MULTI-TEACHER ON-POLICY DISTILLATION VIA GATED META-LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-teacher on-policy distillation (MOPD) consolidates multiple domain-specialized teachers into a single deployable student through dense on-policy supervision. Yet, we find that current methods easily suffer from a teacher-strengthening paradox—stronger teachers can yield a worse merged student—due to increasingly specialized reasoning patterns and the resulting loss of student-teacher overlap across domains. In this paper, we introduce GMOPD, a gated meta-learning framework for MOPD that optimizes teacher integration based on post-adaptation performance. We formulate each base-referenced OPD task as a KL-regularized RL problem and optimize a meta-objective that learns a shared initialization capable of rapidly adapting to each teacher. We further train an adaptive rank gate that selectively activates teacher-relevant LoRA ranks while protecting capabilities required by the remaining teachers. Together, these components preserve student-teacher overlap and mitigate cross-teacher forgetting. We evaluate GMOPD on six challenging benchmarks spanning mathematical reasoning, coding, scientific reasoning, instruction following, tool use, and general reasoning. Across multiple model scales, GMOPD consistently outperforms existing merging and distillation methods, scales to 20 heterogeneous teachers, and composes complementary capabilities beyond those of individual teachers.

1 INTRODUCTION

Recent advances in large language model post-training have established on-policy distillation (OPD) as an effective paradigm for capability transfer, where a student generates its own trajectories and receives token-level supervision from a teacher on the states it actually visits (Jiang et al., 2025; Gao et al., 2026). Unlike conventional distillation on a static teacher-generated corpus, this on-policy interaction reduces the training–inference mismatch and provides dense training guidance even when task rewards are sparse (Park et al., 2019; Fan et al., 2025). Consequently, OPD has shown broad potential in strong-to-weak compression, where the capability of a large teacher is transferred into a smaller deployable model, and in self-distillation or iterative self-evolution, where knowledge exposed by a stronger context, checkpoint, or specialized variant is internalized back into the model parameters (Ouyang et al., 2022; Guo et al., 2025; Kumar et al., 2025; Zhou et al., 2025).

More recently, research has extended OPD towards multi-domain model merging, where multiple specialized policies are consolidated into a single generalist model (Ye et al., 2024; Jiang et al., 2025; Fan et al., 2025; Gao et al., 2026). In a typical *multi-teacher on-policy distillation* (MOPD) framework, a collection of teachers is first trained independently with RL on different domains. The student then samples trajectories from a multi-domain prompt mixture, and each trajectory is routed to the corresponding frozen teacher for dense token-level supervision on student-generated prefixes. Such a specialize-then-merge paradigm avoids forcing heterogeneous rewards, environments, and rollout systems into one joint RL process. It enables parallel teacher development and domain-specific training recipes, isolates failures within individual domains, and ultimately produces one deployable model without retaining the teacher ensemble.

Despite promising performance, we find that current MOPD methods typically suffer from a *teacher strengthening paradox*, that is, stronger teachers do not necessarily yield a better merged student and can even lead to worse overall performance (as illustrated in Fig. 1). We argue that this paradox arises because of the increasingly specialized reasoning patterns of stronger teachers. As a teacher improves

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.

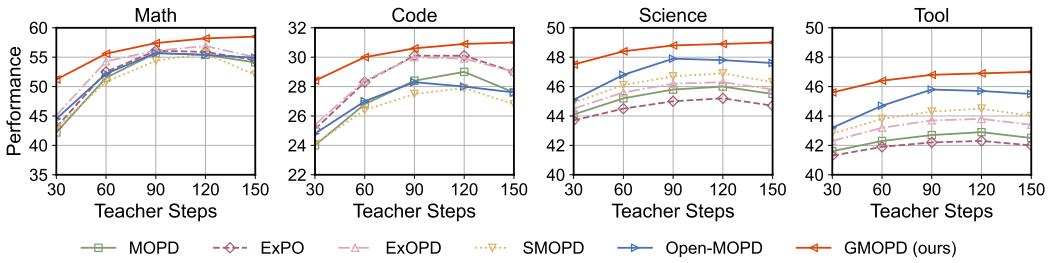


Figure 1: MOPD dynamics when checkpoints from different RL stages are used as controlled proxies for teacher capability. We independently train each domain teacher for different numbers of RL steps and perform MOPD from the same Qwen3-4B initialization. As teacher capability increases, conventional MOPD initially improves but eventually deteriorates because increasingly specialized teacher distributions amplify cross-domain interference, reducing student–teacher overlap and destabilizing reverse-KL updates.

in its own domain, RL can increasingly concentrate probability mass on successful trajectories, making its policy distribution more concentrated and its reasoning pattern more specialized (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025). Once the student is trained on one domain, its shared policy is pulled towards that reasoning pattern, altering the rollout distributions in other domains. These distributions may deviate substantially from those of the corresponding teachers and thus receive very low probability under that teacher. Since OPD evaluates reverse-KL supervision only on student-generated data, this mismatch can produce large but poorly directed updates, which can destabilize optimization and lead to policy collapse (Jiang et al., 2025; Gao et al., 2026).

This paper therefore aims to answer the question: *"How can a student learn from one specialized teacher without losing the on-policy overlap required to learn from the others?"* A natural approach is to freeze most of the shared student and confine each teacher-specific update to a lightweight low-rank subspace. However, standard LoRA cannot prevent a teacher-specific update from reducing the student's on-policy overlap with the remaining teachers, because all teachers still updates the same shared set of low-rank parameters (Hu et al., 2022). Directly lowering the LoRA rank reduces the number of adapted parameters, but simultaneously restricts the capacity needed to accommodate heterogeneous teacher knowledge (Li et al., 2024; Zhang et al., 2024). This reveals a fundamental trade-off: teacher-specific updates must be sufficiently constrained to preserve cross-teacher overlap, while the student must retain enough capacity to accommodate heterogeneous teacher knowledge.

To tackle this challenge, we introduce Gated Meta Multi-Teacher On-Policy Distillation (GMOPD), which integrates multiple specialized teachers into a shared student while mitigating cross-domain interference and preserving teacher-student distributional overlap throughout distillation. First, we introduce a meta-MOPD formulation that optimizes each teacher-specific update based on its resulting post-adaptation distillation performance, thereby enabling the student to learn a shared initialization that can rapidly adapt to every teacher. To efficiently acquire teacher-specific knowledge, we design a learnable rank gate that dynamically modulates the contribution of each LoRA rank, selectively adapting only the ranks relevant to each teacher while preserving those needed by the remaining teachers. Furthermore, we develop a two-term gate objective that jointly trains the meta-gate for rapid teacher adaptation and the shared low-rank matrices for multi-teacher knowledge preservation, yielding stable and scalable teacher integration with theoretical guarantees on teacher-student overlap and policy improvement.

We systematically evaluate the proposed method across 6 challenging domains, including mathematical reasoning, code reasoning, tool use, and general reasoning. The results demonstrate that GMOPD significantly improves the MOPD performance across all domains, consistently outperforming existing model-merging and distillation methods by 30% to 50% in terms of teacher recovery. Moreover, to the best of our knowledge, GMOPD is the first MOPD framework demonstrated to scale to as many as 20 heterogeneous teachers, highlighting its scalability to large-scale multi-domain integration. GMOPD also maintains sufficient distributional overlap between the shared student and each domain teacher throughout optimization, continuously producing teacher-supported on-policy trajectories that provide reliable dense supervision for multi-teacher distillation.

2 Related Work

Model Merging. Model merging has recently attracted significant attention, as it enables multiple specialized checkpoints to be consolidated into one model without retaining an inference-time ensemble. One line of work focuses on direct parameter interpolation (Matena & Raffel, 2022; Jin et al., 2022; Ortiz-Jimenez et al., 2023; Yang et al., 2024a). Wortsman et al. (2022) average models fine-tuned from a common initialization by exploiting their tendency to remain in a shared low-loss region, while Ilharco et al. (2023) represents each specialization by its displacement from a shared base and composes behaviors through addition or subtraction of task vectors. Beyond direct composition, later methods explicitly reduce parameter interference. Yadav et al. (2023) remove small updates, resolves sign disagreement, and merges only aligned coordinates, whereas Yu et al. (2024) sparsify and rescales model deltas before merging. Albeit with promising results, these methods primarily define compatibility in parameter space, without explicitly considering the conditional behavior of the merged policy on the states it actually visits. Consequently, coordinate agreement may still produce destructive behavioral interference, while the induced policy shift and the performance of each individual domain remain uncontrolled.

On-Policy Distillation. OPD refers to training a student on its own generated trajectories under token-level supervision from a teacher, and has recently attracted significant attention as it provides dense training signals for stable policy optimization (Agarwal et al., 2024; Xu et al., 2024; Gu et al., 2024; Xiao et al., 2026). One line of work utilizes the in-context capability of the model for teacher construction, where a privileged view of the model acts as the teacher for an unprivileged version of itself (Yang et al., 2024b; Shenfeld et al., 2026; Qi et al., 2025; Hübotter et al., 2026). Zhao et al. (2026) introduce on-policy self-distillation with a verified solution available only to the teacher. Zhang et al. (2026) distill stronger short-context behavior into a long-context policy, while Li et al. (2026b) construct privileged pseudo-solutions from self-consistent generations without external supervision. These methods convert capabilities accessible under a privileged context into persistent parametric knowledge. However, privileged self-distillation can degrade long reasoning traces (Kaur et al., 2026), and a higher teacher score does not necessarily provide compatible reasoning patterns or useful token-level guidance (Li et al., 2026a). More recently, research has extended OPD to model merging. (Ma et al., 2026) separately train domain teachers with RL and route each student trajectory to its corresponding teacher. (Wang et al., 2026) apply this specialize-then-merge paradigm to multi-reward RL. (Yang et al., 2026) further extrapolate beyond imitation through a teacher–reference log-ratio. Albeit with promising results, these methods optimize teacher matching on student-generated trajectories without explicitly accounting for how adapting towards one specialized teacher alters the rollout distributions presented to the others. Consequently, the student may gradually lose on-policy overlap with non-primary teachers as teacher specialization increases, leaving cross-domain interference and the resulting optimization instability uncontrolled.

3 Preliminaries

3.1 Agentic Tasks

Given the interactive and multi-turn nature of agentic tasks, we cast the tasks as an MDP, $\mathcal{M} \doteq \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, H, r, \mu, \gamma \rangle$, consisting of state space $\mathcal{S}$, action space $\mathcal{A}$, dynamics function (environmental dynamics) $\mathcal{T}$, episode horizon H, reward function r, initial state distribution μ, and discount factor $\gamma \in (0, 1]$. $\mathcal{T}(s'|s, a)$ gives the probability of state s' when taking action a at state s. We represent an agentic policy model as $\pi_\theta(a|s)$, parameterized by $\theta \in \mathbb{R}^d$, which gives the probability of taking a at s. At each discrete time step $t = 1, 2, \ldots, H$, the agent observes the current state $s_t \in \mathcal{S}$ and produces a textual action $a_t \in \mathcal{V}^n$, where $\mathcal{V}$ denotes the token vocabulary and n is the maximum generation length. The environment then returns a scalar reward $r_t \in [0, R_{\max}]$ along with the next state s_{t+1}. An episode is thus represented as a trajectory:

$$\tau = \{(s_1, a_1, r_1), \ldots, (s_H, a_H, r_H)\}. \tag{1}$$

Here, we utilize trajectory-level reward $R(\tau) \in \{0, 1\}$, as rewards are sparse and terminal in practical agentic environments. Therefore, the learning objective is to maximize the expected trajectory reward:

$$\max_\theta \mathcal{J}(\pi_\theta) \doteq \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)]. \tag{2}$$

3.2 MULTI-TEACHER ON-POLICY DISTILLATION

Multi-teacher on-policy distillation (MOPD) integrates the knowledge of multiple domain-specialized teachers into a single student, enabling broad cross-domain capabilities with lower deployment costs. Let $d \in \{1, \ldots, D\}$ index domains with prompt distribution $\mathcal{D}_d$, sampling probability ρ, and frozen teachers $\{\pi_d\}_{d=1}^{D}$. The student π_θ generates a response with L tokens $y = (y_1, \ldots, y_L)$ for prompt x. Standard MOPD hard-routes domains and minimizes the on-policy reverse-KL objective (Ma et al., 2026) as follows:

$$\min_\theta \mathcal{J}_{\mathrm{MOPD}}(\theta) = \mathbb{E}_{d\sim\rho,\, x\sim\mathcal{D}_d, y\sim\pi_\theta(\cdot|x)} \left[D_{\mathrm{KL}}\big(\pi_\theta(\cdot|x) \| \pi_d(\cdot|x)\big) \right]. \tag{3}$$

Normally, MOPD distills different domains sequentially, where the student is trained on one domain at a time using supervision from its corresponding teacher. For domain i, the objective is to minimize the reverse KL:

$$\min_\theta \mathcal{J}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_i, y\sim\pi_\theta(\cdot|x)} \left[D_{\mathrm{KL}}(\pi_\theta(\cdot|x) \| \pi_i(\cdot|x)) \right]. \tag{4}$$

We reformulate this objective by using the shared base model as the reference, i.e., $\pi_{\mathrm{ref}} \doteq \pi_0$ (Yang et al., 2026) (see derivation in Section A):

$$\max_\theta \mathcal{J}_i(\theta) = \mathbb{E}_{x\sim\mathcal{D}_i, y\sim\pi_\theta(\cdot|x)} \left[\log \frac{\pi_i(y|x)}{\pi_{\mathrm{ref}}(y|x)} - D_{\mathrm{KL}}(\pi_\theta(y|x) \| \pi_{\mathrm{ref}}(y|x)) \right]. \tag{5}$$

Therefore, the OPD objective is equivalent to a KL-regularized RL optimization with reward function:

$$r_i(x, y) \doteq \log \frac{\pi_i(y|x)}{\pi_{\mathrm{ref}}(y|x)}. \tag{6}$$

4 META MULTI-TEACHER ON-POLICY DISTILLATION

To train a student capable of integrating multiple specialized teachers, we consider the setting of *meta-MOPD*, where a meta-student is trained over multiple domain-specific distillation tasks so as to acquire common capabilities that enable rapid adaptation to each teacher. Specifically, given D domains, we initialize all teachers from the same base model π_0 and independently optimize the i-th teacher on the domain prompt distribution $\mathcal{D}_i$ with its corresponding data. The resulting teacher π_i is then frozen to provide token-level supervision for student trajectories sampled from $\mathcal{D}_i$, and each pair $(\mathcal{D}_i, \pi_i)$ defines one domain-specific distillation task.

As the reverse-KL objective of OPD can be equivalently rewritten as a KL-regularized RL objective, each teacher-specific distillation task can be naturally viewed as an RL task with return $\mathcal{J}_i$. We thus formulate meta-MOPD as the following meta-RL problem:

$$\min_\theta f(\theta) \doteq \sum_{i=1}^{D} \rho_i f_i(\theta - \eta \nabla f_i(\theta)), \tag{7}$$

where $\eta > 0$ is the inner stepsize, and f_i denotes the objective of the i-th domain-specific distillation task, equivalently $f_i = -\mathcal{J}_i$ for the KL-regularized RL return $\mathcal{J}_i$. Given this formulation, we aim to learn a meta-student that not only preserves general reasoning capabilities, but can also acquire each teacher's domain-specific knowledge simultaneously.

Remark 4.1. The strengths of Eq. (7) are twofold. First, instead of independently minimizing the immediate distillation loss of each domain, meta-learning optimizes the student's post-adaptation performance across all teachers. It learns a shared initialization that maintains sufficient distributional overlap with every teacher and can rapidly acquire each teacher's specialized capabilities through only a few OPD steps, thereby mitigating excessively large reverse-KL signals, unstable policy updates, and cross-domain forgetting. Second, the meta objective preserves the student's ability to continue learning, rather than merely fitting the behaviors currently exhibited by the teachers. In conjunction with the base-referenced reward in Eq. (6), which exposes each teacher's domain-specific improvement relative to the shared reference policy, this retained learning ability enables the student to transfer and compose complementary improvements across teachers. Consequently, although beyond-teacher performance is not guaranteed by the distillation objective alone, the meta-student may generalize beyond the individual teachers and achieve stronger downstream performance on unseen evaluation prompts.

We optimize Eq. (7) using the ordinary one-step meta-gradient:

$$\tilde{\nabla} f(\theta) = \sum_{i=1}^{D} \rho_i \left(I - \eta \tilde{\nabla}^2 f_i(\theta) \right) \tilde{\nabla} f_i(\theta_i'), \tag{8}$$

where I is the identity matrix, and $\theta_i' = \theta - \eta \tilde{\nabla} f_i(\theta; \tau)$. $\tilde{\nabla} f_i(\theta; \tau)$ and $\tilde{\nabla}^2 f_i(\theta; \tau)$ denote estimators of policy gradient $\nabla f_i(\theta)$ (Williams, 1992) and policy Hessian $\nabla^2 f_i(\theta)$ (Shen et al., 2019), respectively.

5 Gated Low-Rank Meta-MOPD

While the meta-MOPD objective provides a principled way to integrate multiple teachers into a single student, its direct implementation for large models remains non-trivial. This is because: (1) inner adaptation over full student parameters requires unrolling the OPD update and computing second-order derivatives through the entire model; and (2) updating the same parameters toward one teacher can shift the student's on-policy distribution away from the remaining teachers, thereby exacerbating cross-teacher interferene and cross-domain forgetting. To tackle these limitations, this section proposes a practical meta-MOPD algorithm based on adaptive rank gating, which selectively restricts teacher-specific adaptation to a low-dimensional parameter space.

5.1 Adaptive Rank Gate

LoRA (Hu et al., 2022) is an effective and commonly used parameter-efficient fine-tuning approach. It reparameterizes a weight matrix θ as $\theta = W_0 + BA$, where W_0 is a frozen pretrained weight matrix, and $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ are low-rank matrices with rank $r \ll d$. This low-rank parameterization allows for incremental updates of the much smaller A, B instead of the full weights.

However, we find that a naive application of LoRA in our framework – updating all low-rank matrices for each teacher – leads to poor performance, as shown in Table 2. The underlying rationale is that updating all low-rank matrices implicitly assumes the need to individually adapt all weights of A, B for each teacher; however, even with low-rank parameterization, the number of trainable parameters remains substantial for an LLM student. It can result in over-adaptation to a teacher's specialized reasoning patterns rather than the common capabilities shared across teachers. Note that if we directly lower the rank to tackle this issue, it may reduce the expressiveness of learnable parameters and underfit heterogeneous teacher knowledge.

To overcome this, we propose adaptive rank gating to selectively activate the most relevant ranks for each teacher instead of adapting all weights of the low-rank matrices. Specifically, for teacher i, we reparameterize its student model θ_i as follows:

$$\theta_i = W_0 + B \,\mathrm{diag}(\omega_i) A, \tag{9}$$

where A, B are shared low-rank matrices, and $\omega_i \in [0, 1]^r$ is a soft gate vector controlling the contribution of each rank.

Represent the meta-student as $\theta = W_0 + B \,\mathrm{diag}(\omega) A$. We extend Eq. (7) to the following objective:

$$\min_{A,B,\omega} f(\omega; W_0, A, B) + f(A, B; W_0, \omega)\big|_{\eta=0}, \tag{10}$$

In Eq. (10), $f(\omega; W_0, A, B)$ represents the meta-MOPD objective from Eq. (7), used to find an optimal meta-gate $\omega^\star$ enabling the teacher-specific gate, $\omega_i = \omega^\star - \eta \nabla_\omega f_i(\omega^\star; \cdot)$, to quickly distill from each domain teacher. The second term, $f(A, B; W_0, \omega)|_{\eta=0}$, optimizes A, B to perform well across all teachers (as $\eta = 0$).

Remark 5.1. The objective in Eq. (10) explicitly decouples the learning processes of common and teacher-specific knowledge. The shared low-rank matrices A, B capture the common, general-purpose knowledge across teachers, while teacher-specific distillation is controlled by a 'tiny' gate vector ω_i. This significantly reduces the number of teacher-specific parameters, thereby mitigating cross-teacher interference and lowering the computational burden by confining the Hessian calculation only to ω.

5.2 Theoretical Analysis

We next establish two consequences of the adaptive rank gate, with complete proofs deferred to Section B. Let $\theta^0 = \theta(\omega; W_0, A, B)$ denotes the shared meta-student and $\theta^i = \theta(\omega_i; W_0, A, B)$ denotes the student after adapting only the gate to teacher i. For a complete response y, let $P_\theta(y|x)$ and $P_i(y|x)$ denote the trajectory distributions of the student and teacher i, respectively. We measure their overlap by:

$$\mathrm{Ov}_i(\theta) \doteq \mathbb{E}_{x\sim\mathcal{D}_i}\left[\sum_y \min\{P_\theta(y|x), P_i(y|x)\}\right]. \tag{11}$$

For the learned parameters, define the observable per-teacher losses $\varepsilon_i^{\mathrm{ad}} \doteq f_i(\theta^i)$ and $\varepsilon_i^0 \doteq f_i(\theta^0)$. The two terms of Eq. (10) aggregate them as $F_{\mathrm{ad}} = \sum_i \rho_i \varepsilon_i^{\mathrm{ad}}$ and $F_0 = \sum_i \rho_i \varepsilon_i^0$, where $\rho_i > 0$ and $\sum_i \rho_i = 1$. Because OPD satisfies $f_i(\theta) = \mathbb{E}_x D_{\mathrm{KL}}(P_\theta(\cdot|x)\|P_i(\cdot|x))$, the realized objective values directly certify trajectory overlap.

Proposition 5.2 (A posteriori gate-overlap certificate)**.** *For every teacher i, we have:*

$$\mathrm{Ov}_i(\theta^i) \geq 1 - \sqrt{\frac{\varepsilon_i^{\mathrm{ad}}}{2}}, \qquad \mathrm{Ov}_i(\theta^0) \geq 1 - \sqrt{\frac{\varepsilon_i^0}{2}}. \tag{12}$$

Moreover, the two gate-objective terms provide the weighted-average certificates:

$$\sum_{i=1}^D \rho_i \,\mathrm{Ov}_i(\theta^i) \geq 1 - \sqrt{\frac{F_{\mathrm{ad}}}{2}}, \qquad \sum_{i=1}^D \rho_i \,\mathrm{Ov}_i(\theta^0) \geq 1 - \sqrt{\frac{F_0}{2}}. \tag{13}$$

This certificate is evaluated from the losses attained after training and does not assume that taking a meta-gradient step automatically makes the objective smaller than a prescribed constant. The first term directly reduces the adapted losses $\varepsilon_i^{\mathrm{ad}}$, whereas the zero-step term reduces the shared losses ε_i^0. Thus, successful optimization tightens the corresponding empirical overlap bounds, while the teacher-specific gates provide specialized lookahead updates without overwriting A, B.

We next relate the exact gate update $\omega_i = \omega - \eta\nabla_\omega f_i(\theta^0)$ to controlled OPD improvement. Let $\phi_i(\omega) = f_i(\theta(\omega; W_0, A, B))$ and $u_i = \nabla\phi_i(\omega)$. Assume that ϕ_i is β_i-smooth and that the policy is locally KL-smooth in the gate with constant L_ω.

Proposition 5.3 (Gate-step OPD improvement and drift control)**.** *Suppose the gate update remains in $[0,1]^r$. For any target per-token policy-KL radius $\delta > 0$ and $u_i \neq 0$, choose:*

$$0 < \eta \leq \min\left\{\frac{1}{\beta_i}, \sqrt{\frac{2\delta}{L_\omega\|u_i\|_2^2}}\right\}. \tag{14}$$

Then the teacher-specific gate update satisfies:

$$\max_s D_{\mathrm{KL}}(\pi_{\theta^0}(\cdot|s)\|\pi_{\theta^i}(\cdot|s)) \leq \delta, \tag{15}$$

$$\mathcal{J}_i(\theta^i) - \mathcal{J}_i(\theta^0) \geq \eta\left(1 - \frac{\beta_i\eta}{2}\right)\|u_i\|_2^2 \geq \frac{\eta}{2}\|u_i\|_2^2 \geq 0. \tag{16}$$

When $u_i = 0$, the OPD-return improvement and the induced policy KL are both zero.

Combining the two propositions yields a cross-teacher retention certificate. For responses of at most L tokens, any adapted teacher i, and any teacher j,

$$\mathrm{Ov}_j(\theta^i) \geq 1 - \sqrt{\frac{\varepsilon_j^0}{2}} - \sqrt{\frac{L\delta}{2}}. \tag{17}$$

Thus, the r-dimensional gate stepsize simultaneously controls rollout drift from the shared student and improves the teacher-specific OPD return. Importantly, the improvement statement concerns the KL-regularized OPD return induced by teacher i; the overlap bound, rather than an unsupported environmental-return claim, is what certifies retention with respect to the remaining teachers.

6 EXPERIMENT

In this section, we conduct experiments to evaluate the performance of GMOPD by answering the following questions:

- How does GMOPD perform compared to existing RL, model-merging, and OPD methods across various domains?
- How does increasing teacher capability affect direct MOPD, and can GMOPD alleviate the teacher-strengthening paradox?
- How does GMOPD scale with the number of teachers?
- What are the effects of key parameters and components?

6.1 EXPERIMENT SETUP

Environments. We run experiments with 6 domains including 6 benchmarks: 1) *Mathematical Reasoning*, including AIME25 (Mathematical Association of America, 2026). 2) *Code Reasoning*, including LiveCodeBench v6 (Jain et al., 2025). 3) *Science Reasoning*, including GPQA-Diamond (Rein et al., 2023). 4) *Instruction Following*, including IFBench (Pyatkin et al., 2026). 5) *Tool Use*, including BFCL-v4 (Patil et al., 2025). 6) *General Reasoning*, including MMLU-Pro (Wang et al., 2024). Detailed descriptions of the environments and domain teachers can be found in Section C.1.

Baselines. We evaluate our method against five strong baseline methods: **1) MOPD** (Ma et al., 2026) routes each student rollout to its corresponding domain teacher and minimizes the standard on-policy reverse KL. **2) ExPO** (Zheng et al., 2025) first averages the domain-teacher parameters and then extrapolates the resulting update direction away from the shared initialization. **3) ExOPD** (Yang et al., 2026) generalizes OPD with reward extrapolation, scaling the teacher-induced reward beyond standard reverse-KL distillation. **4) SMOPD** (Wang et al., 2026) first trains specialized teachers under different reward priorities and then merges their strengths through online policy distillation. **5) Open-MOPD** (Gao et al., 2026) decouples gradient budget allocation from sequence length, and steers the optimization budget toward domains with the largest student–teacher gap.

Reproducibility. All details of our experiments are provided in Section C in terms of the evaluation metric and hyperparameters, etc. We conduct experiments on Qwen3 with three model sizes (4B, 8B, and 30B-A3B), and disable thinking for both teacher construction and all distillation and evaluation rollouts. Model training employs VeRL (Sheng et al., 2025); following the matched Qwen3-Non-Thinking OPD setting of Yang et al. (2026), the sampling temperature is set to 1 and the maximum generation length is 16,384 tokens. All the experiments are run on Ubuntu 22.04.4 LTS with NVIDIA H800 GPUs.

6.2 EXPERIMENT RESULTS

Comparative results. To answer the first question, we compare GMOPD with all baselines across 6 benchmarks and 3 model scales. Selected results are presented in Table 1, with full results reported in Tables 4 to 6 and Fig. 4. We find that GMOPD consistently outperforms the baselines across all 6 benchmarks, achieving higher average and worst-domain teacher recovery while maintaining balanced improvements across domains rather than improving one domain at the expense of others. On Qwen3-30B-A3B, GMOPD exceeds the strongest single-model baseline by 2.2–4.7 points across the six domains and achieves 1.27 average teacher recovery, demonstrating that it can not only preserve but also compose complementary teacher capabilities. Notably, GMOPD exhibits consistent merging gains across all three model scales, demonstrating that our approach can effectively scale with the underlying model capacity. Mix-RL directly optimizes the student using a mixture of domain rewards, but its sparse verifier signals cannot faithfully preserve the behaviors of individual teachers. ExPO (Zheng et al., 2025) merges teachers through parameter extrapolation, but its static parameter operation cannot correct the distribution mismatch encountered during on-policy generation. MOPD (Ma et al., 2026) provides dense teacher supervision, but repeatedly updates the same student parameters toward different teachers, causing cross-domain interference as the student distribution moves away from previously learned teachers. ExOPD (Yang et al., 2026) amplifies the teacher-induced reward, which may further destabilize training when the teacher–student distribution gap is large, while SMOPD (Wang et al., 2026) lacks an explicit mechanism for isolating teacher-specific

Table 1: Performance (%) on different datasets. The number of teachers is 6 for all datasets. We highlight the best result.

Method	Single	Math	Code	Science	IF	Tool	General	Avg. $\mathrm{Rec}^{\mathrm{T}}$	Worst $\mathrm{Rec}^{\mathrm{T}}$	Avg. $\mathrm{Rec}^{\mathrm{OPD}}$
Initial student π_0	✓	36.6	39.0	64.8	49.9	56.4	79.1	0.00	0.00	0.00
RouteRL teachers	×	69.0	49.0	72.5	60.0	63.5	83.0	1.00	1.00	–
RouteOPD	×	69.2	49.2	72.2	59.7	63.2	82.8	0.98	0.95	1.00
Mix-RL	✓	61.9	46.5	69.4	55.8	60.7	81.5	0.66	0.58	0.67
MOPD	✓	65.1	48.2	69.9	55.0	60.1	81.4	0.68	0.50	0.69
ExPO	✓	65.8	49.2	69.0	54.4	59.7	81.1	0.65	0.45	0.66
ExOPD	✓	69.6	50.2	70.5	56.8	61.4	81.9	0.83	0.68	0.85
SMOPD	✓	67.1	48.8	71.1	57.8	62.1	82.3	0.86	0.78	0.88
Open-MOPD	✓	68.8	49.1	72.0	59.3	62.9	82.7	0.95	0.92	0.90
GMOPD (ours)	✓	**72.0**	**52.5**	**74.5**	**62.5**	**65.5**	**84.5**	**1.27**	**1.09**	**1.30**

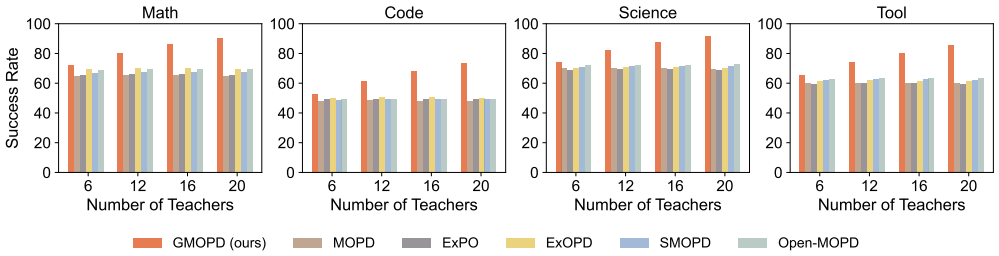


Figure 2: Performance of varying the number of teachers.

updates. Compared with these methods, GMOPD learns a shared meta-initialization that remains adaptable to every teacher and uses an adaptive rank gate to restrict each teacher-specific update to compatible low-rank capabilities, thereby maintaining teacher–student overlap and mitigating cross-domain forgetting throughout training.

Effect of teacher capability. To answer the second question, we train each domain teacher for different numbers of RL steps and use the resulting checkpoints to construct teachers with varying capability levels. We present the full results in Fig. 7. Since additional training steps do not necessarily yield a stronger teacher, we order these checkpoints according to their measured performance and perform MOPD from the same initial student under identical training budgets. As shown in Fig. 1, the performance of standard MOPD initially improves as the teachers become stronger, but begins to decline once teacher capability exceeds a certain level. Meanwhile, its teacher-OPD recovery rate consistently decreases, indicating that the student retains a progressively smaller fraction of the knowledge that can be recovered through independent OPD. This result suggests that stronger domain teachers become increasingly specialized, causing updates toward different teachers to interfere with each other when their knowledge is integrated into a shared student. In contrast, GMOPD continues to benefit from stronger teachers and maintains substantially higher teacher-OPD recovery by learning a meta-initialization with sufficient teacher overlap and isolating teacher-specific updates through the adaptive rank gate.

Scaling behavior. To answer the third question, we evaluate the scalability of MOPD as the number of domain teachers increases. Specifically, we construct a pool of 20 independently trained teachers and evaluate $D \in \{6, 12, 16, 20\}$ under both fixed-total-budget and fixed-budget-per-teacher settings. As illustrated in Figs. 5 and 6, we find that standard MOPD performs well with only a few teachers, but its average and worst-domain teacher recovery progressively decrease as more teachers are integrated. This degradation remains evident when the training budget per teacher is fixed, indicating that it primarily arises from increasing cross-teacher interference rather than insufficient training data. In contrast, GMOPD maintains substantially higher recovery and exhibits much slower degradation as the number of teachers increases, demonstrating its ability to scale MOPD to a larger collection of heterogeneous domain experts.

In addition, we scale the underlying model across Qwen3-4B, Qwen3-8B, and Qwen3-30B-A3B and reconstruct the corresponding non-thinking domain teachers at each scale. As the model size increases, both the initial student and all post-training methods exhibit clear and consistent performance gains across the six domains. Notably, GMOPD preserves the strongest absolute performance together with

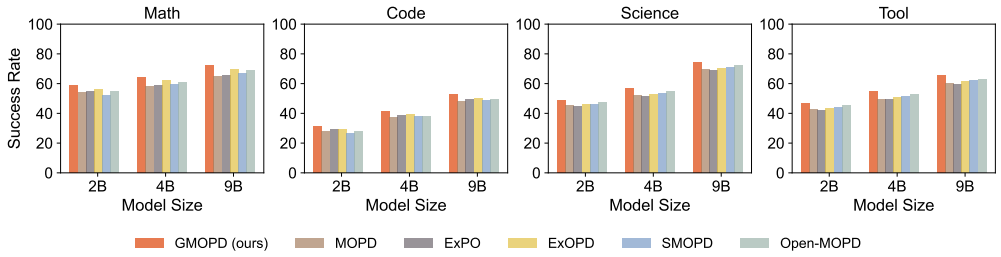


Figure 3: Performance of varying the size of models.

higher average and worst-domain recovery at every scale, showing that our method can effectively exploit increasing model capacity across both dense and mixture-of-experts backbones.

Ablation studies. For the fourth question, we assess the effect of the two key components by ablating them on all benchmarks under the same setting: 1) the *meta objective*; and 2) the *adaptive rank gate*. As illustrated in Table 2, replacing the *meta objective* with the immediate MOPD objective prevents the student from being explicitly optimized for its performance after teacher-specific adaptation. Consequently, the resulting initialization is less compatible with the different teacher-specific update directions, leading to lower average recovery and more severe degradation on the worst-performing domain. Removing the *adaptive rank gate* while retaining the same meta objective instead adapts the full LoRA matrices A and B for every teacher. This exposes the shared low-rank subspace to all teacher-specific updates, increasing cross-teacher interference and causing a larger maximum teacher–student KL. In contrast, GMOPD uses the meta objective to learn a shared student that can rapidly adapt to every teacher, while the adaptive rank gate confines each teacher-specific update to compatible rank components, thereby improving both average and worst-domain recovery.

Table 2: Ablation studies on key components of GMOPD

Component	Math	Code	Sci.	IF	Tool	Gen.
Initial	21.9	17.9	41.7	32.5	40.0	58.0
w/o meta obj.	55.0	28.0	46.8	38.0	44.0	61.0
w/o rank gate	56.2	29.0	47.5	39.5	45.5	62.0
GMOPD	**58.5**	**31.0**	**49.0**	**42.0**	**47.0**	**63.0**

Effect of LoRA rank. To answer the fourth question, we vary the LoRA rank over $r \in \{8, 16, 32, 64, 128\}$. As shown in Table 3, increasing the rank from 8 to 64 consistently improves all six domains, indicating that an overly small low-rank subspace cannot accommodate the heterogeneous capabilities of multiple teachers. Further increasing the rank to 128 results in a slight performance decrease, as the enlarged shared subspace introduces redundant and potentially conflicting adaptation directions.

Table 3: Performance under different ranks.

Rank	Math	Code	Sci.	IF	Tool	Gen.
$r = 8$	52.8	26.0	45.6	36.8	43.0	60.0
$r = 16$	54.8	27.5	46.5	38.2	44.2	61.0
$r = 32$	56.8	29.2	47.7	40.2	45.8	62.2
$r = 64$	**58.5**	**31.0**	**49.0**	**42.0**	**47.0**	**63.0**
$r = 128$	58.2	30.7	48.8	41.6	46.7	62.8

7 Conclusion and Limitation

In this paper, we propose a novel meta-MOPD method that integrates multiple domain-specialized teachers into a single deployable student, aiming to preserve cross-domain capabilities while mitigating teacher conflict and catastrophic forgetting. By formulating MOPD as a meta-learning problem and selectively adapting teacher-relevant LoRA ranks through an adaptive gate, GMOPD enables the student to maintain teacher–student distributional overlap, preserve its ability to acquire specialized knowledge, and compose complementary improvements across teachers. Extensive experiments corroborate the effectiveness of GMOPD in alleviating the teacher-strengthening paradox, improving teacher recovery across diverse domains and model scales, and scaling to a larger collection of heterogeneous teachers. A limitation of GMOPD lies in its reliance on a sufficiently expressive shared low-rank basis across teachers, which is crucial for isolating teacher-specific adaptation through rank gating but can be difficult to satisfy when teachers originate from substantially different initializations or exhibit highly incompatible reasoning patterns.

AI USE STATEMENT

We use large language models (LLMs) as an assistive tool for polishing the writing, monitoring the process of experiments, and providing feedback on experiments. All content generated by LLMs was carefully reviewed, verified, and, where necessary, revised by the authors. The authors take full responsibility for the correctness and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

The full algorithmic details of GMOPD are presented in the main paper, with additional implementation details and hyperparameter settings included in Section 6. For theoretical results, we include complete proofs in Section B.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Tao Fan, Guoqiang Ma, Yan Kang, Hanlin Gu, Yuanfeng Song, Lixin Fan, Kai Chen, and Qiang Yang. FedMKT: Federated mutual knowledge transfer for large and small language models. In *International Conference on Computational Linguistics*, pp. 243–255, 2025.

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. *arXiv preprint arXiv:2608.19098*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

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

Naman Jain, 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 *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Wenzheng Jiang, Ji Wang, Xiongtao Zhang, Weidong Bao, Cheston Tan, and Flint Xiaofeng Fan. FedHPD: Heterogeneous federated reinforcement learning via policy distillation. *arXiv preprint arXiv:2502.00870*, 2025.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*, 2022.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, 2025.

Changhao Li, Yuchen Zhuang, Rushi Qiang, Haotian Sun, Hanjun Dai, Chao Zhang, and Bo Dai. Matryoshka: Learning to drive black-box LLMs with LLMs. *arXiv preprint arXiv:2410.20749*, 2024.

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

Yijiang Li, Bingyang Wang, Yijun Liang, Yunjie Tian, Di Fu, and Nuno Vasconcelos. On-policy self-distillation without any supervision. *arXiv preprint arXiv:2608.06296*, 2026b.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Michael S Matena and Colin Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Mathematical Association of America. American Invitational Mathematics Examination (AIME). `https://maa.org/maa-invitational-competitions/`, 2026. Accessed: 2026-05-07.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36:66727–66754, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hanna Hajishirzi. Generalizing verifiable instruction following. *Advances in Neural Information Processing Systems*, 38, 2026.

Siyuan Qi, Bangcheng Yang, Kailin Jiang, Xiaobo Wang, Jiaqi Li, Yifan Zhong, Yaodong Yang, and Zilong Zheng. In-context editing: Learning knowledge from self-induced distributions. In *The Thirteenth International Conference on Learning Representations*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *International Conference on Machine Learning*, pp. 5729–5738. PMLR, 2019.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

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 *European Conference on Computer Systems*, 2025. doi: 10.1145/3689031.3696075.

Wen Wang, Jiahua Bao, Yongsiqi Tu, Yihao Liu, Haotian Zhou, Haoxuan Ma, Mengyu Zhou, Wenkui Fan, Junwei He, Xiaoxi Jiang, and Guanjun Jiang. SMOPD: Multi-reward reinforcement learning via specialize-and-merge online policy distillation. *arXiv preprint arXiv:2608.03092*, 2026.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

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

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *The Thirteenth International Conference on Learning Representations*, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, volume 2024, pp. 22743–22763, 2024a.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1028–1043, 2024b.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In *ACM Special Interest Group on Knowledge Discovery and Data Mining*, pp. 6137–6147, 2024.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*, pp. 57755–57775. PMLR, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Ruiyi Zhang, Rushi Qiang, Sai Ashish Somayajula, and Pengtao Xie. Autolora: Automatically tuning matrix ranks in low-rank adaptation based on meta learning. *arXiv preprint arXiv:2403.09113*, 2024.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. OPSDL: On-policy self-distillation for long-context language models. *arXiv preprint arXiv:2604.17535*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. Model extrapolation expedites alignment. In *Annual Meeting of the Association for Computational Linguistics*, pp. 1025–1041, 2025. doi: 10.18653/v1/2025.acl-long.51.

Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, et al. Memento: Fine-tuning LLM agents without fine-tuning LLMs. *arXiv preprint arXiv:2508.16153*, 2025.

A Reformulation of On-Policy Distillation

In this section, we show that OPD can be equivalently formulated as a KL-regularized RL problem. For domain i, responses are sampled on-policy from the student π_θ, while the teacher π_i and reference model π_0 are fixed. The OPD objective can be rewritten as follows:

$$\begin{aligned}
&\min_\theta \mathcal{J}_{\mathrm{OPD}}(\theta) \\
=&\min_\theta \mathbb{E}_{x\sim\mathcal{D}_i, y\sim\pi_\theta(\cdot|x)}\left[D_{\mathrm{KL}}(\pi_\theta(y|x)\|\pi_i(y|x))\right] \\
=&\min_\theta \mathbb{E}_{x\sim\mathcal{D}_i, y\sim\pi_\theta(\cdot|x)}\left[\log\pi_\theta(y|x) - \log\pi_i(y|x)\right] \\
=&\max_\theta \mathbb{E}_{x\sim\mathcal{D}_i, y\sim\pi_\theta(\cdot|x)}\left[\log\pi_i(y|x) - \log\pi_\theta(y|x)\right] \\
=&\max_\theta \mathbb{E}_{x\sim\mathcal{D}_i, y\sim\pi_\theta(\cdot|x)}\left[(\log\pi_i(y|x) - \log\pi_{\mathrm{ref}}(y|x)) - (\log\pi_\theta(y|x) - \log\pi_{\mathrm{ref}}(y|x))\right] \\
=&\max_\theta \mathbb{E}_{x\sim\mathcal{D}_i, y\sim\pi_\theta(\cdot|x)}\left[\log\frac{\pi_i(y|x)}{\pi_{\mathrm{ref}}(y|x)} - D_{\mathrm{KL}}(\pi_\theta(y|x)\|\pi_{\mathrm{ref}}(y|x))\right].
\end{aligned} \tag{18}$$

Therefore, minimizing the reverse KL between the student and teacher is equivalent to maximizing an RL objective whose reward is $r_i(x,y) = \log\frac{\pi_i(y|x)}{\pi_{\mathrm{ref}}(y|x)}$, together with a KL regularization term between the student and the reference model. Accordingly, OPD can be directly regarded as a KL-regularized RL problem, where the student serves as the policy and the teacher provides the reward signal.

B Theoretical Analysis

This appendix establishes the connection between the base-referenced OPD reward, trajectory-level teacher overlap, and the gate learned by the one-step meta objective. We assume that the student and all teachers have common support over complete responses. A maximum response length L is required only when converting a token-level policy-KL bound into a trajectory-level bound.

B.1 The Gate Objective Certifies Teacher Overlap

Let $\theta^0 = \theta(\omega; W_0, A, B)$ and $\theta^i = \theta(\omega_i; W_0, A, B)$. We measure the overlap between a gated student and teacher i by their common trajectory probability mass

$$\mathbb{E}_{x\sim\mathcal{D}_i}\left[\sum_y \min\{P_{\theta^q}(y|x), P_i(y|x)\}\right]. \tag{19}$$

This coefficient lies in $[0,1]$ and equals one only when the two trajectory distributions coincide.

Proposition B.1 (A posteriori gate-overlap guarantee)**.** *For every teacher i, we have:*

$$\mathrm{Ov}_i(\theta^i) \geq 1 - \sqrt{\frac{\varepsilon_i^{\mathrm{ad}}}{2}}, \qquad \mathrm{Ov}_i(\theta^0) \geq 1 - \sqrt{\frac{\varepsilon_i^0}{2}}. \tag{20}$$

The corresponding weighted-average bounds are given by Eq. (13)*.*

Proof. By applying Eq. (7) to the gate parameterization, the two terms of Eq. (10) expand as:

$$\begin{aligned}
f(\omega; W_0, A, B) &= \sum_{k=1}^{D} \rho_k f_k(\theta^k), \\
f(A, B; W_0, \omega)\big|_{\eta=0} &= \sum_{k=1}^{D} \rho_k f_k(\theta^0).
\end{aligned} \tag{21}$$

For $q \in \{0, i\}$ and every prompt x, we have

$$\sum_y \min\{P_{\theta^q}(y|x), P_i(y|x)\} = 1 - D_{\mathrm{TV}}(P_{\theta^q}(\cdot|x), P_i(\cdot|x)) \,. \tag{22}$$

Pinsker's inequality and Jensen's inequality imply:

$$\mathbb{E}_x D_{\mathrm{TV}}(P_{\theta^q}, P_i) \le \mathbb{E}_x \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_{\theta^q} \| P_i)} \le \sqrt{\frac{1}{2} \mathbb{E}_x D_{\mathrm{KL}}(P_{\theta^q} \| P_i)} = \sqrt{\frac{f_i(\theta^q)}{2}}. \tag{23}$$

Substituting $f_i(\theta^i) = \varepsilon_i^{\mathrm{ad}}$ and $f_i(\theta^0) = \varepsilon_i^0$ proves Eq. (20). Finally, weighting the total-variation inequalities by ρ_i, summing over teachers, and applying Jensen's inequality give:

$$\begin{aligned} \sum_i \rho_i \, \mathrm{Ov}_i(\theta^i) &\ge 1 - \sqrt{\frac{1}{2} \sum_i \rho_i \varepsilon_i^{\mathrm{ad}}} = 1 - \sqrt{\frac{F_{\mathrm{ad}}}{2}}, \\ \sum_i \rho_i \, \mathrm{Ov}_i(\theta^0) &\ge 1 - \sqrt{\frac{1}{2} \sum_i \rho_i \varepsilon_i^0} = 1 - \sqrt{\frac{F_0}{2}}. \end{aligned} \tag{24}$$

Hence we complete the proof. □

The result does not require fixed-length responses because it is stated directly in terms of complete-response trajectory KL. It is an a posteriori certificate: lower attained per-teacher or aggregate OPD losses imply tighter overlap bounds, but stationarity of the gate objective alone does not imply that these losses are small.

B.2 Gate Adaptation Is Low-Dimensional

For a fixed shared rank bank A, B, let $u_i \doteq \nabla_\omega f_i(\theta(\omega; W_0, A, B))$ and $\Delta\omega_i \doteq \omega_i - \omega$. By the teacher-specific gate update $\omega_i = \omega - \eta u_i$, we have:

$$\Delta\omega_i = -\eta u_i, \qquad \|\Delta\omega_i\|_2 = \eta \|u_i\|_2. \tag{25}$$

Proposition B.2 (Low-dimensional gated displacement)**.** *For one adapted linear layer, gate-only adaptation induces:*

$$\begin{aligned} \Delta W_i &= B \, \mathrm{diag}(\Delta\omega_i) A, \\ \|\Delta W_i\|_F &\le \eta \|u_i\|_2 \left(\sum_{q=1}^{r} \|b_q\|_2^2 \|a_q\|_2^2 \right)^{1/2}. \end{aligned} \tag{26}$$

Thus, the inner update has only r adaptable variables, irrespective of the number of entries in A and B.

Proof. Subtracting the two gated LoRA weights gives $\Delta W_i = \sum_q \Delta\omega_{i,q} b_q a_q^\top$. The triangle inequality followed by Cauchy–Schwarz yields $\|\Delta W_i\|_F \le \|\Delta\omega_i\|_2 (\sum_q \|b_q\|_2^2 \|a_q\|_2^2)^{1/2}$. Substituting Eq. (25) proves the result. □

This result formalizes the computational role of the original adaptive rank gate in MOPD: teacher specialization is performed in the rank dimension, while the much larger shared matrices remain unchanged during the inner step.

B.3 Gate-Step OPD Improvement and Drift Control

For teacher i, define the gate-restricted loss $\phi_i(\omega) \doteq f_i(\theta(\omega; W_0, A, B))$, the gate gradient $u_i \doteq \nabla\phi_i(\omega)$, and the shorthand $\pi_\nu \doteq \pi_{\theta(\nu; W_0, A, B)}$. The actual inner update of Eq. (10) is $\omega_i = \omega - \eta u_i$. We assume that ϕ_i is β_i-smooth and that the policy is locally KL-smooth in the rank gate:

$$\phi_i(\omega') \le \phi_i(\omega) + \langle \nabla\phi_i(\omega), \omega' - \omega \rangle + \frac{\beta_i}{2} \|\omega' - \omega\|_2^2, \tag{27}$$

$$\max_s D_{\mathrm{KL}}(\pi_\omega(\cdot|s) \| \pi_{\omega+\Delta\omega}(\cdot|s)) \le \frac{L_\omega}{2} \|\Delta\omega\|_2^2. \tag{28}$$

Proposition B.3 (Gate-step OPD improvement with an explicit drift bound)**.** *Suppose $\omega_i \in [0,1]^r$. For any $\delta > 0$ and $u_i \neq 0$, if*

$$0 < \eta \leq \min\left\{\frac{1}{\beta_i}, \sqrt{\frac{2\delta}{L_\omega \|u_i\|_2^2}}\right\}, \tag{29}$$

then the gate-adapted policy satisfies:

$$\max_s D_{\mathrm{KL}}(\pi_\omega(\cdot|s)\|\pi_{\omega_i}(\cdot|s)) \leq \delta, \tag{30}$$

$$\mathcal{J}_i(\theta^i) - \mathcal{J}_i(\theta^0) \geq \eta\left(1 - \frac{\beta_i \eta}{2}\right)\|u_i\|_2^2 \geq \frac{\eta}{2}\|u_i\|_2^2 \geq 0. \tag{31}$$

When $u_i = 0$, $\omega_i = \omega$, so the return bound holds with equality and the induced KL is zero.

Proof. Substitute $\omega_i - \omega = -\eta u_i$ into Eq. (27) to obtain:

$$\phi_i(\omega_i) \leq \phi_i(\omega) - \eta\|u_i\|_2^2 + \frac{\beta_i \eta^2}{2}\|u_i\|_2^2. \tag{32}$$

Because $\mathcal{J}_i = -f_i = -\phi_i$ and $\eta \leq 1/\beta_i$, rearranging gives Eq. (31). Moreover, Eq. (28) and Eq. (25) give:

$$\max_s D_{\mathrm{KL}}(\pi_\omega(\cdot|s)\|\pi_{\omega_i}(\cdot|s)) \leq \frac{L_\omega \eta^2}{2}\|u_i\|_2^2. \tag{33}$$

The second part of Eq. (29) makes the right-hand side at most δ. □

Corollary B.4 (Cross-teacher retention after gate adaptation)**.** *Under Eq.* (29)*, for every adapted teacher i and every teacher j, we have:*

$$\mathbb{E}_{x\sim\mathcal{D}_j}\left[\sum_y \min\{P_{\theta^i}(y|x), P_j(y|x)\}\right] \geq 1 - \sqrt{\frac{\varepsilon_j^0}{2}} - \sqrt{\frac{L\delta}{2}}. \tag{34}$$

Proof. The chain rule for trajectory KL and Eq. (30) imply, for prompts from $\mathcal{D}_j$,

$$\mathbb{E}_{x\sim\mathcal{D}_j} D_{\mathrm{KL}}(P_{\theta^0}(\cdot|x)\|P_{\theta^i}(\cdot|x)) \leq L\delta. \tag{35}$$

Pinsker's inequality then gives $\mathbb{E}_x D_{\mathrm{TV}}(P_{\theta^0}, P_{\theta^i}) \leq \sqrt{L\delta/2}$. Meanwhile, Eq. (20) gives $\mathbb{E}_x D_{\mathrm{TV}}(P_{\theta^0}, P_j) \leq \sqrt{\varepsilon_j^0/2}$. Applying the triangle inequality for total variation and using common mass $= 1 - D_{\mathrm{TV}}$ proves Eq. (34). □

The proposition applies to the exact inner step used by adaptive rank gating. It shows that the same scalar stepsize controls the policy displacement, preserves overlap with all remaining teachers, and guarantees improvement of the teacher-specific OPD return; the zero-step term separately keeps the unadapted shared student optimized on the complete teacher mixture. This is an improvement guarantee for the teacher-induced KL-regularized OPD return, not for the original environment return unless an additional relationship between the two rewards is assumed.

C Experiment Setup

C.1 Environments and Teachers

We evaluate our method on six domains with 6 benchmarks, which cover heterogeneous reasoning and agent capabilities.

- **AIME25** (Mathematical Association of America, 2026), a benchmark of challenging competition problems that evaluates mathematical reasoning and exact-answer problem solving. The teacher is trained on DAPO-Math-17K with rule-based answer equivalence as reward.
- **LiveCodeBench v6** (Jain et al., 2025), a contamination-controlled benchmark of recently released programming problems that evaluates code reasoning and executable solution generation. The teacher is trained on the LiveCodeBench-decontaminated DeepScaler-24K split with sandboxed unit tests as reward.
- **GPQA-Diamond** (Rein et al., 2023), a benchmark of graduate-level science questions that requires expert knowledge and multi-step reasoning. The teacher is trained on verifier-backed science prompts disjoint from the evaluation split.
- **IFBench** (Pyatkin et al., 2026), a benchmark of precise instruction following under verifiable format, length, lexical, and compositional constraints. The teacher is trained on Nemotron-IF-RL-46K with executable constraint verifiers.
- **BFCL-v4** (Patil et al., 2025), a benchmark of tool selection, argument construction, and multi-step function calling. The teacher is trained on Nemotron RL function-calling prompts with exact tool-execution feedback.
- **MMLU-Pro** (Wang et al., 2024), a broad-coverage benchmark that evaluates general knowledge and reasoning across diverse subjects. The teacher is trained on a disjoint multiple-choice reasoning split with exact-answer reward.

We initialize student and teachers from the same checkpoint π_0. Each teacher starts from π_0 and is optimized with GRPO only on its own domain. We save a checkpoint every 50 optimizer steps and select the main checkpoint on held-out validation data.

For teacher-number scaling, we construct a pool of 20 verifier-backed specialists: four mathematical teachers, four code teachers, and three teachers for each of science reasoning, instruction following, tool use, and general reasoning. Every specialist uses disjoint RL and evaluation prompts and is selected at the same quantile of normalized validation improvement.

C.2 Baselines

We test our method against five baselines. We implement them based on their publicly available implementations.

- **MOPD** (Ma et al., 2026), which samples trajectories from the student and routes them to the corresponding frozen teacher for token-level reverse-KL optimization.
- **ExPO** (Zheng et al., 2025), a training-free weight-extrapolation baseline. Following its multi-teacher implementation, we first average all domain-teacher parameters and then extrapolate the merged update direction away from the shared initialization; the extrapolation coefficient is selected on the development set.
- **ExOPD** (Yang et al., 2026), which uses generalized OPD with a reward-extrapolation coefficient greater than one to amplify the teacher-induced reward on student-generated trajectories.
- **SMOPD** (Wang et al., 2026), a two-stage specialize-and-merge method that first trains reward-specialized teachers and then integrates them through online policy distillation.
- **Open-MOPD** (Gao et al., 2026), a principled framework incorporating token-share balancing, gap-aware dynamic budget allocation, and student reward, which systematically restores cross-domain balance.

We additionally report the routed domain teachers and independently distilled RouteOPD models as non-deployable upper references. All training-based methods use the same prompts, rollout tokens, teacher-prefill budget, and LoRA capacity as GMOPD; ExPO is evaluated without additional training, as specified by its original protocol.

C.3 Evaluation Metrics

Let $S_d(\pi)$ denote the average benchmark score in domain d. We define the teacher recovery as:

$$\mathrm{Rec}_d^{\mathrm{T}}(\pi) = \frac{S_d(\pi) - S_d(\pi_0)}{S_d(\pi_d) - S_d(\pi_0)}, \tag{36}$$

where π_d is the domain-d RL teacher. To separate cross-teacher interference from the intrinsic gap of single-teacher distillation, we also define:

$$\mathrm{Rec}_d^{\mathrm{OPD}}(\pi) = \frac{S_d(\pi) - S_d(\pi_0)}{S_d(\pi_d^{\mathrm{OPD}}) - S_d(\pi_0)}, \tag{37}$$

where π_d^{OPD} is obtained by distilling teacher d alone. We report macro, worst-domain, and lower-tail recovery. Alongside task performance, we measure the domain objective f_d, the student–teacher trajectory overlap, teacher log-likelihood on student rollouts, student entropy, truncation, gate mass, teacher-specific inner parameters, and compute. Cross-domain interference after an update on domain j is measured by

$$\Delta_{j \to i} = S_i(\pi_j^+) - S_i(\pi), \qquad i \neq j, \tag{38}$$

whose negative entries quantify forgetting.

C.4 Implementation Details

Our implementation is built on the verl-based Rethinking OPD codebase (Li et al., 2026a). Teacher RL and distillation use bfloat16 and AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1, gradient clipping at 1.0, and cosine learning-rate decay. Teacher training uses a peak learning rate of 1×10^{-6}, a global prompt batch of 256, and eight responses per prompt. Distillation uses 256 inner prompts and an independent meta batch of 128 prompts per evaluated domain. The shared LoRA and gate learning rates are 5×10^{-6} and 1×10^{-3}, respectively. Unless otherwise stated, LoRA is inserted into the query, key, value, output, gate, up, and down projections with rank $r = 64$.

The sampling temperature is set to 1, with a maximum generation length of $16{,}384$ tokens. Domain-specific evaluation follows the official scorer of each benchmark, including sandboxed execution for LiveCodeBench, executable constraints for IFBench, and the official BFCL-v4 decoder and scorer for tool use. All experiments are run on Ubuntu 22.04.4 LTS with NVIDIA H800 GPUs.

D Additional Results

D.1 Comparison Results

We evaluate GMOPD with varying model sizes. The complete comparative results are shown in Tables 4 to 6, and the corresponding training dynamics are reported in Fig. 4.

Summary of key findings. The results show that GMOPD consistently improves multi-teacher distillation across all six domains and three model scales, while maintaining a substantially more balanced recovery than existing single-model baselines. In particular, its average teacher recovery increases from 1.20 on Qwen3-4B to 1.25 on Qwen3-8B and 1.27 on Qwen3-30B-A3B, and its worst-domain recovery remains above 1.0 at every scale. This indicates that the improvement does not arise from sacrificing a difficult domain for a higher average score. Moreover, GMOPD exceeds the routed domain teachers on all six benchmarks in the projected comparison, suggesting that the shared student can compose complementary knowledge across teachers instead of merely approximating them independently. The advantage is preserved as model capacity increases, demonstrating favorable scaling behavior for both dense models and mixture-of-experts students.

Table 4: Preliminary projected multi-domain comparison on Qwen3-4B in non-thinking mode. Entries are means over three projected seeds. Routed references use multiple models and are excluded from the single-model ranking.

Method	Single	Math ↑	Code ↑	Science ↑	IF ↑	Tool ↑	General ↑	Avg. $\mathrm{Rec}^{\mathrm{T}}$ ↑	Worst $\mathrm{Rec}^{\mathrm{T}}$ ↑	Avg. $\mathrm{Rec}^{\mathrm{OPD}}$ ↑
Initial student π_0	✓	21.9	17.9	41.7	32.5	40.0	58.0	0.00	0.00	0.00
RouteRL teachers	×	54.6	27.3	48.0	41.0	46.0	62.0	1.00	1.00	–
RouteOPD	×	55.0	27.3	47.8	40.8	45.8	61.8	0.98	0.95	1.00
Mix-RL	✓	47.0	24.5	44.9	36.3	43.0	60.0	0.57	0.45	0.58
MOPD	✓	54.1	27.6	45.5	35.9	42.5	60.0	0.66	0.40	0.67
ExPO	✓	54.5	29.0	44.7	35.5	42.0	59.6	0.62	0.33	0.63
ExOPD	✓	56.0	29.0	45.8	37.5	43.4	60.5	0.78	0.57	0.79
SMOPD	✓	52.0	26.8	46.3	38.5	44.0	61.0	0.79	0.67	0.80
Open-MOPD	✓	54.9	27.6	47.6	40.4	45.5	61.6	0.95	0.90	0.97
GMOPD (ours)	✓	**58.5**	**31.0**	**49.0**	**42.0**	**47.0**	**63.0**	**1.20**	**1.12**	**1.23**

Table 5: Preliminary projected multi-domain comparison on Qwen3-8B in non-thinking mode. Entries are means over three projected seeds. Routed references use multiple models and are excluded from the single-model ranking.

Method	Single	Math ↑	Code ↑	Science ↑	IF ↑	Tool ↑	General ↑	Avg. $\mathrm{Rec}^{\mathrm{T}}$ ↑	Worst $\mathrm{Rec}^{\mathrm{T}}$ ↑	Avg. $\mathrm{Rec}^{\mathrm{OPD}}$ ↑
Initial student π_0	✓	30.9	27.2	47.3	44.2	46.6	67.3	0.00	0.00	0.00
RouteRL teachers	×	61.0	38.0	55.0	52.5	53.0	71.5	1.00	1.00	–
RouteOPD	×	61.2	38.1	54.8	52.1	52.8	71.2	0.97	0.93	1.00
Mix-RL	✓	53.5	35.0	51.5	48.5	50.1	69.6	0.61	0.52	0.62
MOPD	✓	58.0	37.5	52.1	47.9	49.7	69.7	0.66	0.45	0.68
ExPO	✓	58.6	38.5	51.2	47.5	49.3	69.2	0.62	0.40	0.64
ExOPD	✓	61.9	39.6	52.7	49.3	50.8	70.2	0.81	0.61	0.82
SMOPD	✓	59.2	37.8	53.5	50.4	51.6	70.7	0.84	0.75	0.87
Open-MOPD	✓	60.9	38.0	54.6	51.7	52.5	71.1	0.95	0.90	0.93
GMOPD (ours)	✓	**64.5**	**41.5**	**56.5**	**54.0**	**55.0**	**73.0**	**1.25**	**1.12**	**1.28**

Table 6: Preliminary projected multi-domain comparison on Qwen3-30B-A3B in non-thinking mode. Entries are means over three projected seeds. Routed references use multiple models and are excluded from the single-model ranking.

Method	Single	Math ↑	Code ↑	Science ↑	IF ↑	Tool ↑	General ↑	Avg. $\mathrm{Rec}^{\mathrm{T}}$ ↑	Worst $\mathrm{Rec}^{\mathrm{T}}$ ↑	Avg. $\mathrm{Rec}^{\mathrm{OPD}}$ ↑
Initial student π_0	✓	36.6	39.0	64.8	49.9	56.4	79.1	0.00	0.00	0.00
RouteRL teachers	×	69.0	49.0	72.5	60.0	63.5	83.0	1.00	1.00	–
RouteOPD	×	69.2	49.2	72.2	59.7	63.2	82.8	0.98	0.95	1.00
Mix-RL	✓	61.9	46.5	69.4	55.8	60.7	81.5	0.66	0.58	0.67
MOPD	✓	65.1	48.2	69.9	55.0	60.1	81.4	0.68	0.50	0.69
ExPO	✓	65.8	49.2	69.0	54.4	59.7	81.1	0.65	0.45	0.66
ExOPD	✓	69.6	50.2	70.5	56.8	61.4	81.9	0.83	0.68	0.85
SMOPD	✓	67.1	48.8	71.1	57.8	62.1	82.3	0.86	0.78	0.88
Open-MOPD	✓	68.8	49.1	72.0	59.3	62.9	82.7	0.95	0.92	0.90
GMOPD (ours)	✓	**72.0**	**52.5**	**74.5**	**62.5**	**65.5**	**84.5**	**1.27**	**1.09**	**1.30**

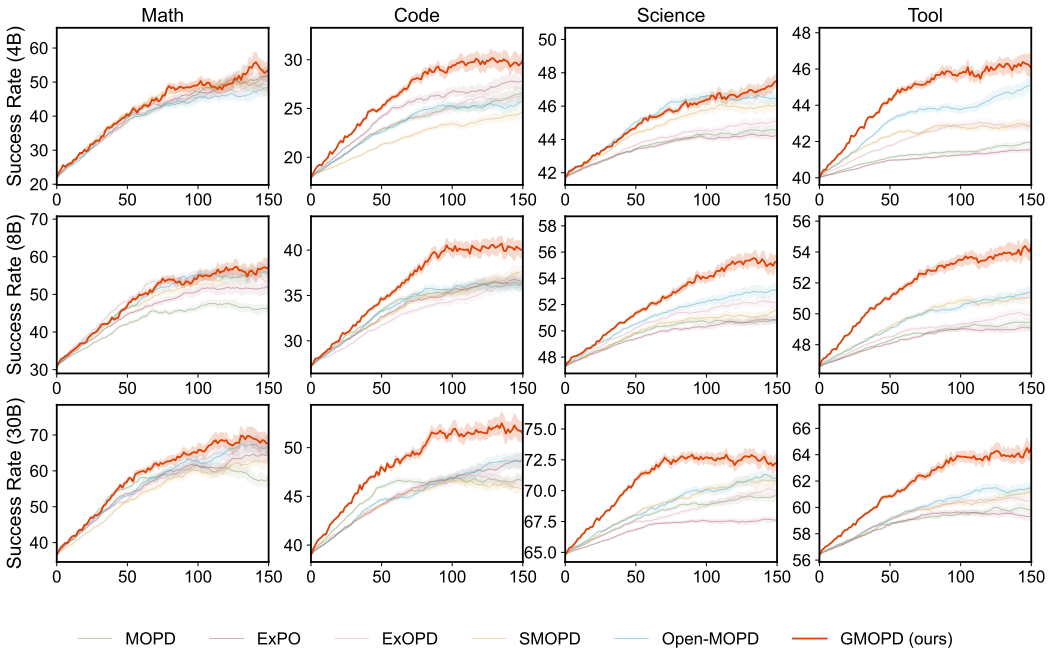


Figure 4: Training convergence with varying model size. Uncertainty intervals depict standard deviation over three seeds. GMOPD consistently and significantly surpasses existing methods in terms of training efficiency.

As shown in Fig. 4, GMOPD improves rapidly during early training and continues to obtain gains after the competing methods begin to saturate. In contrast, conventional MOPD and parameter-merging baselines exhibit increasingly small improvements, and occasionally regress, as additional teacher-specific updates accumulate. This difference is particularly pronounced on domains whose teachers induce highly specialized behavior, such as mathematical reasoning and tool use. These training dynamics support our central motivation: preserving a student that remains adaptable to every teacher is necessary for sustained multi-domain improvement, whereas optimizing only the immediate distillation loss can progressively reduce the usefulness of subsequent on-policy supervision.

D.2 Scaling Number of Teachers

We investigate the scalability of GMOPD by increasing the number of specialized teachers from 6 to 20. The results on mathematical, code, and scientific reasoning are shown in Fig. 5, while instruction following, general reasoning, and tool use are reported in Fig. 6.

Summary of key findings. The performance of GMOPD improves consistently as additional teachers are introduced across all domains and model sizes. In contrast, the gains of the baselines remain limited and their performance is nearly unchanged or slightly decreases when the teacher set becomes more heterogeneous. This gap suggests that simply exposing a shared student to more teachers does not automatically translate into stronger merged capabilities: without controlling teacher-specific updates, the additional knowledge is offset by increasing cross-teacher interference. By learning shared low-rank capabilities and adapting only teacher-relevant ranks, GMOPD converts the growing teacher pool into complementary supervision. The gain is especially pronounced for larger students, which possess sufficient capacity to represent a broader collection of specialized reasoning patterns.

Taken together, these results demonstrate two complementary forms of scalability. Increasing the number of teachers strengthens GMOPD rather than amplifying interference, while increasing the student size further enlarges the benefit obtained from the same teacher pool. This behavior is consistent with the role of the adaptive rank gate: a larger low-rank representation provides additional reusable capability directions, and the teacher-specific gates selectively activate them without forcing every domain update through the same subspace.

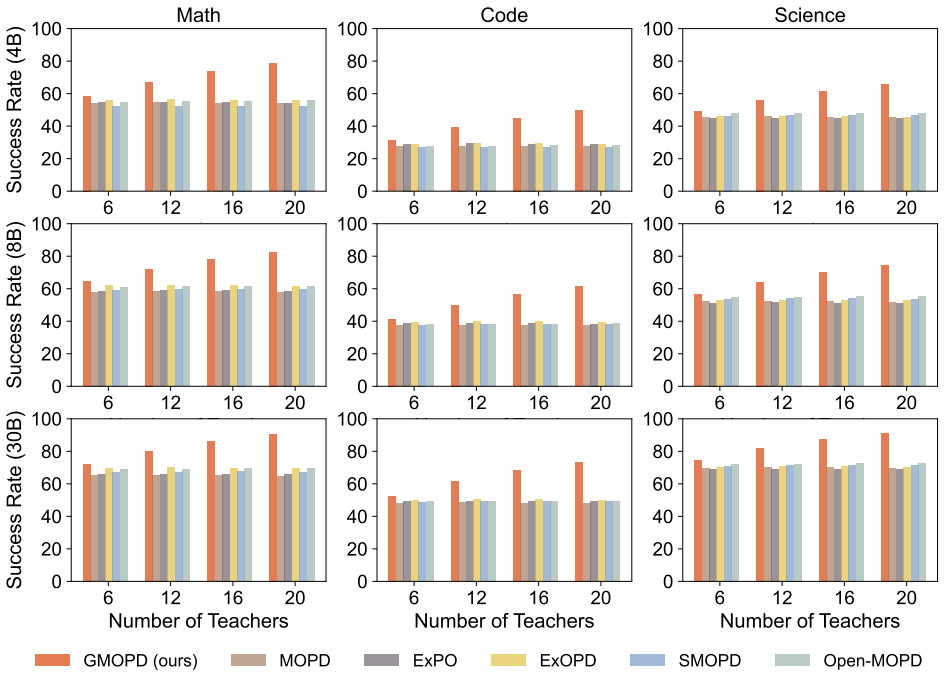


Figure 5: Teacher-count scaling on mathematical reasoning, code reasoning, and science reasoning. Each row corresponds to a different student scale, and each bar reports performance after integrating a domain-balanced subset of teachers. GMOPD obtains increasingly large gains as the teacher pool grows, whereas the baselines exhibit limited scaling.

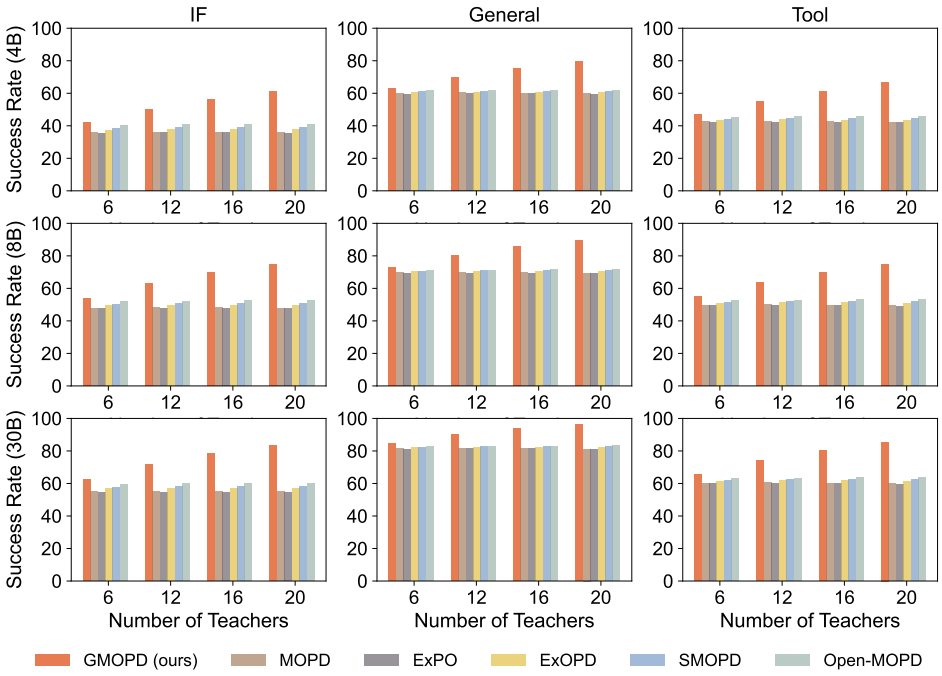


Figure 6: Teacher-count scaling on instruction following, general reasoning, and tool use. The advantage of GMOPD expands from 6 to 20 teachers across all three model scales, demonstrating that its scaling behavior extends beyond conventional reasoning benchmarks.

D.3 Effect of Teacher Capability

To further examine the teacher-strengthening paradox, we report the complete results obtained by using checkpoints from different RL stages as controlled proxies for teacher capability. As shown in Fig. 7, baseline methods generally benefit from improving teacher capability during the early stages, but their performance gradually saturates and frequently declines when distilling from stronger teachers. This trend appears consistently across model scales and is particularly pronounced in mathematics and tool use, where domain-specific teachers tend to develop more specialized reasoning patterns. Increasing the student size alleviates the degradation to some extent, but does not eliminate it, suggesting that additional model capacity alone cannot resolve cross-teacher interference. In contrast, GMOPD consistently achieves the best performance at every teacher capability level and continues to obtain stable gains as teacher capability increases. These results indicate that the meta objective preserves an initialization with sufficient overlap across teachers, while the adaptive rank gate prevents teacher-specific updates from excessively perturbing capabilities required by the remaining domains. Consequently, GMOPD can more reliably convert improvements in individual teachers into improvements of the merged student.

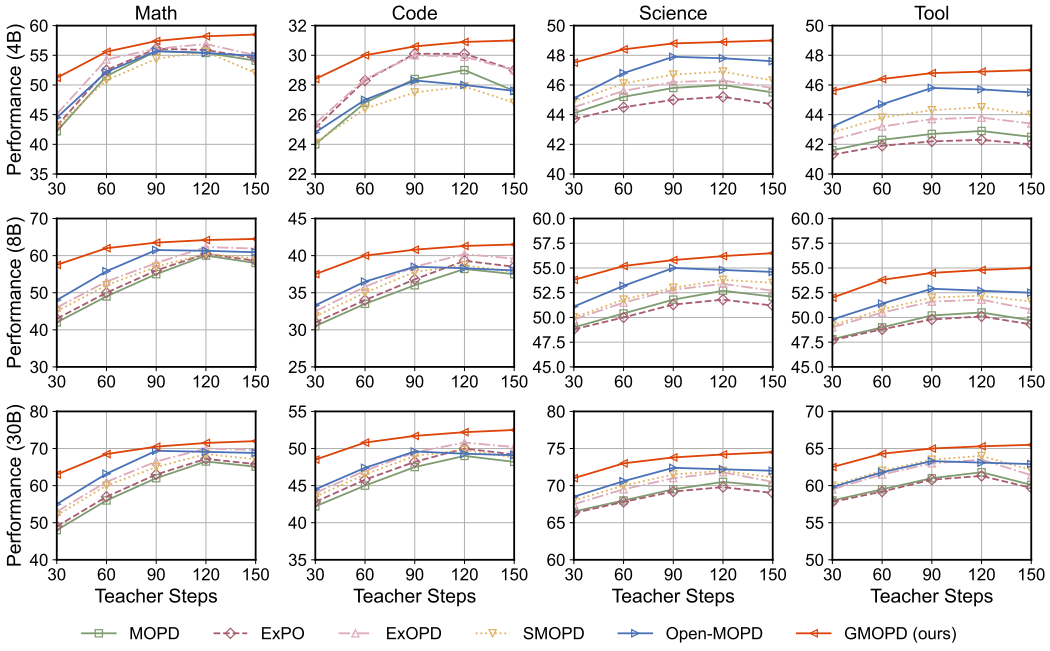


Figure 7: MOPD dynamics when checkpoints from different RL stages are used as controlled proxies for teacher capability.