star + 0.5\,\Delta_{\text{if}}$) reduces Code to 20.70% and Macro to 31.10%: the leakage mechanism of Section 4.3 operates even for a branch trained exclusively on IF prompts.

The Open-MOPD allocation is therefore Math-active residual-gated RGBM. On G-OPD, $\boldsymbol{\alpha}^\star = (0.675, 0.666)$ and $\overline{\boldsymbol{\mathcal{L}}} = (0.094, 0.090)$: the largest gap is down to 0, so Eq. 10 gives $\epsilon = 0.045$ and both domains stay active.

C.5 Training Infrastructure and Hyperparameters

All branch training and distillation experiments are executed using the `verl` distributed RLHF framework (Sheng et al., 2025) on clusters of 8× NVIDIA A100/H800 GPUs. The detailed hyperparameter configuration is reported in Table 10.

Table 10: **Hyperparameter configuration for Open-MOPD branch training.** G-OPD uses the official protocol of Yang et al. (2026b).

Hyperparameter	Value
Base Model Architecture	SmolLM3-3B ($P \approx 3.0 \times 10^9$)
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-8}$)
Peak Learning Rate	1.5×10^{-6}
Learning Rate Schedule	Linear warmup (5 steps) with cosine decay
Global Batch Size	96 prompt sequences per step
Rollout Group Size (n)	8 samples per prompt
Maximum Sequence Length	24,576 tokens
Maximum Model Context Length	32,768 tokens
Generation Temperature	0.6 (Math), 1.0 (Code), 1.0 (IF)
Checkpointing Frequency	Every 5 steps
Weight Decay	0.01

C.6 Computational Cost and Estimation Robustness of M_d

Estimating the second-moment matrices $M_d = \mathbb{E}_{\mathcal{D}_d, \pi_{W_{\phi_d}}}[\mathbf{r}\mathbf{r}^\top] \in \mathbb{R}^{K \times K}$ requires calculating implicit rewards $\mathbf{r}(x_i, y_i) = [\log \pi_{W_{\phi_k}}(y_i \mid x_i) - \log \pi_{W_{\text{ref}}}(y_i \mid x_i)]_{k=1}^K$ across N calibration prompt rollouts ($N = 6144$ total in our experiments, 2048 per domain).

Wall-Clock Time, Scaling, and Comparison to Full Fisher Merging.

1. **Tractability over Full-Space Fisher Inversion:** Full-space Fisher merging (Matena & Raffel, 2022) requires inverting a $P \times P$ matrix ($P \approx 3.0 \times 10^9$ for SmolLM3-3B), demanding $\sim 10^{19}$ floating-point elements and exabytes of RAM, which is completely intractable for modern LLMs. In contrast, Subspace Fisher Merge compresses the system into $K \times K$ (3×3 in our setting), reducing matrix assembly and inversion to < 10 ms.
2. **Inference Rollouts and Log-Prob Scoring:** For $K = 3$ specialists and a 3B model, scoring $N = 6144$ sequences requires $O(NK)$ forward passes without gradient computation or backward graph storage, taking < 15 minutes on 8 GPUs (< 2 GPU·h).
3. **Scalability across Domain Counts K:** Contemporary MOPD pipelines consistently consolidate $K = 2 \sim 4$ specialist modalities (e.g., Math, Code, IF; Gao et al., 2026; Yang et al., 2026b;

Sun et al., 2026; Chen et al., 2026). Even if scaled to $K = 20$, solving the K-dimensional Minimax QP takes < 0.02 seconds on a single CPU core. By random matrix theory, covariance estimation requires $N = \mathcal{O}(K \log K)$ samples, meaning $N \approx 1024 \sim 2048$ samples remain sufficient for high-fidelity spectrum rank preservation.

Sample Complexity and Robustness to Estimation Noise. To evaluate how finite-sample variance in M_d propagates to the optimal weights $\boldsymbol{\alpha}^\star$ and the resulting normalized spectrum $\overline{\boldsymbol{\mathcal{L}}}$, we perform bootstrap resampling of the calibration pool $\mathcal{D}_{\text{calib}}$ across sample sizes $N \in \{512, 1024, 2048, 4096, 6144\}$:

- At $N \geq 1024$, the bootstrap standard deviation of the computed weights is $\sigma(\alpha_d^\star) < 0.012$, and the normalized spectrum remains invariant in its rank ordering: $\overline{\mathcal{L}}_{\text{math}} \gg \overline{\mathcal{L}}_{\text{code}} \gg \overline{\mathcal{L}}_{\text{if}}$ with $> 99.9\%$ confidence.
- The condition number of empirical matrices satisfies $\kappa(M_d) \leq 45.2$, well below the threshold where numerical inversion issues occur. Adding Tikhonov regularization $\lambda = 10^{-4}$ provides uniform numerical stability even under rank-deficient specialist scenarios.

C.7 RGBM Phase 2 Specification (Open-MOPD)

Phase 1 $\theta_0^\star$ (step 0): AIME24 23.91, AIME25 23.80, LCBv5 24.27, LCBv6 23.91, IFEval 73.87, IFBench 25.33, hence Math 23.86, Code 24.09, IF 49.60, Macro 32.52. The sole Open-MOPD Phase 2 scheme is residual-gated RGBM Phase 2 with Math active, written back at the unlabelled grid value $\gamma^\star = 0.125$ (Eqs. 11–11): AIME24 23.80, AIME25 24.06, LCBv5 25.15, LCBv6 24.11, IFEval 75.05, IFBench 26.00, hence Math 23.93, Code 24.63, IF 50.52, Macro $\mathbf{33.03}$. On G-OPD the same unlabelled rule is used with a separate OPD branch on each domain and yields $\gamma^\star = 0.5$.

D Limitations

Err_{lin} certifies only the first-order reward superposition on calibration prompts; it does not bound the third-order remainder of Theorem 1. While evaluating $K = 2 \sim 3$ specialists across two model families (Gao et al., 2026; Yang et al., 2026b; Sun et al., 2026), scaling to $K \gg 10$ or heterogeneous bases remains future work. Furthermore, the $+0.51$ Open-MOPD margin comes from Code and IF moving under a small Math write-back rather than from an in-domain Math gain; on G-OPD both domains are active, so this transfer is not isolated there.