NO PAIN, MORE GAIN: ITERATIVE MERGING FOR EFFECTIVE MULTI-TEACHER ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-teacher on-policy distillation (MOPD) combines independently developed domain teachers into a single student by distilling their predictions on student-generated samples. We study a setting where teachers share a reference model but undergo different post-training procedures, and find that MOPD can struggle to recover some teacher capabilities. Because distillation occurs on student-generated prefixes, the student initialization can strongly affect subsequent recovery. However, initial benchmark performance is not a reliable predictor of a good MOPD initialization. For example, merge initialization can start below SFT warm-up yet finish higher after MOPD. We further find that effective merging depends on both the relative teacher contributions and the overall merge scale, with some strong configurations lying outside the simplex of convex parameter averaging. Thus, selecting a good merge initialization requires evaluating not only its immediate performance but also the learning it enables under MOPD, making one-shot coefficient search difficult. We propose **I**terative **M**erging for **MOPD** (**IM-MOPD**), which starts from a uniform merge and progressively adds task-vector increments for under-recovered domains during distillation. In a 5-domain setting, IM-MOPD achieves higher average normalized recovery than MOPD with either uniform merge initialization or SFT warm-up, showing that effective teacher contributions can be determined progressively during training.

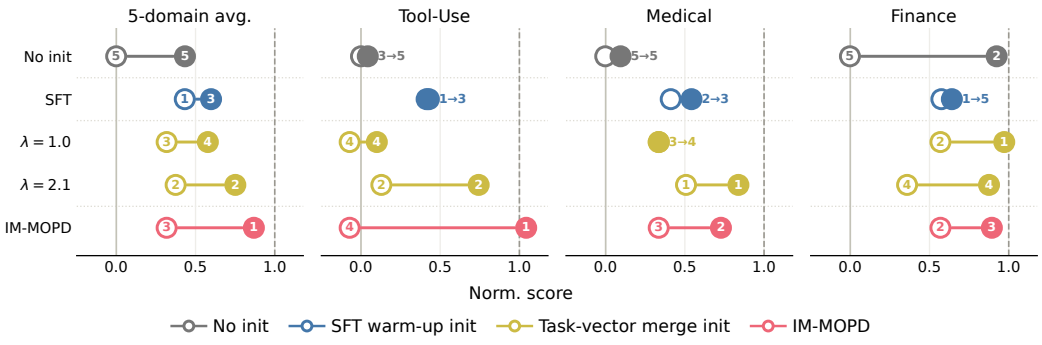


Figure 1: **Student initialization matters for MOPD, but stronger initial performance does not necessarily lead to better outcomes.** On Qwen3-4B in the 5-domain setting, MOPD initialized from the shared reference model struggles especially with the SFT-trained Tool Use and Medical teachers, while successfully recovering the RL-trained Finance teacher. Merge initialization can start below SFT warm-up yet finish higher after MOPD. Here, λ denotes the global merge scale, defined as the sum of the merge coefficients. IM-MOPD further improves performance in a single run, without an additional training stage or a separate coefficient search. Open and filled circles denote initial and final normalized scores, respectively (reference: 0; teacher: 1). Experimental details are provided in Appendix B.1.

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.

1 INTRODUCTION

Building a unified language model requires integrating capabilities acquired through different data and training procedures. Different domains often benefit from different post-training paradigms and recipes, including supervised fine-tuning (SFT), reinforcement learning (RL), or their combinations, with domain-specific choices of data and optimization (Zhang et al., 2024; Grattafiori et al., 2024; Chu et al., 2025; Blakeman et al., 2025). These experts may be developed across different timelines, motivating a modular workflow of independent specialization and subsequent integration. Multi-teacher on-policy distillation (MOPD) supports this workflow by transferring the capabilities of domain teachers into a single student. Each teacher provides token-level supervision at the prefixes generated by the student. Recent work demonstrates the practical value of this approach for combining independently developed capabilities (Ma et al., 2026; Blakeman et al., 2026).

However, it remains unclear whether MOPD can effectively integrate teachers with different post-training histories. Recent work suggests that MOPD can struggle to transfer teacher capabilities when the teacher and student have substantially different SFT histories even though they share the same reference model (Li et al., 2026; Blakeman et al., 2026; Zhu et al., 2026). This is particularly important for MOPD, where teacher supervision is provided at student-generated prefixes. Because MOPD distills on student-generated prefixes, the initialization determines the states on which distillation begins and can therefore affect subsequent learning. One way to mitigate this issue is to apply a light SFT warm-up before MOPD (Blakeman et al., 2025). Such warm-up can bring the student closer to the teachers. However, it still requires designing mixed-domain SFT, including the data mixture and training schedule. This motivates a way to prepare the student for MOPD while preserving the benefits of independent teacher development.

We show that merge initialization provides such an alternative. Following the notation of task arithmetic (Ilharco et al., 2023), we represent the weight difference between each teacher and the shared reference model as a task vector and initialize the student by adding a weighted combination of these vectors to the reference model. SFT warm-up incurs additional training cost and requires balancing multiple domain objectives again, weakening the benefit of developing multi-teachers independently. Merge initialization reduces this preparation to selecting and combining teacher task vectors, improving subsequent MOPD without an additional stage of mixed-domain training. Interestingly, initial benchmark performance does not reliably identify a good MOPD initialization. In Figure 1, SFT warm-up starts above merge initialization but finishes below it after MOPD. We further find that effective merging depends on both domain-wise ratios and the global merge scale $\lambda = \sum_i \alpha_i$. In some settings, $\lambda > 1$ improves post-MOPD recovery, placing effective initializations outside the simplex of convex parameter averaging. Selecting a good merge initialization therefore requires considering both domain-wise ratios and the global scale, while its quality cannot be reliably assessed from initial performance alone.

Rather than directly searching over continuous merge coefficients, we adapt teacher contributions during training. Starting from a uniform merge initialization, we alternate MOPD with small task-vector corrections for under-recovered domains. This replaces a domain-wise continuous coefficient search with repeated merge-or-not decisions using a shared correction size, allowing both the relative domain weights and total task-vector contribution to evolve with the student. Teacher-continuation analysis (Section 4.3) shows that teachers perform better with prefixes generated by the merged student than with those from the SFT warm-up student, even when the merged student itself performs similarly or worse. This suggests that the merged student can benefit more from teacher supervision during MOPD. Empirically, our method achieves normalized recovery scores of 86.9 on 4B and 76.0 on 1.7B, compared with 59.0 and 46.5 for SFT warm-up, respectively.

Our contributions are as follows:

- **Initialization quality beyond initial performance.** We show that higher initial performance does not necessarily lead to higher post-MOPD performance, and provide additional prefix-level evidence that benchmark scores do not fully characterize a good MOPD initialization.(Figure 1, Figure 5).

- **Merge initialization without additional SFT.** We show that merging teacher task vectors improves capability recovery after MOPD over base initialization and can outperform the

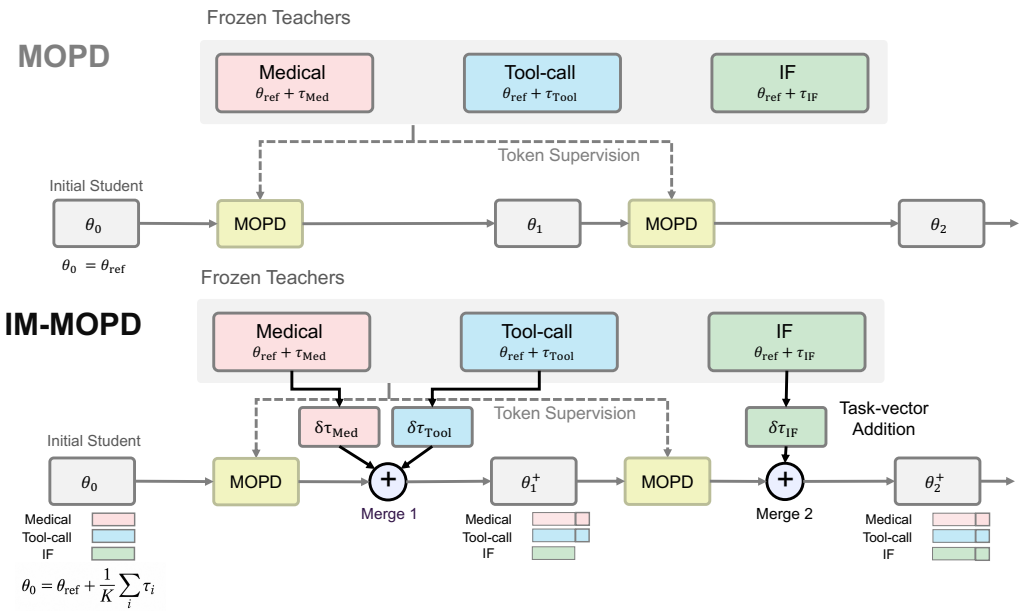


Figure 2: **Overview of IM-MOPD.** From a general SFT model θ_{ref} and K domain teachers, IM-MOPD starts from a uniform merge of task vectors $\tau_i = \phi_i - \theta_{\text{ref}}$ and alternates MOPD with selective merging: domains below the recovery threshold γ receive a fixed increment $\delta \sum_{i \in \mathcal{M}_t} \tau_i$ before the next round of distillation as a direct weight-space correction for an enhanced initialization. For specific details, see Algorithm 1.

evaluated SFT warm-up baseline. This initialization requires neither an additional SFT stage nor access to the experts' training examples (Figure 1, Table 1).

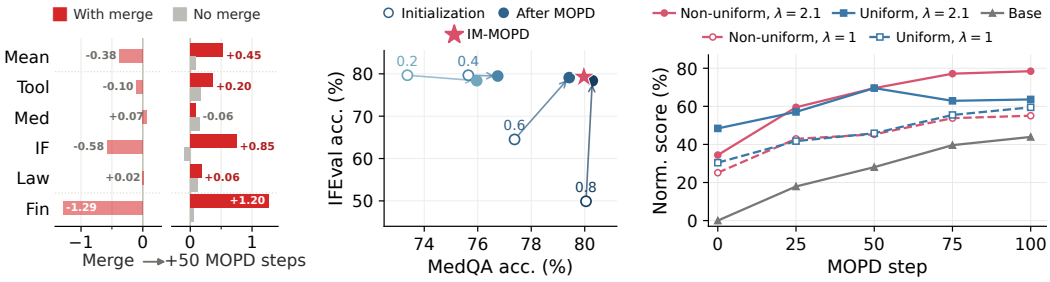
- **Iterative merging during MOPD.** We propose **IM-MOPD**, which periodically adds fixed task-vector increments for domains whose recovery remains below a threshold. In a 5-domain setting, IM-MOPD achieves higher average normalized recovery than MOPD baselines using SFT warm-up or uniform merge initialization (Table 1, Figure 4).

2 Preliminaries

We consider a same-origin multi-teacher setting with K domain teacher policies $\{\pi_{\phi_i}\}_{i=1}^K$, parameterized by $\{\phi_i\}_{i=1}^K$, and their corresponding prompt distributions $\{\mathcal{D}_i\}_{i=1}^K$. The student policy π_θ, parameterized by θ, and all domain teachers share the same architecture and originate from a common reference checkpoint θ_{ref}, after which the teachers are subsequently specialized through different domain-specific post-training procedures. Our goal is to integrate the capabilities of these independently specialized teachers into a single student policy π_θ.

Multi-Teacher On-Policy Distillation (MOPD). On-policy distillation trains a student on its own generated responses, with a teacher providing supervision at the prefixes visited by the student (Gu et al., 2024; Agarwal et al., 2024). MOPD extends this procedure to multiple domain teachers by routing each prompt to its corresponding teacher (Ma et al., 2026). For a prompt $x \sim \mathcal{D}_i$, the student generates a response $y = (y_1, \ldots, y_T) \sim \pi_\theta(\cdot \mid x)$. At token position t, the student-generated prefix is $s_t = (x, y_{<t})$. MOPD minimizes the token-averaged reverse KL between the student and the corresponding teacher over these student-generated trajectories:

$$\mathcal{L}_{\text{MOPD}}(\theta) = \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_{\substack{x \sim \mathcal{D}_i \\ y \sim \pi_\theta(\cdot|x)}} \left[\frac{1}{T}\sum_{t=1}^{T} D_{\text{KL}}(\pi_\theta(\cdot \mid s_t) \,\|\, \pi_{\phi_i}(\cdot \mid s_t)) \right]. \tag{1}$$



(a) 5-domain merge effect. (b) 2-domain grid search. (c) Merge ratios and scale.

Figure 3: **Selecting a fixed merge initialization requires accounting for subsequent learning, ratios, and scale. (a)** From the same 5-domain MOPD checkpoint, we continue training with or without a merge intervention. Although merging can reduce performance immediately, the intervened branch achieves higher recovery after 50 additional MOPD updates. **(b)** 2-domain Medical-IF coefficient grid vs. IM-MOPD. Open/filled circles denote performance before/after MOPD for fixed merge initializations; numbers indicate the Medical coefficient, with the remainder assigned to IF, and the star denotes IM-MOPD. **(c)** 5-domain MOPD under different merge initialization. Both ratios and scale affect post-MOPD recovery.

In practice, we optimize this objective using the policy-gradient formulation of on-policy distillation for computational efficiency; implementation details are provided in Appendix A.2. Because supervision is conditioned on student-generated prefixes, initialization shapes the prefixes encountered during distillation and thus the subsequent MOPD trajectory.

Model Merging. Model merging combines compatible model checkpoints directly in parameter space without additional training (Izmailov et al., 2018; Wortsman et al., 2022). In our setting, all teachers share the reference checkpoint θ_{ref}, allowing each teacher to be represented by a task vector $\tau_i = \phi_i - \theta_{\text{ref}}$ (Ilharco et al., 2023). We parameterize a merged model as $\theta_{\text{merge}}(\alpha) = \theta_{\text{ref}} + \sum_{i=1}^{K} \alpha_i \tau_i$, where $\alpha_i \geq 0$ controls the contribution of teacher i. We define the global merge scale as $\lambda = \sum_i \alpha_i$. Standard weighted parameter averaging corresponds to $\lambda = 1$, with uniform averaging given by $\alpha_i = 1/K$. We do not otherwise constrain λ to one, allowing configurations with larger total teacher contribution. A merged model can be used as the initial MOPD student, $\theta_0 = \theta_{\text{merge}}(\alpha)$.

3 Method

Overview. We study how student initialization influences the integration of domain teachers with heterogeneous post-training histories under MOPD. All teachers originate from a shared reference model θ_{ref} but are independently specialized through SFT or RLVR. Details of teacher construction and the experimental setup are provided in Appendix A.1 and A.2. Following Ma et al. (2026), let $s_i(\theta)$ denote the evaluation performance of model θ on domain i. We define the normalized domain score as $\tilde{s}_i(\theta) = \frac{s_i(\theta) - s_i(\theta_{\text{ref}})}{s_i(\phi_i) - s_i(\theta_{\text{ref}})}$, such that the shared reference model has score 0 and the corresponding domain expert has score 1. We also report the average normalized score $\tilde{s}(\theta) = \frac{1}{K} \sum_{i=1}^{K} \tilde{s}_i(\theta)$.

3.1 Observations

Merge initialization provides an effective starting point for MOPD. As shown in Figure 1, MOPD from the shared reference model struggles to recover the capabilities of SFT-trained teachers such as Medical and Tool-Use, while approaching teacher-level performance on Finance. This uneven recovery suggests that the choice of student initialization can affect how effectively different teacher capabilities are transferred. SFT warm-up partially alleviates this issue, but introduces an

additional training cost and requires separate choices of data mixture and training schedule. We find that initializing the student by merging teacher task vectors substantially improves subsequent MOPD recovery without any additional training. This suggests that model merging can serve as a simple and low-cost way to prepare students for MOPD.

A stronger student is not necessarily a better MOPD initialization. In Figure 1, SFT warm-up yields higher initial performance than the scaled merge initialization ($\lambda = 2.1$) on Tool-Use, yet the ordering reverses after MOPD. IM-MOPD further improves recovery despite not starting from the strongest benchmark performance. A similar pattern appears for merge interventions during training. In Figure 3a, we compare continued MOPD with and without intermediate merge interventions from the same checkpoint. Although a merge intervention can temporarily reduce normalized performance, the intervened branch achieves higher performance after 50 additional MOPD updates. These results indicate that immediate benchmark performance alone is insufficient for selecting states for subsequent MOPD.

Selecting good merge coefficients becomes increasingly difficult as the number of domains grows. Even with two domains, the coefficient sweep in Figure 3b yields different post-MOPD trade-offs, and the best initialization by average normalized score is not the best after MOPD. Evaluating a candidate therefore requires considering the learning it enables. With 5-domains, Figure 3c shows that both merge ratios and scale matter: non-uniform weighting improves recovery at the larger scale, while normalizing the coefficients to sum to one weakens the result. Increasing the uniform scale helps but does not match the stronger non-uniform configuration. Good coefficients thus require jointly choosing domain-wise ratios and total contribution, with each candidate assessed through MOPD. As the teacher set grows, this makes a fixed coefficient search increasingly costly, motivating progressive merge-or-not decisions during training.

3.2 Iterative Merging for MOPD (**IM-MOPD**)

Motivated by the observations above, we replace one-shot merge coefficient search with an iterative procedure that interleaves MOPD updates with small, fixed-scale additions of selected teacher task vectors (Figure 2). IM-MOPD initializes the student with a uniform merge, $\theta_0 = \theta_{\text{ref}} + \frac{1}{K}\sum_{i=1}^{K} \tau_i$, and alternates distillation with selective task-vector corrections for domains whose capabilities remain insufficiently recovered. This allows the effective merge weights to adapt throughout training without determining the full coefficient set in advance.

Algorithm 1 IM-MOPD

Require: θ_{ref} reference model; $\{\phi_i\}_{i=1}^{K}$ teachers; $\tau_i = \phi_i - \theta_{\text{ref}}$ task vectors; n MOPD iterations; H correction interval; δ fixed correction scale; γ threshold
$\theta \leftarrow \theta_{\text{ref}} + \frac{1}{K}\sum_{i=1}^{K} \tau_i$
for $t = 1, \ldots, n$ **do**
  $\theta \leftarrow \text{MOPD}(\theta; \{\phi_i\}_{i=1}^{K})$
  if $t \bmod H = 0$ **and** $t < n$ **then**
    $\mathcal{M} \leftarrow \{i \mid \tilde{r}_i(\theta) \le \gamma\}$
    $\theta \leftarrow \theta + \delta \sum_{i \in \mathcal{M}} \tau_i$
  end if
end for
return θ

At intervals of H MOPD updates, we assess domain recovery on a validation set. Let $\tilde{r}_i(\theta)$ denote the normalized score evaluated on validation data for domain i, with the reference model and corresponding teacher anchoring zero and one. We select domains whose recovery is at or below a threshold γ, forming $\mathcal{M} = \{i \mid \tilde{r}_i(\theta) \le \gamma\}$, and update the student as $\theta \leftarrow \theta + \delta \sum_{i \in \mathcal{M}} \tau_i$. Here, $\delta > 0$ is a fixed correction scale shared across domains and intervention points. Algorithm 1 summarizes the procedure.

MOPD resumes from the corrected student, and subsequent merge decisions reflect the recovery achieved through further learning. A domain may receive additional corrections if its recovery remains below the threshold. These task-vector additions allow both the relative merge ratios and the total merge scale to evolve during training, without constraining the cumulative coefficients to sum to one. By interleaving these corrections with distillation, IM-MOPD progressively shapes the student from which further teacher-guided learning proceeds.

Table 1: **MOPD results across 5-domains at Qwen3-4B and Qwen3-1.7B.** IM-MOPD achieves the highest average normalized score among the compared student initialization methods at both model sizes, without an additional SFT stage. Domain Teacher reports the performance of the domain specialist for each domain.

Method	MedQA	CaseHOLD	FinQA	IFBench	τ^2	Norm.
Qwen3-4B-OT3	69.5	61.8	59.1	24.9	5.3	0.0
Domain Teacher	80.7	73.4	75.0	59.4	66.7	100.0
MOPD	69.1	64.7	74.3	56.9	7.9	42.8
SFT Warm-up + MOPD	73.9	68.2	71.1	53.4	31.6	59.0
Uniform Merge + MOPD	72.7	67.6	74.7	53.6	11.4	53.8
IM-MOPD	76.6	70.9	74.0	57.8	69.3	**86.9**
Qwen3-1.7B-OT3	44.1	44.6	42.4	18.4	7.9	0.0
Domain Teacher	58.5	69.6	62.4	46.4	28.1	100.0
MOPD	49.0	52.2	59.1	36.8	0.0	34.9
SFT Warm-up + MOPD	52.7	56.6	54.8	32.6	10.5	46.5
Uniform Merge + MOPD	51.2	59.9	58.6	34.6	4.4	46.3
IM-MOPD	53.7	59.4	55.9	42.9	28.1	**76.0**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Reference model and teachers. We first fine-tune Qwen3-4B-Base and Qwen3-1.7B-Base (Yang et al., 2025) on OpenThoughts3 (Guha et al., 2026) to obtain a common reference checkpoint for the student and domain teachers at each model size. We then train 5 domain teachers separately from this reference: Medical, Law, and Tool Use through SFT, and Finance and Instruction Following (IF) through reinforcement learning with verifiable rewards (RLVR). For detailed training and data configurations, please refer to Appendix A.1.

Baselines. We compare our method with three representative initialization baselines. *MOPD* starts from the common reference model (Ma et al., 2026). *Uniform Merge* starts from the equal-weight average of the five teachers and serves as the static merge initialization baseline, isolating the effect of iterative adaptation from the choice of a more sophisticated one-shot merging rule. *SFT Warm-up* first trains the student on mixed teacher-generated samples before MOPD. More details on the baselines are provided in Appendix A.2–A.4.

Iterative-merging configurations. Both model sizes start from uniform merging and use the same validation-based recovery rule. We set the recovery threshold to $\gamma = 0.6$ for 4B and $\gamma = 0.4$ for 1.7B, the fixed correction scale to $\delta = 0.3$ for both, and the correction interval to $H = 25$. Every H MOPD updates, we add the corresponding teacher task vector for domains whose recovery is at or below γ. Appendix A.4 provides the validation protocol and correction schedules.

Evaluations. We evaluate Medical on MedQA (Jin et al., 2021), Law on CaseHOLD (Zheng et al., 2021), Finance on FinQA (Chen et al., 2021), Instruction Following on IFBench (Pyatkin et al., 2025), and Tool Use on τ^2-Telecom (Barres et al., 2025). We report answer accuracy for MedQA and CaseHOLD, numeric-answer accuracy for FinQA, prompt-level loose accuracy for IFBench, and task success for telecom tasks. For the main results, we report pass@1 for Tool Use and avg@3 for the other benchmarks, together with the average normalized score across the five domains. Evaluation protocols are provided in Appendix A.5.

4.2 MAIN RESULTS

Merge initialization improves recovery, but leaves uneven domain transfer. Table 1 shows that starting MOPD from a uniform merge substantially improves aggregate recovery over starting from

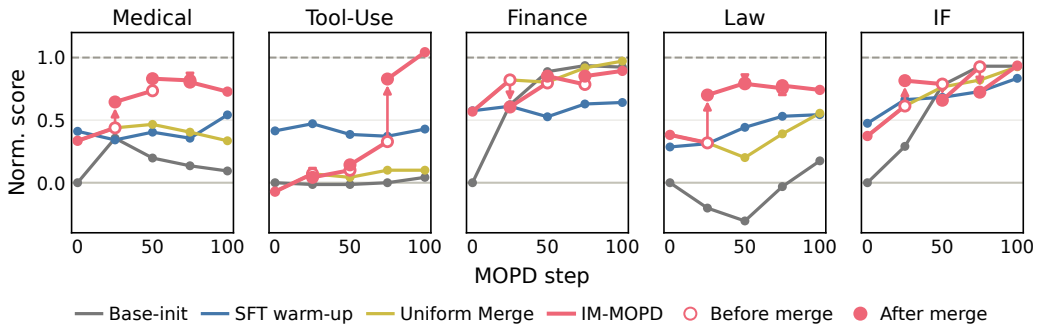


Figure 4: **Iterative merging improves recovery in domains left behind by fixed initialization.** On Qwen3-4B, IM-MOPD starts from Uniform Merge and progressively improves under-recovered domains during MOPD. The largest gains are observed in Medical, Tool Use, and Law, while performance in already well-recovered domains remains broadly stable.

the reference model, without requiring an additional SFT stage. SFT warm-up provides a further improvement in aggregate recovery, but requires an additional training stage. Despite these gains, both approaches leave some domains substantially under-recovered, most notably Tool Use. Thus, a better initialization improves MOPD overall, but does not by itself ensure balanced capability integration across domains.

Iterative merging improves aggregate capability integration. For Qwen3-4B, iterative merging increases the average normalized recovery from 53.8% with uniform initialization to 86.9%. For Qwen3-1.7B, IM-MOPD reaches 76.0%, compared with 46.5% for SFT warm-up, and recovers Tool Use to the teacher level. Across both model sizes, IM-MOPD achieves higher scores than SFT warm-up on all five benchmarks. Relative to uniform initialization, the largest improvements are observed in Medical and Tool Use, while performance on Finance and Instruction Following remains broadly comparable. These results show that iterative corrections can improve recovery in domains that remain under-recovered after initialization without substantially degrading domains already transferred effectively by MOPD.

4.3 Ablations and Analysis

Table 2: **Merge-timing ablation on Qwen3-4B.** We take the final cumulative merge coefficients produced by IM-MOPD and apply them either entirely at initialization or entirely after MOPD. Both controls underperform the original IM-MOPD schedule, showing that when the same final merge coefficients are applied matters.

Merge timing	MedQA	CaseHOLD	FinQA	IFBench	τ^2	Norm.
No merge (Base MOPD)	69.1	64.7	74.3	56.9	7.9	42.8
All at initialization	76.3	70.8	74.0	52.4	57.0	79.0
All after MOPD	68.1	71.0	59.6	38.2	10.5	23.3
Iterative (IM-MOPD)	76.6	70.9	74.0	57.8	69.3	86.9

Iterative merging versus merging only at initialization. We compare IM-MOPD with an initialization-only variant that merges teacher task vectors using the same cumulative coefficients, followed by the same number of MOPD updates (Table 2). Iterative merging reaches 86.9% normalized recovery, versus 79.0% for the initialization-only variant, with the largest gain in Tool Use. With the cumulative coefficients held fixed, distributing task-vector additions across distillation is more effective than applying them all at the start.

Merging and MOPD reinforce each other. After MOPD from the reference model with no additional initialization, we apply the final cumulative merge coefficients obtained by IM-MOPD to

the trained student. As shown in Table 2, this late merging achieves only 23.3% normalized recovery, below Base MOPD and far below IM-MOPD. Together with the initialization-only control, this result shows that the gains of IM-MOPD cannot be explained simply by the final task-vector displacement induced by merging. The same merge coefficients are substantially more effective when interleaved with MOPD, indicating that subsequent teacher-guided updates after each correction are important for the final recovery.

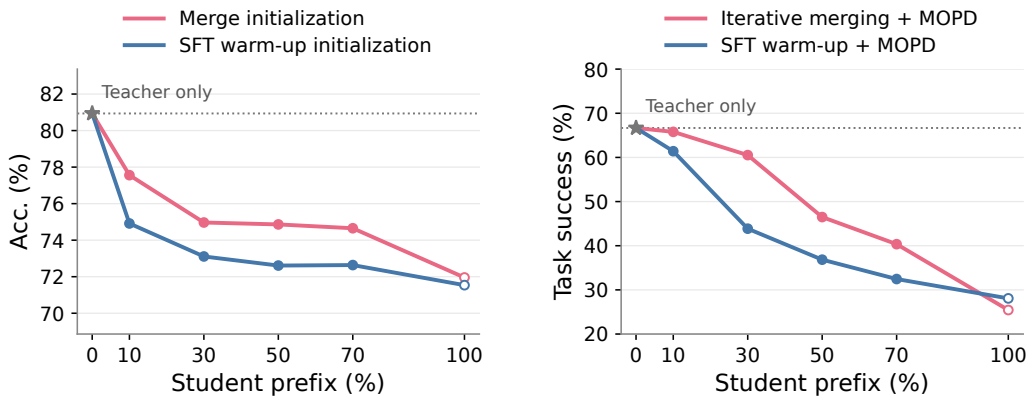
Hyperparameter sensitivity. We evaluate sensitivity to the recovery threshold $\gamma \in \{0.4, 0.6\}$ and correction size $\delta \in \{0.1, 0.2, 0.3, 0.4\}$ in Table 3. Qwen3-4B maintains over 80% normalized recovery across tested correction sizes, while Qwen3-1.7B performs best with $\gamma = 0.4$ and $\delta = 0.3$. Effective settings can thus be identified from a small candidate set. More detailed results are provided in Appendix B.2.

Table 3: **IM-MOPD sensitivity to γ and δ.** Average normalized recovery (%) across 5 domains.

Model	γ	$\delta = .2$	$\delta = .3$	$\delta = .4$
Qwen3-4B	0.6	80.3	86.9	86.1

Model	γ	$\delta = .1$	$\delta = .2$	$\delta = .3$
Qwen3-1.7B	0.4	62.2	57.2	76.0
	0.6	57.7	65.0	69.8

Teacher continuation from student-generated prefixes. Prior work evaluates the reliability of teacher supervision in OPD by continuing from student-generated prefixes (Li et al., 2026). We adopt this diagnostic to assess teacher–student compatibility for merged and SFT warm-up students. Each student generates an initial prefix, after which the corresponding domain teacher completes the response. We vary the fraction generated by the student and evaluate task performance (Figure 5). Higher continuation performance indicates greater compatibility between the student-generated prefix and the teacher's subsequent generation. Protocol details are provided in Appendix B.3.



(a) Medical (MedQA).

(b) Tool Use (τ^2-Telecom).

Figure 5: **Teacher continuation is stronger from merge-based student prefixes on Qwen3-4B.** In Medical, teacher continuation is consistently stronger from merge-initialized prefixes than from SFT-warm-up prefixes despite similar student-only performance. In Tool Use, IM-MOPD yields higher continuation success from short prefixes despite lower student-only performance at the evaluated checkpoint. The 0% and 100% endpoints denote teacher-only and student-only performance, respectively.

Across Medical and Tool Use, teacher continuation performance drops even with short prefixes from the SFT warm-up student, whereas prefixes from the merged student better preserve it over the early portion of the trajectory (Figure 5). These results are consistent with prior findings that teacher–student compatibility matters for effective OPD (Li et al., 2026). This advantage persists even when the merged student's standalone performance is similar to or lower than that of the SFT warm-up student. Together, these results suggest that merging helps the student generate prefixes from which teachers can more successfully continue, allowing the student to benefit more from teacher supervision during MOPD.

5 Related Works

5.1 On-Policy Distillation for Large Language Models (LLMs)

Knowledge distillation (KD) traditionally transfers teacher knowledge by matching predictive distributions on a fixed training distribution (Hinton et al., 2015). Kim & Rush, 2016 extended the paradigm to apply sequence-level KD from teacher-generated sequences. However, such offline supervision exposes the student primarily to teacher- or data-induced prefixes, creating a mismatch between train and inference setups. On-policy distillation (OPD) addresses this mismatch by instead querying the teacher on student-generated trajectories (Agarwal et al., 2024), and is commonly paired with reverse-KL objectives that naturally optimize the student distribution over its own visited states (Gu et al., 2024). This formulation has become increasingly relevant for reasoning LLMs, where the context has scaled with the reasoning length (Xu et al., 2026; Blakeman et al., 2026; Ma et al., 2026; Team et al., 2026). Yet, the benefit of on-policy supervision depends on the teacher remaining informative on states visited by the student: as teacher–student distributions diverge, teacher feedback can become less effective or unreliable (Xu et al., 2025; Li et al., 2026; Blakeman et al., 2026). Recent methods therefore selectively constrain or adapt distillation according to teacher–student alignment (Xing et al., 2026), highlighting distributional compatibility as a central challenge in effective OPD.

5.2 Consolidating Domain Specialists

As post-training increasingly targets heterogeneous capabilities often through reinforcement learning (RL), consolidating these divergent capabilities into a single model has become a central challenge. Existing approaches differ in how capabilities are acquired and integrated. Mixed RL jointly optimizes a shared policy over a mixture of domains, enabling direct capability integration but coupling domains with potentially heterogeneous data, rewards, rollout lengths, and optimization dynamics (Blakeman et al., 2025). Cascade RL instead applies domain-specific RL stages sequentially, allowing each stage to use domain-specific environments and optimization configurations (Wang et al., 2025). While carefully designed cascades can largely preserve previously acquired capabilities, capability drift can accumulate across increasingly specialized stages, motivating intermediate consolidation or stabilization (Yang et al., 2026). Alternatively, domain specialists can be trained independently and subsequently consolidated through parameter merging (Team et al., 2025) or multi-teacher on-policy distillation (MOPD) (Ma et al., 2026). Our work builds on this specialist-consolidation paradigm and investigates whether iterative merging can progressively integrate heterogeneous capabilities while maintaining effective teacher–student alignment.

6 Conclusion

We study MOPD when domain teachers share a common reference model but undergo different post-training procedures. We find that initial benchmark performance is an unreliable criterion for selecting MOPD initializations, and that effective merge initialization depends on both relative teacher weights and the overall merge scale. Motivated by these observations, we introduce IM-MOPD, which starts from a uniform merge and progressively adds task-vector corrections for domains that remain under-recovered during distillation. Across five domains at both 4B and 1.7B scales, IM-MOPD achieves higher aggregate capability recovery than uniform merging, and SFT warm-up, without requiring an additional SFT stage. Merge-timing ablations further show that applying the same final cumulative merge coefficients entirely before or after MOPD is less effective than interleaving the corresponding corrections with distillation. Together, these results suggest that teacher contributions are more effectively determined progressively during MOPD than fixed entirely before training.

Limitations. IM-MOPD assumes that the student and teachers share an architecture and originate from a common reference checkpoint. Its task-vector updates therefore cannot be applied directly when teacher and student model sizes differ. Our experiments cover five domains within the Qwen3 family at 1.7B and 4B; broader applicability to larger models and more diverse teacher sets remains to be evaluated. The method also requires periodic validation to identify under-recovered domains, introducing evaluation overhead.

AI USE STATEMENT

We used generative AI assistants to assist with code implementation and data-processing scripts, as well as language polishing, translation, and editing of the manuscript. All research decisions, experimental design, analyses, interpretation of results, and scientific claims were made by the authors. Separately, LLMs were used as part of our experimental pipeline to generate training samples for domain teachers and SFT warm-up and to simulate users in the Tool Use evaluation (Appendices A.1, A.3, and A.5).

ETHICS STATEMENT

This study uses existing benchmark data and simulated tool-use environments. Distillation and merging can transfer factual errors, biases, and unsafe behavior from teacher models. Our medical, legal, and financial evaluations measure benchmark performance and do not establish readiness for professional decision-making. Applying these models in such settings or to real-world tools requires separate safety assessment and appropriate human oversight.

REPRODUCIBILITY STATEMENT

Algorithm 1 specifies IM-MOPD. Appendix A.1 documents teacher training and data sources, while Appendices A.2–A.3 describe MOPD and SFT warm-up settings. Validation data and merge schedules are given in Appendix A.4, and evaluation protocols in Appendix A.5. Appendices B.2–B.4 provide additional ablation results, the student-prefix diagnostic procedure, and two-domain results.

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 *International Conference on Learning Representations*, pp. 21246–21263, 2024.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. *arXiv preprint arXiv:2506.07982*, 2025.

Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, et al. Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. *arXiv preprint arXiv:2512.20848*, 2025.

Aaron Blakeman, Aaron Thomas, Aastha Jhunjhunwala, Abhibha Gupta, Abhinav Khattar, Adam Rajfer, Adi Renduchintala, Adil Asif, Aditya Vavre, Adriana Flores Miranda, et al. Nemotron 3 ultra: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. *arXiv preprint arXiv:2606.15007*, 2026.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, 2021.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. In *International Conference on Learning Representations*, volume 2026, pp. 108059–108130, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14):6421, 2021.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing Verifiable Instruction Following. In *Conference on Neural Information Processing Systems*, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Meituan LongCat Team, Anchun Gui, Bei Li, Bingyang Tao, Bole Zhou, Borun Chen, Chao Zhang, Chengcheng Han, Chenhui Yang, Chi Zhang, et al. Introducing longcat-flash-thinking: A technical report. *arXiv preprint arXiv:2509.18883*, 2025.

Boxin Wang, Chankyu Lee, Nayeon Lee, Sheng-Chieh Lin, Wenliang Dai, Yang Chen, Yangyi Chen, Zhuolin Yang, Zihan Liu, Mohammad Shoeybi, et al. Nemotron-cascade: Scaling cascaded reinforcement learning for general-purpose reasoning models. *arXiv preprint arXiv:2512.13607*, 2025.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998, 2022.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust Region On-Policy Distillation. *arXiv preprint arXiv:2606.01249*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pp. 64616–64646, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, et al. Nemotron-cascade 2: Post-training llms with cascade rl and multi-domain on-policy distillation. *arXiv preprint arXiv:2603.19220*, 2026.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. In *International Conference on Learning Representations*, pp. 44694–44713, 2024.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *International Conference on Artificial Intelligence and Law*, pp. 159–168, 2021.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026.

A Experimental Details

Training was performed on NVIDIA A100 80GB GPUs.

A.1 Teacher Training

We first fine-tune Qwen3-4B-Base[1] and Qwen3-1.7B-Base[2] on OpenThoughts3[3] to obtain a common reference at each model size. Domain teachers are initialized from the corresponding reference. Tables 4 and 5 summarize their training configurations. For GRPO, batch size is given as prompts $\times$ completions.

Table 4: Training configurations of the 4B domain teachers.

Teacher	Training	Batch	Length	Peak LR	Tokens (M)
Medical	SFT	16	8,192	10^{-5}	426
Law	SFT	128	20,480	10^{-5}	269
Tool Use	SFT	64	32,768	$5 \cdot 10^{-6}$	796
Finance	GRPO	32×8	16,384	$3 \cdot 10^{-6}$	315
IF	GRPO	64×8	32,768	$3 \cdot 10^{-6}$	483

Table 5: Training configurations of the 1.7B domain teachers.

Teacher	Training	Batch	Length	Peak LR	Tokens (M)
Medical	SFT	16	8,192	10^{-5}	426
Law	SFT	64	20,480	10^{-5}	269
Tool Use	SFT	64	32,768	$5 \cdot 10^{-6}$	796
Finance	GRPO	32×8	16,384	$3 \cdot 10^{-6}$	275
IF	GRPO	64×8	32,768	$3 \cdot 10^{-6}$	424

SFT data. Medical uses 40,853 correct, deduplicated samples generated by Qwen3.6-35B-A3B from MedQA training prompts. Law uses 56,123 correct, deduplicated samples generated by the same model from CaseHOLD and bar-exam prompts. Tool Use uses 32,758 retail, airline, and telecom samples from `inclusionAI/AReaL-tau2-data`[4], after removing overlength samples and telecom task-ID overlap. Tool schemas are preserved, and supervision covers the final assistant turn with earlier thinking removed.

SFT optimization. SFT uses the Trainer implementation in Hugging Face Transformers[5]. All three domains use packing, zero weight decay, gradient-norm clipping at 1, and cosine decay with 5% warm-up. Tool Use additionally uses a 10% learning-rate floor.

RL data. Finance uses 7,565 numeric-answer prompts from the training splits of FinQA and TAT-QA.[6] IF uses 11,728 verifiable instruction-following prompts from the `rlvr1` blend of Nemotron-RL-Ultra-Training-Blends.[7]

GRPO. Finance and IF use NeMo-RL[8] GRPO with numeric-answer and instruction-constraint rewards, respectively. Both use AdamW with zero weight decay, ten learning-rate warm-up updates,

[1] https://huggingface.co/Qwen/Qwen3-4B-Base
[2] https://huggingface.co/Qwen/Qwen3-1.7B-Base
[3] https://huggingface.co/datasets/open-thoughts/OpenThoughts3-1.2M
[4] https://huggingface.co/datasets/inclusionAI/AReaL-tau2-data
[5] https://github.com/huggingface/transformers
[6] https://github.com/NExTplusplus/TAT-QA
[7] https://huggingface.co/datasets/nvidia/Nemotron-RL-Ultra-Training-Blends
[8] https://github.com/NVIDIA-NeMo/RL

and a constant rate thereafter. Training uses clipped policy updates and overlength penalties, without a reference KL penalty. Sampling uses temperature 1 and top-p 1.

A.2 MOPD Training Configuration

MOPD data. Medical uses 8,384 four- and five-option MedQA training prompts. Law uses 31,563 CaseHOLD and bar-exam prompts.[9] Finance and IF reuse the RL prompt pools described in Appendix A.1. Tool Use uses 31,718 dialogue contexts from the retail, airline, and telecom portions of AReaL-tau2-data, with each context serving as a prompt for the next assistant turn.

MOPD optimization. All student methods use 100 MOPD updates, a batch of 128 prompts, uniform domain sampling, and one completion per prompt. Policy-gradient distillation clips advantage magnitude at 5 and masks truncated completions, without task-reward or reference-KL terms. The learning rate is 10^{-6}, and completions are capped at 16,384 tokens. Sampling uses temperature 1 and top-p 1.

A.3 SFT Warm-up Baseline

The warm-up baseline trains on mixed samples generated by the five teachers at the corresponding model size until the student reaches approximately 30% normalized recovery. MOPD then starts from this warm-up checkpoint (Table 6).

Table 6: SFT warm-up configurations.

Student	Training	Batch	Length	Peak LR	Tokens (M)
4B	SFT	32	32,768	10^{-5}	167
1.7B	SFT	32	32,768	10^{-5}	171

Warm-up uses packed sequences and cosine learning-rate decay with 5% warm-up.

A.4 Main Experiment Settings

Main results use $\gamma = 0.6$ for 4B and $\gamma = 0.4$ for 1.7B, with $\delta = 0.3$ for both. Corrections are applied after updates 25, 50, and 75 according to the recovery rule in Algorithm 1. Table 7 lists the selected domains for each configuration.

Validation data. For Tool Use, we sample 128 tasks from the full dataset. For each other domain, we sample 256 prompts from a validation set held out from its training pool.

Table 7: Realized merge schedules. T/M/L/F/I denote Tool Use, Medical, Law, Finance, and IF.

Size	γ	δ	Update 25	Update 50	Update 75
4B	0.6	0.2	T,M,L,I	T,M	T
4B	0.6	0.3	T,M,L,I	T	T
4B	0.6	0.4	T,M,L,I	T	$\emptyset$
1.7B	0.4	0.1	T,M,L,I	T,M	T
1.7B	0.4	0.2	T,M,F,I	T	T
1.7B	0.4	0.3	T,M,I	T	T,I
1.7B	0.6	0.1	T,M,L,F,I	T,M,I	T,M,I
1.7B	0.6	0.2	T,M,L,F,I	T,M,I	T
1.7B	0.6	0.3	T,M,L,F,I	T,M,I	T,F,I

[9] Bar-exam data: `https://huggingface.co/datasets/reglab/barexam_qa`.

A.5 Evaluation Protocol

Evaluation uses temperature 1, top-p 1, and disabled top-k. Single-turn tasks use zero presence penalty: MedQA/CaseHOLD/FinQA have 1,273/5,314/1,147 examples and a 26,624-token completion cap; IFBench has 300 prompts and a 32,768-token cap. We use MedQA, CaseHOLD, FinQA, IFBench, and IFEval for the non-tool domains.[10]

Telecom evaluation uses the official v1.0.1 harness[11] on 114 base-split tasks, with a 32,768-token context and a 200-step limit. The user simulator is `gpt-4.1-2025-04-14` at temperature 0; empty responses and execution or context-limit errors count as failures. The split includes 20 task IDs flagged for upstream training overlap, so teacher-data filtering does not exclude upstream exposure.

B Additional Experiments

B.1 Initialization Comparison

Table 8 gives the five-domain results underlying Figure 1. The $\lambda = 1.0$ merge assigns coefficient 0.2 to each teacher; the $\lambda = 2.1$ merge uses $(0.5, 0.3, 0.2, 0.3, 0.8)$ in the order Medical, Law, Finance, IF, and Tool Use. SFT warm-up and IM-MOPD follow Appendices A.3 and A.4, respectively. All post-MOPD scores use 100 updates.

Table 8: Five-domain performance for the initializations in Figure 1. Each entry shows before / after MOPD. IF is evaluated with IFEval prompt-level strict accuracy.

Initialization	MedQA	CaseHOLD	FinQA	IFEval	τ^2
No init	69.5 / 69.1	61.8 / 64.7	59.1 / 74.3	37.5 / 74.9	5.3 / 7.9
SFT warm-up	72.6 / 73.9	65.4 / 68.2	69.1 / 71.1	56.6 / 71.0	30.7 / 31.6
Merge ($\lambda = 1.0$)	71.7 / 72.7	66.5 / 67.6	69.0 / 74.7	52.5 / 74.7	0.9 / 11.4
Merge ($\lambda = 2.1$)	73.7 / 77.5	67.9 / 68.9	66.0 / 73.5	52.7 / 66.2	13.2 / 50.9
IM-MOPD	71.7 / 76.6	66.5 / 70.9	69.0 / 74.0	52.5 / 75.0	0.9 / 69.3

B.2 Ablation Details

Tables 9–10 report the complete benchmark results underlying the main-text ablations. Both scales are analyzed using the validation recovery rule described in Appendix A.4.

Table 9: 4B correction-size sweep with $\gamma = 0.6$.

Correction	MedQA	CaseHOLD	FinQA	IFBench	τ^2	Norm.
$\delta = 0.2$	79.1	70.3	73.9	58.7	36.8	80.3
$\delta = 0.3$	76.6	70.9	74.0	57.8	69.3	86.9
$\delta = 0.4$	78.9	71.8	73.0	54.2	59.6	86.1

B.3 Student-prefix Teacher-continuation Diagnostics

We vary the student-generated fraction $\alpha \in \{0.1, 0.3, 0.5, 0.7\}$; the 100% endpoints are evaluated using the student alone.

[10] `https://huggingface.co/datasets/GBaker/MedQA-USMLE-4-options`
`https://huggingface.co/datasets/casehold/casehold`
`https://github.com/czyssrs/FinQA`
`https://huggingface.co/datasets/allenai/IFBench_test`
`https://huggingface.co/datasets/google/IFEval`

[11] `https://github.com/sierra-research/tau2-bench`

Table 10: 1.7B recovery-threshold and correction-size sweep.

(γ, δ)	MedQA	CaseHOLD	FinQA	IFBench	τ^2	Norm.
$(0.4, 0.1)$	53.7	62.9	59.0	34.8	14.0	62.2
$(0.4, 0.2)$	52.2	59.5	59.3	34.0	14.0	57.2
$(0.4, 0.3)$	53.7	59.4	55.9	42.9	28.1	76.0
$(0.6, 0.1)$	54.6	62.1	58.6	35.2	8.8	57.7
$(0.6, 0.2)$	55.0	63.9	58.2	38.4	12.3	65.0
$(0.6, 0.3)$	56.9	64.7	50.5	41.8	19.3	69.8

Medical. For each recorded MedQA response of length L, the teacher continues after the first $\lfloor \alpha L \rfloor$ student-generated tokens. The prefix length is therefore determined separately for each sample.

Tool Use. In τ^2-Telecom, we instead use a per-turn budget $N = \mathrm{round}(\alpha \bar{L})$, where $\bar{L}$ is the student's mean tokens per assistant turn in a standalone evaluation. The teacher continues when the student reaches this budget. We use the mean because changing one turn in a multi-turn simulation changes subsequent interactions, so later turns cannot be matched to their original sample-specific lengths. Because the budget uses the mean turn length, some shorter turns may be generated entirely by the student.

B.4 Two-domain Initialization and Merge Timing

Table 11 summarizes the MedQA and IFEval results. IFEval reports prompt-level strict accuracy over 541 prompts. Norm. averages normalized recovery over these two benchmarks.

Table 11: Two-domain performance before and after MOPD. Simplex coefficients are ordered as Medical and IF. The last row uses uniform initialization and mid-training merging.

	Initialization			After MOPD		
Initialization / method	MedQA	IFEval	Norm.	MedQA	IFEval	Norm.
$(0.2, 0.8)$	73.4	79.7	76.7	76.0	78.4	86.4
$(0.4, 0.6)$	75.6	79.7	86.7	76.7	79.5	91.3
$(0.6, 0.4)$	77.4	64.5	75.4	79.4	79.1	102.5
$(0.8, 0.2)$	80.0	49.9	68.9	80.3	78.4	105.4
IM-MOPD	78.3	68.8	84.8	80.0	79.3	105.2