

Counteraction-Aware Multi-Teacher On-Policy Distillation for General Capability Recovery with Domain Preservation

Tianlei Chen^{1,*‡}, Jiao Ou^{1,*†}, Ziyuan Liu¹, Ruiming Tang^{1,†}, Jian Liang¹, Han Li¹

¹Kuaishou Technology, Beijing, China

tianlei_chen@163.com, ojiao1111@gmail.com, tangruiming@kuaishou.com

Abstract

Domain specialization can improve LLM behavior, but often weakens the general capabilities inherited from the original model. Recent Multi-Teacher On-Policy Distillation (MOPD) pipelines recover model capabilities by supervising student-generated trajectories with teacher feedback, but typically assume teacher-aligned prompt coverage, requiring prompts to match the teachers’ training distributions. This assumption is difficult to satisfy when the general teacher is an open-source model whose post-training data are unknown. Instead of attempting to reconstruct this hidden distribution, we study general capability recovery with readily available proxy general prompts. We identify two failure modes of vanilla MOPD in this incomplete-coverage situation: recovery-preservation counteraction from mixing conflicting recovery and preservation gradients, and weak-signal flattening from uniformly averaging samples with unequal correction demand. We propose **Counteraction-Aware Multi-Teacher On-Policy Distillation (CaMOPD)**, which addresses these issues with decoupled alternating training and gap-based sample selection. CaMOPD allocates dedicated updates to general recovery, periodically performs domain-preservation updates, and selects samples with larger averaged token-level teacher-student log-probability gaps to concentrate correction signals. Across role-play dialogue and medical reasoning QA scenarios, CaMOPD outperforms all baselines in general capability recovery while maintaining domain-specific behavior. Gradient coherence analyses further support the intended effect of CaMOPD in producing more coherent correction signals.

1 Introduction

The remarkable success of large language models (LLMs) has accelerated their deployment across vertical domains, from role-play (Wang et al., 2025) to clinical decision support (Baichuan-M3 Team, 2026). However, domain specialization faces a fundamental paradox: the process of specialization itself often weakens the general capabilities that made these models valuable in the first place, creating a need for capability preservation (McCloskey and Cohen, 1989; Kumar et al., 2022). Classical mitigation strategies mix general-domain data or use replay mechanisms to reduce forgetting (Lopez-Paz and Ranzato, 2017; Rolnick et al., 2019; Dong et al., 2024; Huang et al., 2024). However, these approaches heavily rely on massive amounts of high-quality original data and introduce challenges in optimizing training schedules.

Recently, multi-teacher on-policy distillation (MOPD) has attracted attention as a way to recover diverse capabilities into a single student model by supervising student-generated trajectories with token-level teacher feedback (Yang et al., 2026; GLM-5-Team, 2026). However, existing MOPD pipelines often assume teacher-aligned prompt coverage: the prompts used during MOPD should sufficiently match the data distribution that produced the teachers (LLM-Core Xiaomi, 2026; Baichuan-M3 Team, 2026). This premise places a heavy burden on data collection and can be expensive, time-consuming, or impossible when teacher’s original post-training data are unavailable.

To avoid over-reliance on costly collection of the teacher’s original data, we consider using proxy prompts to recover general capability. Specifically, starting from an open-source general model, we post-train it into a vertical-domain model, after which MOPD is applied to recover its lost general capability: the original open-source model serves as the general teacher, and the specialized model

*Equal contribution.

†Corresponding authors.

‡Work completed during an internship at Kuaishou Technology.

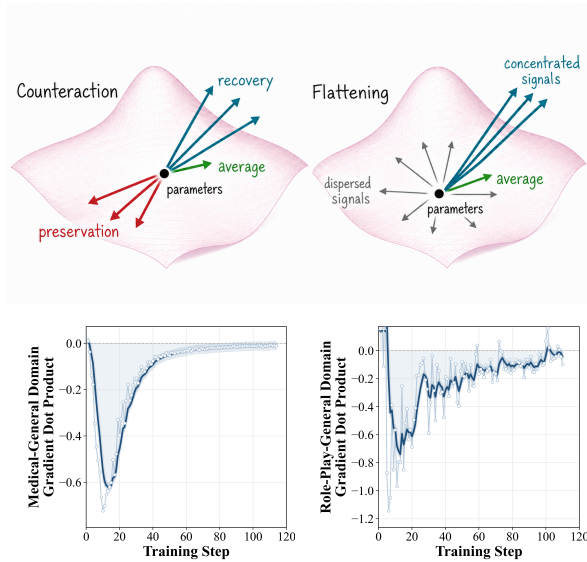


Figure 1: Conceptual overview and training-dynamics. **Top:** Two failure modes of MOPD under incomplete-coverage teacher feedback: recovery-preservation **counteraction** and weak-signal **flattening** within a feedback source. **Bottom:** Training dynamics of cross-domain gradient dot products for Medical Reasoning QA and Role-Play Dialogue. **The predominantly negative values empirically support the recovery-preservation counteraction failure mode.**

serves as both the student initialization and the domain teacher. In this situation, domain prompts are available from the specialization pipeline, while the general teacher’s original post-training data are unknown. Instead of attempting to reconstruct or cover this hidden distribution, we use readily available public general-domain prompts as **proxy prompts**, which may incompletely cover the open-source teacher’s true training distribution. Under this incomplete-coverage setting, we identify two MOPD failure modes, as illustrated in Figure 1: **recovery-preservation counteraction**, where gradients for general capability recovery and domain behavior preservation point in different directions and counteract when averaged in the same update, as reflected by predominantly negative cross-domain gradient dot products during training; and **weak-signal flattening**, where a fraction of gradients from different prompts under the same teacher are small and directionally dispersed, flattening the large-norm gradients that are most useful for training.

In this paper, we propose **Counteraction-Aware Multi-Teacher On-Policy Distillation (CaMOPD)** to address these two failure modes.

First, inspired by task-level alternating optimization in multi-task learning (Luong et al., 2016; Akbari et al., 2023), CaMOPD introduces **decoupled alternating training** to address recovery-preservation counteraction: by default, it performs three general-recovery steps followed by one domain-preservation step. This alternating schedule allocates dedicated updates to general recovery while periodically preserving domain behavior, thereby mitigating gradient counteraction between recovery and preservation. **Second**, CaMOPD uses **gap-based sample selection**. It selects samples with higher rankings in token-level teacher-student log-probability gaps. This concentrates updates on prompts with stronger correction demand, and our gradient coherence analysis shows that the selected subsets produce more aligned update directions. We validate CaMOPD in two stylistically different vertical-domain settings, role-play dialogue and medical reasoning QA. Evaluations show that, under the premise of preserving the acquired domain behavior, CaMOPD outperforms all baselines in general capability recovery. Furthermore, the success across such distinct scenarios verifies the broad applicability and generalization of our method.

Our contributions are summarized as follows:

- We identify two MOPD failure modes under proxy prompts: **recovery-preservation counteraction** and **weak-signal flattening**.
- We propose **CaMOPD**, integrating **decoupled alternating training** and **gap-based sample selection** to address gradient counteraction and concentrate high-demand correction signals.
- Evaluations across two **stylistically distinct** vertical domains validate our analysis, demonstrating the effectiveness of CaMOPD.

2 Related work

2.1 On-Policy Distillation

On-policy distillation (OPD) queries a teacher on student-generated outputs, making it a natural fit for LLM post-training on model-generated trajectories (Agarwal et al., 2024; Gu et al., 2024; Qwen Team, 2025; DeepSeek-AI, 2026). However, OPD signals are not uniformly useful across trajectories or tokens. One line of work studies signal imbalance, teacher-student compatibility, and noisy

or low-coherence teacher feedback, showing that OPD can introduce unreliable update directions (Li et al., 2026; Fu et al., 2026; Xu et al., 2025, 2026). Another line improves OPD by calibrating, relaxing, selecting, or reweighting local supervision at the token or sample level (Zheng et al., 2026; Ko et al., 2026; Zhang et al., 2026; Kim and Baek, 2026; Huang et al., 2025; Xie et al., 2026; Xu et al., 2026). Unlike these methods that refine local signals, CaMOPD introduces gap-based sample selection to explicitly improve cross-sample gradient coherence.

2.2 Multi-Teacher On-Policy Distillation

Multi-teacher OPD (MOPD) extends OPD by using multiple teachers as token-level guidance sources, and has been used in several recent LLM post-training pipelines. MiMo-V2-Flash and DeepSeek-V4 use MOPD-style distillation for base-model post-training and broad capability integration (LLM-Core Xiaomi, 2026; DeepSeek-AI, 2026). Baichuan-M3, Nemotron-Cascade 2, and GLM-5 apply related ideas to vertical-domain or staged post-training, with the latter two using cross-stage distillation to integrate or recover capabilities that degrade across training stages (Baichuan-M3 Team, 2026; Yang et al., 2026; GLM-5-Team, 2026). Unlike these pipelines that assume fully aligned teacher data, CaMOPD targets an incomplete-coverage setting with proxy prompts, focusing on reducing cross-domain gradient conflicts and elevating signal coherence.

2.3 Mitigating Catastrophic Forgetting

Capability degradation during specialization is closely related to catastrophic forgetting in continual adaptation (Luo et al., 2023). Existing mitigation strategies commonly mix domain data with general-domain or pretraining data (Dong et al., 2024; Béthune et al., 2025), or use replay mechanisms to revisit prior-task or synthesized examples (Lopez-Paz and Ranzato, 2017; Rolnick et al., 2019; Lesort, 2020; Huang et al., 2024). CaMOPD shares the capability-preservation objective of replay-based approaches, while organizing recovery and preservation through separate alternating updates.

3 Method

This section aims to address the general capability recovery after domain specialization problem under a practical incomplete-coverage setting. As

illustrated in Figure 2, Part A depicts the proxy-prompt training setup (detailed in Section 3.1); Part B illustrates the recovery-preservation counteraction and weak-signal flattening failure modes of Vanilla MOPD (analyzed in Section 3.2); and Part C presents our CaMOPD method, which resolves these issues via decoupled alternating training and gap-based sample selection (introduced in Section 3.3).

3.1 Preliminaries

Problem Setup. As illustrated in Figure 2 (Part A), we consider general capability recovery after domain specialization under an incomplete-coverage setting. Given a domain-specialized model, we aim to recover its general capabilities while preserving domain behaviors using proxy general prompts $\mathcal{D}_g$ and domain prompts $\mathcal{D}_d$. The original open-source model serves as the general teacher T_g , while the domain-specialized model serves as both the domain teacher T_d and the initialization for the target student policy π_θ .

Technical Formulation of MOPD. Under the multi-teacher on-policy distillation (MOPD) paradigm (LLM-Core Xiaomi, 2026), the training student policy π_θ is optimized on rollouts $y = (y_1, \dots, y_L)$ of length L generated by the inference policy μ_θ given prompt $x \sim \mathcal{D}_b$ ($b \in \{g, d\}$). Here, μ_θ denotes the inference policy obtained by loading the current training parameters θ into the inference engine. Let $h_t = (x, y_{<t})$. The token-level teacher feedback defines the advantage function $\hat{A}_{b,t}(\theta)$ as the teacher-student log-probability gap:

$$\hat{A}_{b,t}(\theta) = \text{sg} \left[\log \frac{\pi_{T_b}(y_t | h_t)}{\pi_\theta(y_t | h_t)} \right], \quad (1)$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation. The branch-wise Vanilla MOPD loss uses the training-inference importance weight $w_t(\theta)$, defined as $\text{sg}[\pi_\theta(y_t | h_t)/\mu_\theta(y_t | h_t)]$ when the ratio lies in $[\epsilon_{\text{low}}, \epsilon_{\text{high}}]$, and as 0 otherwise:

$$\mathcal{L}_{\text{MOPD}}^b(\theta) = -\mathbb{E}_{x \sim \mathcal{D}_b, y \sim \mu_\theta(\cdot|x)} \left[\frac{1}{L} \sum_{t=1}^L w_t(\theta) \hat{A}_{b,t}(\theta) \log \pi_\theta(y_t | h_t) \right], \quad (2)$$

3.2 Failure Modes of Vanilla MOPD

We identify two optimization failure modes that Vanilla MOPD can induce under the incomplete-coverage setting.

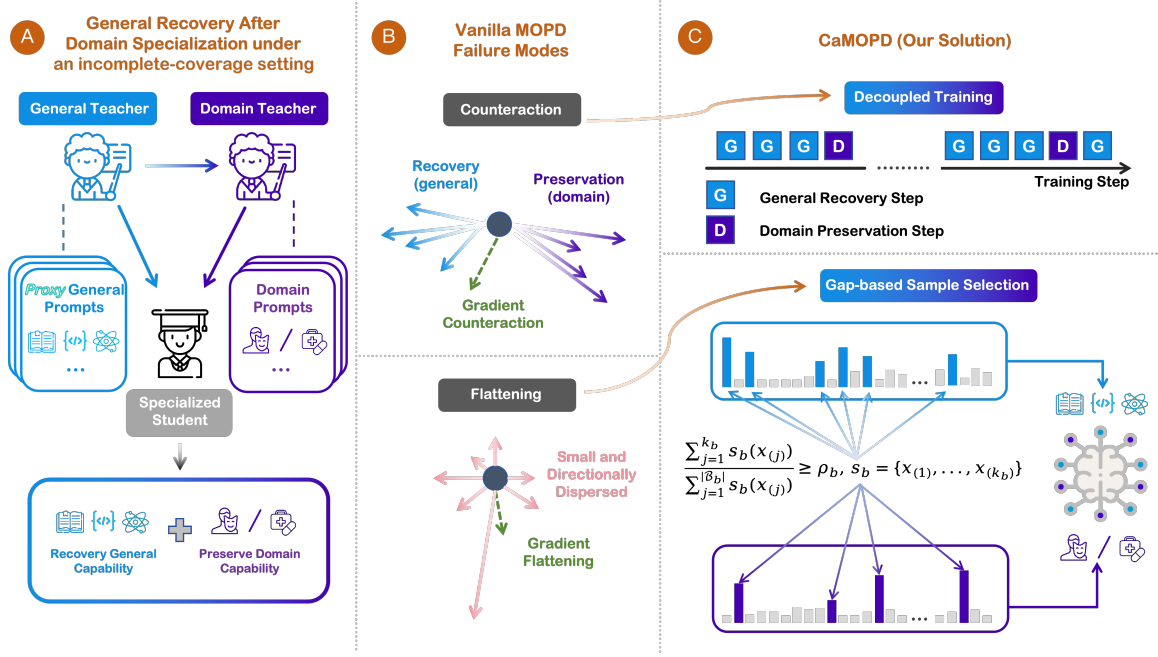


Figure 2: Overview of the general capability recovery after domain specialization problem, the Vanilla MOPD failure modes, and the CaMOPD design. **A:** General capability recovery after domain specialization under an **incomplete-coverage setting**. The original open-source model acts as the general teacher on proxy general prompts, while the specialized model serves as both the student and the domain teacher on domain prompts. **B:** Vanilla MOPD mixes both feedback sources in every update, which can cause recovery-preservation **counteraction** when general and domain gradients are misaligned, and weak-signal **flattening** when low-demand or dispersed samples dilute concentrated corrections within a branch. **C:** CaMOPD addresses these issues with decoupled alternating training, together with gap-based sample selection that concentrates each branch on samples with larger teacher-student correction demand.

Recovery-Preservation Counteraction. The first failure mode occurs between the two data streams. Vanilla MOPD unconditionally aggregates gradients from the general-prompt and domain-prompt branches during every parameter update. The training dynamics of Vanilla MOPD show that the general-recovery and domain-preservation gradients generally point in conflicting directions, as reflected by predominantly negative cross-domain gradient dot products in Figure 1. Consequently, part of the recovery direction is implicitly counteracted by the preservation direction within the same mixed update.

Weak-Signal Flattening. The second failure mode occurs within each feedback source. Full-batch averaging assigns equal aggregation weight to all sampled examples, allowing low-demand samples to dilute the contribution of high-demand samples. More critically, examples with weak effective learning signals often exhibit poor cross-sample gradient coherence, so their dispersed directions further weaken the coherent gradients that

are most useful for training.

3.3 Counteraction-Aware MOPD (CaMOPD)

To resolve the aforementioned bottlenecks, we introduce a scheduled and score-gated distillation strategy that decouples cross-domain aggregation and refines within-branch correction density while fully preserving the MOPD importance sampling core.

Decoupled Alternating Training. Inspired by task-level alternating optimization in multi-task learning (Luong et al., 2016; Akbari et al., 2023), CaMOPD addresses recovery-preservation counteraction by assigning the two signals to separate updates under a periodic training schedule:

$$\mathcal{S} = \underbrace{G, \dots, G}_{n_g}, D, \quad (3)$$

where n_g general recovery steps (G) are followed by a single domain preservation step (D). By updating only the active branch $b = \mathcal{S}(t)$ at each step, general and domain gradients are computed at

different parameter points, reducing direct mixed-update counteraction between recovery and preservation signals.

Role-specific Gap-based Sample Selection. To mitigate weak-signal flattening, CaMOPD selects samples based on their correction demand within the active branch. For each rollout y_i of length L_i , let $h_{i,t} = (x_i, y_{i,<t})$. We compute the initial token-level gap before the parameter update, $\Delta_{b,i,t} = \log \pi_{T_b}(y_{i,t} | h_{i,t}) - \log \pi_{\theta}(y_{i,t} | h_{i,t})$, as a label-free indicator.

General recovery uses the token-average absolute gap to measure deviation from the general teacher:

$$s_g(x_i) = \frac{1}{L_i} \sum_{t=1}^{L_i} |\Delta_{g,i,t}|. \quad (4)$$

Conversely, domain preservation uses the token-average positive gap to estimate preservation demand:

$$s_d(x_i) = \frac{1}{L_i} \sum_{t=1}^{L_i} [\Delta_{d,i,t}]_+. \quad (5)$$

Here, $[z]_+ = \max(z, 0)$ denotes the positive part of z . This asymmetric scoring reflects the different roles of the two branches: general recovery treats any large teacher-student deviation as a useful recovery signal, whereas domain preservation prioritizes tokens that the domain teacher assigns higher likelihood than the current student, avoiding over-penalization of already plausible domain-specific behaviors. Given a batch $\mathcal{B}_b$ from the active branch b and a gap-score mass target $\rho_b \in (0, 1]$, we sort samples by $s_b(x_i)$ in descending order, yielding $x_{(1)}, \dots, x_{(|\mathcal{B}_b|)}$. For $k \in \{1, \dots, |\mathcal{B}_b|\}$, we choose the smallest prefix length k_b that covers the target cumulative gap-score mass:

$$k_b = \min \left\{ k : \frac{\sum_{j=1}^k s_b(x_{(j)})}{\sum_{j=1}^{|\mathcal{B}_b|} s_b(x_{(j)})} \geq \rho_b \right\}. \quad (6)$$

The selected subset is then the corresponding prefix $S_b = \{x_{(1)}, \dots, x_{(k_b)}\}$, and the loss calculation is performed on this subset. To intensify the optimization drive on these prioritized instances, we introduce a branch-specific scaling ratio r_b (e.g., $r_g = 2, r_d = 1$) to scale their generated teacher feedback. For sample x_i , let $w_{i,t}(\theta)$ and $\hat{A}_{b,i,t}(\theta)$ denote the corresponding token-level importance weight and advantage, respectively. The

final CaMOPD objective for active branch b is formulated as:

$$\mathcal{L}_b^{\text{CaMOPD}}(\theta) = -\frac{1}{|S_b|} \sum_{x_i \in S_b} \left[\frac{1}{L_i} \sum_{t=1}^{L_i} w_{i,t}(\theta) r_b \hat{A}_{b,i,t}(\theta) \log \pi_{\theta}(y_{i,t} | h_{i,t}) \right]. \quad (7)$$

The complete end-to-end procedure is summarized in Algorithm 1.

Algorithm 1 Counteraction-Aware Multi-Teacher On-Policy Distillation (CaMOPD)

Require: Initial student π_{θ_0} , teachers $\{\pi_{T_g}, \pi_{T_d}\}$, data streams $\{\mathcal{D}_g, \mathcal{D}_d\}$, schedule $\mathcal{S}$, number of training steps N , learning rate η , gap-score mass targets $\{\rho_g, \rho_d\}$, recovery scale r_g

- 1: **for** training step $t = 0, \dots, N - 1$ **do**
- 2: Set active branch $b \leftarrow \mathcal{S}(t)$
- 3: Set branch scale $r_b \leftarrow r_g$ if $b = g$, otherwise $r_b \leftarrow 1$
- 4: Sample prompts $\mathcal{B}_b \sim \mathcal{D}_b$
- 5: Load θ_t into the inference engine to obtain μ_{θ_t}
- 6: Generate on-policy responses $y_i \sim \mu_{\theta_t}(\cdot | x_i)$ for $x_i \in \mathcal{B}_b$
- 7: Compute initial token gaps $\Delta_{b,i,\tau} = \log \pi_{T_b}(y_{i,\tau} | h_{i,\tau}) - \log \pi_{\theta_t}(y_{i,\tau} | h_{i,\tau})$
- 8: **if** $b = g$ **then**
- 9: Score each sample by $s_i \leftarrow \frac{1}{L_i} \sum_{\tau=1}^{L_i} |\Delta_{g,i,\tau}|$
- 10: **else**
- 11: Score each sample by $s_i \leftarrow \frac{1}{L_i} \sum_{\tau=1}^{L_i} [\Delta_{d,i,\tau}]_+$
- 12: **end if**
- 13: Sort samples by s_i in descending order
- 14: Select the smallest prefix length k_b covering ρ_b of cumulative gap-score mass and set $S_b = \{x_{(1)}, \dots, x_{(k_b)}\}$
- 15: Compute $\mathcal{L}_b^{\text{CaMOPD}}(\theta_t)$ on S_b using Eq. (7)
- 16: Update $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta} \mathcal{L}_b^{\text{CaMOPD}}(\theta_t)$
- 17: **end for**

4 Experiments

4.1 Experimental Setup

Role-Play Dialogue Instantiation. We use Qwen3-4B-Instruct-2507 as the base model and general teacher T_g . The role-play domain teacher T_d is obtained by supervised fine-tuning the same base model on CoSER (Wang et al., 2025). During MOPD, the general-recovery branch uses 10K Nemotron (Bercovich et al., 2025) proxy prompts, while the domain-preservation branch uses 10K held-out long-turn CoSER prompts.

Medical Reasoning QA Instantiation. We use Qwen3-8B as the base model and general teacher

Table 1: Results for the Role-Play dialogue instantiation on general capability benchmarks and Role-Play domain benchmarks. Higher is better for all metrics. Bold marks the best result among baselines. CaMOPD denotes our method.

Model	General capability										Role-Play domain capability				
	GPQA Diamond	Zebra Logic	HMMT 25	LCB v5	LCB v6	IF Eval	Arena HP	Arena CW	LiveBench 1125	Avg.	Story Cons.	Anthro.	Char. Fid.	Story Qual.	Avg.
General Teacher	61.49	79.50	33.12	34.77	20.57	83.18	36.50	65.80	65.60	53.39	34.74	23.09	23.89	41.98	30.92
Role-Play Teacher	54.29	31.50	26.67	25.09	16.00	67.65	9.70	4.20	42.80	30.88	42.27	33.42	38.75	49.89	41.08
Off-policy KD	60.35	68.50	28.54	30.47	23.43	75.97	23.60	11.20	57.00	42.12	41.25	32.15	38.23	49.61	40.31
Vanilla MOPD	61.99	71.30	29.17	31.54	22.29	79.67	25.50	13.70	58.90	43.78	43.57	34.91	41.45	51.50	42.86
Relaxed OPD	63.51	71.40	31.25	32.26	22.86	79.67	30.10	17.30	59.00	45.26	45.42	33.35	41.53	52.84	43.28
SeleTKD	60.98	70.60	31.04	34.41	24.00	78.93	27.10	16.90	58.90	44.76	43.87	32.28	39.46	51.54	41.79
CaMOPD	62.63	76.30	31.04	37.28	24.00	80.41	33.20	33.30	63.50	49.07	47.81	33.96	44.65	53.47	45.00

Table 2: Results for the Medical Reasoning QA instantiation on general capability benchmarks and Medical domain benchmarks. Higher is better for all metrics. Bold marks the best result among baselines. CaMOPD denotes our method.

Model	General capability										Medical domain capability		
	GPQA Diamond	Zebra Logic	HMMT 25	LCB v5	LCB v6	IF Eval	Arena HP	Arena CW	LiveBench 1125	Avg.	MedQA USMLE	MedXpertQA Text	Avg.
General Teacher	62.50	84.80	45.00	58.42	31.43	84.66	24.80	36.50	52.10	53.36	79.03	17.22	48.13
Medical Teacher	62.37	71.60	28.33	48.75	27.43	48.98	10.90	36.40	34.40	41.02	86.17	22.90	54.54
Off-policy KD	60.10	70.30	34.38	50.54	24.57	48.61	13.20	32.60	41.70	41.78	85.23	22.41	53.82
Vanilla MOPD	60.10	72.60	33.96	51.25	24.00	48.06	14.80	33.00	42.80	42.29	86.02	23.55	54.79
Relaxed OPD	61.49	73.20	32.92	53.76	29.14	48.61	15.40	34.90	40.70	43.35	85.94	24.12	55.03
SeleTKD	58.46	72.10	33.54	54.48	29.71	49.17	15.20	35.60	41.80	43.34	85.78	23.51	54.65
CaMOPD	61.99	72.60	34.58	56.63	29.71	59.89	16.40	31.80	49.00	45.84	85.78	23.63	54.71

T_g . The domain-specialized initialization and domain teacher T_d are both Intelligent-Internet/II-Medical-8B (Intelligent Internet, 2025), an open-source medical reasoning model post-trained from Qwen3-8B with medical SFT and DAPO RL (Yu et al., 2025). During MOPD, the general-recovery branch uses 10K Nemotron proxy prompts, and the domain-preservation branch uses 10K prompt-only medical examples. More data and training details are provided in Appendix A.

Evaluation. For general capability, we evaluate knowledge and reasoning with GPQA-Diamond (Rein et al., 2024), ZebraLogic (Lin et al., 2025), and HMMT25 from MathArena (Balunović et al., 2025); coding with LiveCodeBench v5 and LiveCodeBench v6 (Jain et al., 2025); instruction following with IF-Eval (Zhou et al., 2023); preference-based generation with the Hard Prompt and Creative Writing subsets of Arena-Hard v2 (Li et al., 2025); and aggregate live evaluation with LiveBench 1125 (White et al., 2025). For role-play capability, following CoSER (Wang et al., 2025), we evaluate Storyline Consistency, Anthropomorphism, Character Fidelity, and Storyline Quality. For medical reasoning QA capability, we evaluate MedXpertQA Text and MedQA-USMLE 4-option

test accuracy. Appendix B details the scoring protocol for the general and medical reasoning QA benchmarks.

4.2 Baselines

For each domain, we compare Vanilla MOPD, CaMOPD, and two OPD methods reproduced under the Vanilla MOPD training mechanism with the same student initialization, teachers, data streams, batch size, rollout budget, optimizer settings, and training steps. We also include Off-policy KD, which applies supervised fine-tuning to mixed responses generated by the general and domain teachers. Relaxed OPD (Ko et al., 2026) interprets the teacher-student log-likelihood ratio as a token reward and relaxes strict imitation with reward clipping and entropy-based token sampling. SeleTKD (Huang et al., 2025) uses a propose-and-verify token selection rule, applying full distillation loss to accepted tokens while masking or down-weighting rejected tokens.

4.3 Main Results

Tables 1 and 2 report the main comparisons for role-play and medical reasoning QA specialization on both general benchmarks and domain-specific

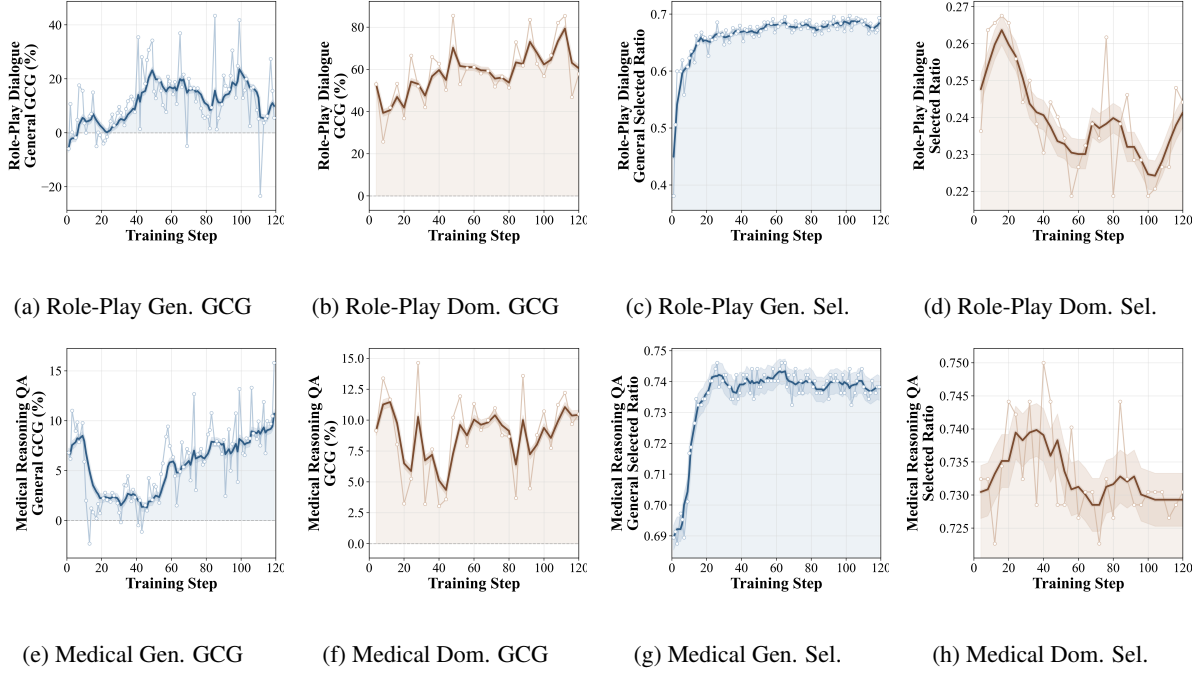


Figure 3: Training dynamics and gap-based sample selection analysis. Panels (a)–(d) show Role-Play dialogue, and panels (e)–(h) show Medical Reasoning QA. Within each row, the four panels report general-branch GCG, domain-branch GCG, general-branch selected fraction, and domain-branch selected fraction, respectively. GCG denotes gradient coherence gain of the selected subset over the full candidate batch, and selected-fraction plots report the retained fraction under the same 80% gap-score mass target.

evaluations. Across these two stylistically different instantiations, CaMOPD achieves strong general-capability recovery while maintaining the specialized domain behavior, demonstrating the effectiveness and generality of our counteraction-aware design.

General Capability Recovery. CaMOPD consistently improves recovery over baselines on the general benchmarks. In the Role-Play dialogue instantiation, CaMOPD raises the general average from the strongest baseline score of 45.26 to 49.07. At the individual-benchmark level, CaMOPD achieves the best or tied-best results on seven of nine general benchmarks. For example, relative to the strongest baseline on each benchmark, it improves ArenaHard v2 Creative Writing from 17.30 to 33.30, ZebraLogic from 71.40 to 76.30, and LiveBench 1125 from 59.00 to 63.50. In the Medical Reasoning QA instantiation, CaMOPD raises the general average from the strongest baseline score of 43.35 to 45.84. At the individual-benchmark level, CaMOPD achieves the best or tied-best results on seven of nine general benchmarks. For example, relative to the strongest baseline on each benchmark, it improves IF-Eval from 49.17 to 59.89,

LiveBench 1125 from 42.80 to 49.00, and LiveCodeBench v5 from 54.48 to 56.63. These results demonstrate the effectiveness of CaMOPD across both specialization settings.

Domain Capability Preservation. MOPD-based training preserves the specialized domain behavior across both instantiations. Our method and the MOPD-based baselines all remain close to or above the corresponding domain teachers on the domain metrics, indicating that the MOPD training mechanism itself is sufficient to protect the acquired domain capability.

4.4 Training Dynamics and Selection Analysis

We analyze retained sample fractions and within-branch gradient coherence to characterize how gap-based sample selection concentrates updates on samples with larger correction demand and reduces within-branch signal dilution.

Selection Ratio Under The Same Mass Target.

Figures 3c, 3d, 3g, and 3h report the retained sample fractions under the same 80% gap-score mass target. The general branch requires broad coverage in both domains. The domain branch is sparse in role-play dialogue but remains broader in medical

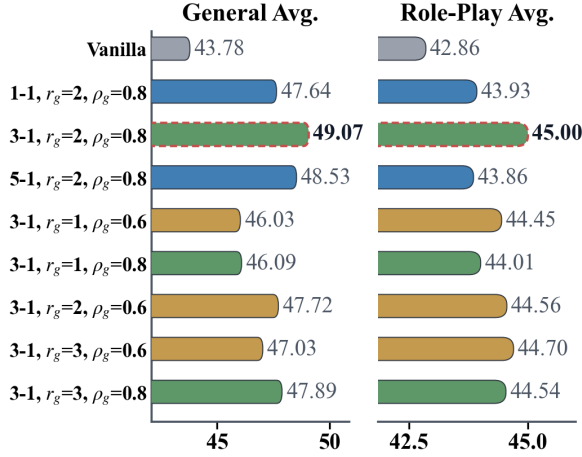


Figure 4: Bar-chart summary of the hyperparameter study. General Avg. is the arithmetic mean over the nine general benchmarks in Table 7; Role-Play Avg. uses the domain-average scores.

reasoning QA, suggesting that domain correction demand can vary substantially across specialized tasks.

Within-branch Gradient Coherence. To examine whether this selection is associated with reduced weak-signal flattening, we measure the within-branch gradient coherence of the parameter updates. For a sample set A , we define

$$\text{Coh}(A) = \frac{\|\sum_{i \in A} g_i\|}{\sum_{i \in A} \|g_i\|},$$

where g_i is the parameter gradient of sample i . Higher values indicate more aligned gradient directions. We report gradient coherence gain (GCG) as the relative improvement of the selected subset over the full candidate batch:

$$\text{GCG} = \frac{\text{Coh}(A_{\text{selected}}) - \text{Coh}(A_{\text{full}})}{\text{Coh}(A_{\text{full}})} \times 100\%.$$

Figures 3a, 3b, 3e, and 3f plot the GCG of the selected subset over the full candidate batch. GCG is predominantly positive across both specialization settings, with especially large gains in the role-play domain branch and smaller but consistent gains in medical reasoning QA. Together with the selection-ratio analysis, this suggests that gap-based sample selection reduces weak-signal flattening and produces more coherent update directions across both specialization settings.

4.5 Hyperparameter Analysis

We further examine the recovery-to-preservation schedule, the recovery scaling ratio r_g , and the gap-score mass target ρ_g in the role-play setting. As

Table 3: Matched ablations across Role-Play dialogue and Medical Reasoning QA. Gen. Avg. is the arithmetic mean over the nine general benchmarks, and Dom. Avg. is the corresponding domain average. Higher is better, and bold marks the best result.

Variant	Role-Play Dialogue		Medical Reasoning QA	
	Gen. Avg.	Dom. Avg.	Gen. Avg.	Dom. Avg.
CaMOPD	49.07	45.00	45.84	54.71
w/o selection	47.21	43.46	44.94	54.61
w/o selection and scaling	46.51	42.60	44.63	54.59
w/o selection, scaling, and decoupling	46.10	42.00	43.19	54.55

summarized in Figure 4, the 3-1 schedule gives the best recovery-preservation trade-off; the 5-1 schedule achieves similar general recovery, but domain performance starts to decline. For recovery scaling and gap-based sample selection, $r_g = 2$ and $\rho_g = 0.8$ gives the strongest aggregate general recovery, combining broader sample coverage with amplified gradients on selected high-demand samples. We therefore use the 3-1 schedule with $r_g = 2$ and $\rho_g = 0.8$ in the main experiments. The full metric table and detailed analysis are provided in Appendix C.

4.6 Ablation Study

Table 3 presents a matched ablation chain that successively isolates gap-based sample selection, recovery scaling, and decoupled alternating training across both specialization settings. Each adjacent pair changes one component while keeping the remaining training settings fixed. *CaMOPD* drops the low-gap tail under a 3-1 alternating schedule and applies $r_g = 2$ to the high-gap samples; *w/o selection* follows the same schedule, applies $r_g = 2$ to the high-gap samples, and leaves the low-gap samples unscaled; *w/o selection and scaling* uses all samples without recovery scaling; and *w/o selection, scaling, and decoupling* uses mixed updates with a 3:1 ratio of general to domain samples. Across both settings, the general and domain averages decrease at each successive ablation. The full per-metric results are provided in Appendix D.

5 Conclusion

We studied general capability recovery after domain specialization under an incomplete-coverage setting, where only proxy general prompts are available for the general teacher. We identified two failure modes of Vanilla MOPD: recovery-preservation counteraction and weak-signal flat-

tening. To address them, we proposed CaMOPD, which combines alternating recovery/preservation updates with gap-based sample selection. Experiments on role-play dialogue and medical reasoning QA show that CaMOPD improves general recovery over baselines while maintaining domain capability. Gradient and selection analyses further suggest that CaMOPD reduces mixed-update counteraction and produces more coherent correction signals.

Limitations

Our experiments instantiate CaMOPD with 4B and 8B models in the role-play and medical reasoning QA domains, respectively. Due to computational constraints, we do not evaluate larger model scales in this work; extending CaMOPD to larger specialized models is an important direction for future exploration. In addition, we use a fixed 32K response-length budget for on-policy rollouts. This setting is chosen to avoid artificially limiting the output capacity of long-reasoning models during distillation, but it does not study how different rollout length budgets affect the recovery-preservation trade-off. Future work can examine CaMOPD under shorter and adaptive rollout budgets.

Ethical Considerations

This work studies post-training methods for recovering general capabilities while preserving domain-specific behavior. In the medical reasoning QA setting, the models and evaluations are used only for research on medical question answering and educational reasoning benchmarks. The resulting systems should not be interpreted as providing professional medical advice, diagnosis, or treatment recommendations. Any deployment in clinical or patient-facing settings would require substantially stronger validation, expert oversight, safety guardrails, and compliance with applicable medical regulations. Our training and evaluation use publicly available datasets and open-source models. We do not use private patient records or personally identifiable medical information.

Acknowledgments

We thank the reviewers for their constructive feedback. We also thank our colleagues on Kuaishou’s infrastructure team for their support with the training and evaluation infrastructure.

References

- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. [On-policy distillation of language models: Learning from self-generated mistakes](#). In *International Conference on Learning Representations*, pages 21246–21263.
- Hassan Akbari, Dan Kondratyuk, Yin Cui, Rachel Hornung, Huisheng Wang, and Hartwig Adam. 2023. [Alternating gradient descent and mixture-of-experts for integrated multimodal perception](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Baichuan-M3 Team. 2026. [Baichuan-M3: Modeling clinical inquiry for reliable medical decision-making](#). *arXiv preprint arXiv:2602.06570*.
- Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. [MathArena: Evaluating LLMs on uncontaminated math competitions](#). In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.
- Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shafaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, and 117 others. 2025. [Llama-Nemotron: Efficient reasoning models](#). *arXiv preprint arXiv:2505.00949*.
- Louis Béthune, David Grangier, Dan Busbridge, Eleonora Gualdoni, Marco Cuturi, and Pierre Ablin. 2025. [Scaling laws for forgetting during finetuning with pretraining data injection](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 4020–4042. PMLR.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. [HuatuoGPT-o1, towards medical complex reasoning with LLMs](#). *arXiv preprint arXiv:2412.18925*.
- DeepSeek-AI. 2026. [DeepSeek-V4: Towards highly efficient million-token context intelligence](#). *arXiv preprint arXiv:2606.19348*.
- Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. [How abilities in large language models are affected by supervised fine-tuning data composition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–198, Bangkok, Thailand. Association for Computational Linguistics.
- Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. 2026. [Revisiting on-policy distillation: Empirical failure modes and simple fixes](#). *arXiv preprint arXiv:2603.25562*.

- GLM-5-Team. 2026. [GLM-5: from vibe coding to agentic engineering](#). *arXiv preprint arXiv:2602.15763*.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. [MiniLLM: Knowledge distillation of large language models](#). In *International Conference on Learning Representations*.
- Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. 2025. [SelectKD: Selective token-weighted knowledge distillation for LLMs](#). *arXiv preprint arXiv:2510.24021*.
- Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024. [Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1416–1428. Association for Computational Linguistics.
- Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2026. [m1: Unleash the potential of test-time scaling for medical reasoning with large language models](#). In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pages 369–383. PMLR.
- Intelligent Internet. 2025. [II-Medical-8B: Medical reasoning model](#). <https://huggingface.co/Intelligent-Internet/II-Medical-8B>. Model card.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. [LiveCodeBench: Holistic and contamination free evaluation of large language models for code](#). In *International Conference on Learning Representations*, pages 58791–58831.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. [What disease does this patient have? a large-scale open domain question answering dataset from medical exams](#). *Applied Sciences*, 11(14):6421.
- Minsang Kim and Seung Baek. 2026. [Explain in your own words: Improving reasoning via token-selective dual knowledge distillation](#). In *International Conference on Learning Representations*, pages 103174–103196.
- Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. 2026. [Scaling reasoning efficiently via relaxed on-policy distillation](#). *arXiv preprint arXiv:2603.11137*.
- Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. [Fine-tuning can distort pretrained features and underperform out-of-distribution](#). In *International Conference on Learning Representations*.
- Timothée Lesort. 2020. *Continual Learning: Tackling Catastrophic Forgetting in Deep Neural Networks with Replay Processes*. Ph.D. thesis, Institut Polytechnique de Paris.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025. [From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 34209–34231. PMLR.
- Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. 2026. [Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe](#). *arXiv preprint arXiv:2604.13016*.
- Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. [ZebraLogic: On the scaling limits of LLMs for logical reasoning](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 37889–37905. PMLR.
- LLM-Core Xiaomi. 2026. [MiMo-V2-Flash technical report](#). *arXiv preprint arXiv:2601.02780*.
- David Lopez-Paz and Marc’Aurelio Ranzato. 2017. [Gradient episodic memory for continual learning](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. [An empirical study of catastrophic forgetting in large language models during continual fine-tuning](#). *arXiv preprint arXiv:2308.08747*.
- Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. [Multi-task sequence to sequence learning](#). In *International Conference on Learning Representations*.
- Michael McCloskey and Neal J. Cohen. 1989. [Catastrophic interference in connectionist networks: The sequential learning problem](#). In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. [MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering](#). In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.
- Qwen Team. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.

- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. [GPQA: A graduate-level google-proof Q&A benchmark](#). In *First Conference on Language Modeling*.
- David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. [Experience replay for continual learning](#). In *Advances in Neural Information Processing Systems*, volume 32.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. [HybridFlow: A flexible and efficient RLHF framework](#). In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.
- Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025. [CoSER: Coordinating LLM-based persona simulation of established roles](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 64822–64858. PMLR.
- Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddhartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. [LiveBench: A challenging, contamination-limited LLM benchmark](#). In *International Conference on Learning Representations*, pages 91595–91631.
- Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. [MedReason: Eliciting factual medical reasoning steps in LLMs via knowledge graphs](#). *arXiv preprint arXiv:2504.00993*.
- Xurong Xie, Zhucun Xue, Jiafu Wu, Jian Li, Yabiao Wang, Xiaobin Hu, Yong Liu, and Jiangning Zhang. 2026. [LLM-oriented token-adaptive knowledge distillation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34070–34078.
- Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. 2025. [Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling](#). In *International Conference on Learning Representations*.
- Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, and Zhipeng Wang. 2026. [PACED: Distillation and on-policy self-distillation at the frontier of student competence](#). *arXiv preprint arXiv:2603.11178*.
- Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, Renjie Pi, Grace Lam, Nayeon Lee, Alexander Bukharin, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2026. [Nemotron-Cascade 2: Post-training LLMs with cascade RL and multi-domain on-policy distillation](#). *arXiv preprint arXiv:2603.19220*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. [DAPO: An open-source LLM reinforcement learning system at scale](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 113222–113244. Curran Associates, Inc.
- Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler, II, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. 2026. [Fast and effective on-policy distillation from reasoning prefixes](#). *arXiv preprint arXiv:2602.15260*.
- Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. [SWIFT: A scalable lightweight infrastructure for fine-tuning](#). *arXiv preprint arXiv:2408.05517*.
- Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. 2026. [SCOPE: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting](#). *arXiv preprint arXiv:2604.10688*.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *arXiv preprint arXiv:2311.07911*.
- Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. [MedXpertQA: Benchmarking expert-level medical reasoning and understanding](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80961–80990. PMLR.

A Dataset Appendix: Dataset Details and Processing

A.1 Training Sources

For the role-play experiments, we start from Qwen3-4B-Instruct-2507 (Qwen Team, 2025). We use two teachers during MOPD. The general teacher T_g is the original Qwen3-4B-Instruct-2507 model, and the role-play teacher T_d is obtained by supervised fine-tuning the same base model on CoSER. CoSER (Wang et al., 2025)¹ is a role-play dataset built from 771 novels and provides multi-turn, multi-character conversations with character profiles, dialogue context, plot summaries, and character experiences. The official dataset card reports a formatted SFT file, `sft_conversations_sharegpt.json`, together with a held-out test file and full book-level JSON files.

For the medical reasoning QA experiments, we instantiate the same recovery-under-preservation setup with a publicly released medical post-trained model. The general teacher is Qwen/Qwen3-8B (Qwen Team, 2025), and the domain-specialized initialization and domain teacher are both Intelligent-Internet/II-Medical-8B (Intelligent Internet, 2025)². The model card lists II-Medical-8B under the Apache-2.0 license and describes it as a medical reasoning model obtained by applying medical supervised fine-tuning and DAPO RL optimization (Yu et al., 2025) on top of Qwen/Qwen3-8B. Accordingly, II-Medical-8B provides the domain-specialized initialization and teacher in this instantiation.

For general-domain recovery, we use Nemotron-Post-Training-Dataset-v1 (Bercovich et al., 2025)³, a large SFT corpus released with five splits: chat, code, math, stem, and tool calling. The official card reports 25,659,642 examples in total, with 746,622 chat examples, 1,896,395 code examples, 2,044,407 math examples, 20,662,167 stem examples, and 310,051 tool-calling examples. We sample 10K general prompts proportionally from these categories.

¹<https://huggingface.co/datasets/Neph0s/CoSER>

²<https://huggingface.co/Intelligent-Internet/II-Medical-8B>

³<https://huggingface.co/datasets/nvidia/Nemotron-Post-Training-Dataset-v1>

A.2 CoSER Split for Specialization and Domain Preservation

We first rank CoSER conversations by the number of dialogue turns. The 10K conversations with the largest number of turns are reserved as unseen role-play domain data for MOPD. For each reserved conversation, we use all but the final turn as the prompt/context and leave the final continuation to be generated on-policy by the current student. These examples are never used in role-play SFT. The remaining 300K CoSER conversations are used to fine-tune Qwen3-4B-Instruct-2507 and produce the role-play teacher.

Using held-out prompts ensures that domain-preservation updates are computed on contexts excluded from role-play SFT. Long-turn conversations require the model to maintain persona, dialogue state, and storyline consistency over rich contexts, providing a challenging preservation setting.

A.3 Nemotron Sampling for General Recovery

Because the pretraining and post-training data of the open-source general teacher are unknown, we use Nemotron as a proxy for the general distribution. We sample 10K prompts from its 25.66M examples while preserving the official category proportions, yielding the incomplete-coverage recovery setting studied in our experiments.

A.4 Medical Reasoning QA Model and Domain-Preservation Prompts

The medical reasoning QA instantiation uses II-Medical-8B as both the starting student and the preservation teacher. This mirrors the role-play setup at the level of the recovery objective: CaMOPD starts from a domain-specialized model, recovers general capability from the corresponding Qwen3-8B teacher, and periodically performs domain-preservation updates on medical reasoning QA prompts to reduce domain drift.

The medical reasoning QA domain-preservation stream is prompt-only. It does not use gold answers during MOPD/CaMOPD training: the current student first generates on-policy continuations, and the domain teacher provides token-level log-probability feedback on those student trajectories. We construct the preservation prompts from medical question-answering and reasoning sources related to the II-Medical data ecosystem: 4K prompts

Nemotron general prompt example.

category: math/code/stem/chat/tool_calling
messages:
 [user] a task prompt, such as a math problem, programming challenge, STEM question, chat request, or tool-use instruction.
 [assistant] a synthetic reference response generated by public/open models.
metadata: source information and, for tool-calling data, tool schema or call metadata.

Figure 5: Nemotron general prompt structure used for the general recovery branch.

CoSER Role-Play conversation example.

conversations:
 [system] role instruction and character/background profile, e.g., “Your role is to be [character]” plus **scenario context**.
 [user/character] previous dialogue turns and actions.
 [assistant/character] the next in-character response.
auxiliary fields: character profiles, scenario, topic, plot summary, speaking characters, and book-level metadata in the full CoSER files.

Figure 6: CoSER Role-Play conversation structure used for Role-Play specialization and domain preservation.

from MedMCQA (Pal et al., 2022)⁴, 3K from MedReason (Wu et al., 2025)⁵, 2K English prompts from Medical-R1-Distill-Data (Chen et al., 2024)⁶, and 1K from m23k-tokenized (Huang et al., 2026)⁷. After selection and formatting, the resulting preservation set contains 10K prompts.

A.5 Data Examples

Figures 5–7 show the field structure and training usage for the general, role-play, and medical reasoning QA streams.

A.6 Resulting Data Roles

Table 4 summarizes how each data source is used in the training pipeline. We separate specialization data, domain-preservation prompts, and proxy general-recovery prompts to avoid mixing their roles in the MOPD/CaMOPD setup.

⁴<https://huggingface.co/datasets/openlifescienceai/medmcqa>

⁵<https://huggingface.co/datasets/UCSC-VLAA/MedReason>

⁶<https://huggingface.co/datasets/FreedomIntelligence/Medical-R1-Distill-Data>

⁷<https://huggingface.co/datasets/UCSC-VLAA/m23k-tokenized>

Medical domain-preservation prompt template.

raw fields: question, optional options or opa/opb/opc/opd
prompt:
 [system] You are a careful medical reasoning assistant. Provide educational medical information, mention uncertainty when relevant, and recommend professional medical care for diagnosis or treatment decisions.
 [user] {question}
 A. {opa}
 B. {opb}
 C. {opc}
 D. {opd}
 Please reason step by step and give the final answer choice at the end.
open-ended variant: for sources with only question or prompt/text, the user message is {question} followed by “Please reason step by step and answer clearly. Do not present the answer as a substitute for professional medical advice.”
data_type: medical
training use: gold answers and explanations are not placed in the prompt field; teacher feedback is computed on student-generated continuations.

Figure 7: Medical domain-preservation prompt template constructed from source dataset fields.

A.7 Training Protocol

We perform SFT using the ms-swift framework (Zhao et al., 2024) and implement the MOPD training pipeline using verl (Sheng et al., 2025).

All MOPD-based methods generate one response per prompt with the current student and compute token-level teacher feedback on student trajectories. We train for 120 steps with a prompt batch size of 512 and a learning rate of 2×10^{-6} . Prompts are truncated at 4096 tokens and rollouts are capped at 32768 tokens. Unless otherwise specified, CaMOPD uses a 3-1 recovery-to-preservation schedule, gap-score mass targets of $\rho_g = \rho_d = 0.8$, and a general recovery scale of $r_g = 2$, with the domain scale fixed at $r_d = 1$. We run all training on a 4-node cluster with 8 NVIDIA H100 GPUs per node. Each run uses 32 H100 GPUs for approximately 7 hours, corresponding to about 224 total GPU hours.

Gradient Computation. For the analyses reported in Figures 1 and 3, we compute gradients with respect to all model parameters at each analyzed training step and aggregate the resulting statistics across parameter shards under tensor and pipeline parallelism.

Table 4: Dataset roles in CaMOPD. Domain-preservation prompts provide preservation gradients for the active vertical domain, while the general branch uses a compact proportional sample from Nemotron to recover general capability under a limited MOPD budget.

Data	Size	Used for	Notes
CoSER remainder	300K	Role-Play SFT teacher	Trains T_d
CoSER long-turn heldout	10K	Role-Play domain preservation	Unseen by Role-Play SFT
Medical prompt mixture	10K	Medical domain preservation	Prompt-only; no labels used
Nemotron proportional sample	10K	General recovery	Proxy general prompts

Table 5: Evaluation settings and scoring protocols for the general benchmarks. T denotes the evaluated model’s generation temperature, n denotes the number of generated samples, and k denotes the pass@ k aggregation parameter. LLM-based verification is used only for objective answer extraction and equivalence checking; preference-based generation benchmarks use GPT-4.1 as a pairwise judge.

Benchmark	Generation setting	Scoring/aggregation	Judge or checker
GPQA-Diamond	0-shot; $T=0.7$; $n=4$ top_p=0.95; top_k=20 rep.=1.05; max 32K tokens	pass@1/accuracy estimate with $k=1$	Qwen2.5-14B-Instruct verifier; $T=0.2$
HMMT25	$T=0.6$; $n=16$ top_p=0.95; top_k=20 rep.=1.05; max 32K tokens	pass@1/accuracy estimate with $k=1$	Qwen2.5-14B-Instruct verifier; $T=0.2$
ZebraLogic	$T=0$; $n=1$ max 32K tokens	Puzzle/grid correctness	ZebraLogic rule-based grid evaluator
LiveCodeBench v5	0-shot; $T=0$; $n=1$ max 32K tokens	Code pass rate	Official execution-based evalu- ator
LiveCodeBench v6	0-shot; $T=0$; $n=1$ max 32K tokens	Code pass rate	Official execution-based evalu- ator
IF-Eval	$T=0$; $n=1$ max 16K tokens	Instruction-following score	Official deterministic rule checker
Arena-Hard v2 (HP)	$T=0$; $n=1$ max 32K tokens	Pairwise preference score	GPT-4.1 judge via Arena-Hard pipeline; $T=0.0$
Arena-Hard v2 (CW)	$T=0$; $n=1$ max 32K tokens	Pairwise preference score	GPT-4.1 judge via Arena-Hard pipeline; $T=0.0$
LiveBench 1125	$T=0$; $n=1$ max 32K tokens	Official mixed benchmark score	Official exact, rule-based, sym- bolic, and execution scorers

B Evaluation Protocols

This appendix describes how the general-capability, role-play, and medical reasoning QA benchmarks are scored. For general capability, we use GPQA-Diamond (Rein et al., 2024), HMMT25 from MathArena (Balunović et al., 2025), ZebraLogic (Lin et al., 2025), LiveCodeBench v5 and LiveCodeBench v6 (Jain et al., 2025), IF-Eval (Zhou et al., 2023), the Hard Prompt and Creative Writing subsets of Arena-Hard v2 (Li et al., 2025), and LiveBench 1125 (White et al., 2025). Table 5 summarizes the generation settings, aggregation rules, and judges or checkers used for these general benchmarks. We distinguish three evaluation modes. First, objective benchmarks with short canonical answers are scored by extracting the model’s final answer and comparing it with the reference answer. Second, code and instruction-

following benchmarks are evaluated by their official execution or rule-based checkers. Third, preference-based generation benchmarks use an LLM judge because the target quality criteria are preference-based rather than exact-answer based.

Answer Verification for Objective Benchmarks.

For GPQA-Diamond and HMMT25, we use an LLM-based answer verifier implemented with Qwen2.5-14B-Instruct. The verifier is given the original prompt, the model response, and the ground-truth answer. It is instructed to identify the final answer in the response and output both an extracted answer and a binary decision indicating whether it matches the reference. Correctness is determined by this binary equivalence decision, with Qwen2.5-14B-Instruct serving as the answer extractor and equivalence verifier.

Concretely, the verifier prompt wraps each

example as `<prompt>`, `<response>`, and `<ground_truth_answer>` fields and asks for `<extracted_answer>` and `<decision>` tags. A prediction is counted as correct if and only if the parsed decision is yes. For multi-sample settings, we generate n candidate responses and report the $k = 1$ pass@1/accuracy estimate obtained from their binary decisions.

On a validation subset of 100 model outputs, with 50 outputs from each benchmark, the verifier decisions agreed with the official rule-based extraction for all examples, supporting its reliability for answer extraction and equivalence checking.

Rule-based and Execution-based Benchmarks. ZebraLogic is evaluated with the ZebraLogic grid evaluator, which parses the predicted grid and compares it against the puzzle solution. LiveCodeBench v5 and v6 are evaluated with the benchmark execution pipeline: generated programs are extracted and run against the associated tests, and pass rates are reported following the benchmark protocol. IF-Eval is scored with its official instruction-following checker, which applies deterministic constraints to the generated responses. LiveBench 1125 is evaluated with the official LiveBench ground-truth evaluation protocol, which applies the appropriate exact, rule-based, symbolic, or execution-based scorer to each task family.

Preference-based Generation Benchmarks. Arena-Hard v2 requires model-based evaluation because its outputs are judged by pairwise preference. We use GPT-4.1 as the judge and follow the Arena-Hard automatic judging pipeline. We report the Hard Prompt and Creative Writing subsets separately as HP and CW.

B.1 Role-Play Evaluation Protocols

For role-play preservation, we follow the official Given-Circumstance Acting (GCA) evaluation protocol from CoSER (Wang et al., 2025). The protocol simulates role-playing dialogues from the provided character profiles, scenario, and reference dialogue, and evaluates the generated dialogue using the CoSER scoring logic. In our experiments, Qwen2.5-72B-Instruct serves as the fixed environment model, next-speaker predictor, and judge, while the evaluated model varies. We also evaluate the models in Table 1 using the official CoSER GPT-4o-based GCA configuration and obtain the same model ranking. Given the substantial cost of multi-turn dialogue simulation and judging, we

use the Qwen2.5-72B-Instruct configuration for all reported main, ablation, and hyperparameter evaluations. We report the four standard GCA dimensions—Storyline Consistency, Anthropomorphism, Character Fidelity, and Storyline Quality—and use their arithmetic mean as the role-play average.

B.2 Medical Reasoning QA Evaluation Protocols

For the medical reasoning QA instantiation, we evaluate domain preservation with two objective benchmarks: MedXpertQA Text (Zuo et al., 2025) and MedQA-USMLE (Jin et al., 2021).

MedXpertQA Text. MedXpertQA Text is a text-only expert-level medical multiple-choice benchmark from the Text subset of MedXpertQA (Zuo et al., 2025)⁸. We evaluate the full test split, which contains 2,450 examples. The model is prompted to choose a single option and end with an explicit final answer. We extract the predicted option letter from the final-answer pattern and compute exact-match accuracy against the reference label.

MedQA-USMLE. MedQA-USMLE (Jin et al., 2021) is a four-option medical exam benchmark. We use the GBaker/MedQA-USMLE-4-options test split, which contains 1,273 examples. The model is prompted as a multiple-choice question-answering system and is required to end with `Final answer: <letter>`. We extract the predicted option and compute exact-match accuracy.

C Hyperparameter Analysis

We further study the training schedule and two hyperparameters in the general recovery branch. The recovery-to-preservation schedule controls how often general recovery is updated relative to domain preservation. The recovery scaling ratio r_g multiplies the gap signal of selected general-branch samples in Eq. (7). The gap-score mass target ρ_g controls how much total recovery gap-score mass must be covered by the selected samples in Eq. (6). The main-text configuration uses a 3-1 recovery-to-preservation schedule, $r_g = 2$, and $\rho_g = 0.8$. We compare it with Vanilla MOPD and schedule, scaling, and mass-target variants, keeping the teachers, data, and other optimization settings unchanged.

⁸<https://huggingface.co/datasets/TsinghuaC3I/MedXpertQA>

Table 6: Evaluation protocol for Medical Reasoning QA benchmarks. All scores are objective exact-match metrics.

Benchmark	Evaluation set	Scoring/aggregation
MedXpertQA Text	Full text test split; 2,450 examples	Multiple-choice exact match from parsed final-answer letter
MedQA-USMLE	GBaker/MedQA-USMLE-4-options test split; 1,273 examples	Four-option exact match from parsed final-answer letter

Table 7: General and Role-Play capability under different recovery schedules and general-recovery hyperparameters. Higher is better for all metrics.

		1-1	5-1	3-1	3-1	3-1	3-1	3-1	3-1
	Vanilla	$r_g=2$	$r_g=2$	$r_g=1$	$r_g=1$	$r_g=2$	$r_g=2$	$r_g=3$	$r_g=3$
Metric	MOPD	$\rho_g=0.8$	$\rho_g=0.8$	$\rho_g=0.6$	$\rho_g=0.8$	$\rho_g=0.6$	$\rho_g=0.8$	$\rho_g=0.6$	$\rho_g=0.8$
<i>General capability</i>									
GPQA-Diamond	61.99	61.11	60.23	61.87	61.62	61.62	62.63	60.98	62.50
ZebraLogic	71.30	74.10	76.70	74.40	74.20	74.50	76.30	74.50	75.50
HMMT25	29.17	31.46	29.79	31.46	30.83	31.25	31.04	30.42	31.88
LiveCodeBench v5	31.54	37.99	37.28	34.41	35.13	34.77	37.28	34.41	35.48
LiveCodeBench v6	22.29	23.43	23.43	24.00	20.57	24.00	24.00	22.29	22.86
IF-Eval	79.67	79.48	79.85	78.74	78.93	80.78	80.41	79.67	80.96
Arena-Hard v2 (HP)	25.50	30.60	35.20	27.60	26.60	32.60	33.20	30.00	29.30
Arena-Hard v2 (CW)	13.70	27.60	31.10	22.10	25.50	27.30	33.30	28.00	30.20
LiveBench 1125	58.90	63.00	63.20	59.70	61.40	62.70	63.50	63.00	62.30
<i>Role-Play domain capability</i>									
Storyline Consistency	43.57	45.29	45.30	46.79	44.46	46.14	47.81	44.58	47.47
Anthropomorphism	34.91	36.58	35.98	34.24	35.50	34.34	33.96	35.65	34.05
Character Fidelity	41.45	41.76	42.62	42.18	42.85	44.66	44.65	44.69	42.39
Storyline Quality	51.50	52.10	51.55	54.58	53.22	53.08	53.47	53.88	54.25
Average	42.86	43.93	43.86	44.45	44.01	44.56	45.00	44.70	44.54

Effect of Recovery-to-Preservation Schedule.

Under $r_g = 2$ and $\rho_g = 0.8$, the 3-1 schedule gives the best recovery-preservation trade-off. The 5-1 schedule achieves similar general recovery, but its role-play average is lower than the 3-1 setting, indicating that domain capability starts to decline when recovery updates become too frequent. The 1-1 schedule is more conservative for recovery and also yields a lower domain average than 3-1.

Effect of Recovery Scaling and Gap-score Mass

Target. For a fixed 3-1 schedule, $r_g = 2$ and $\rho_g = 0.8$ gives the strongest aggregate general recovery. The larger gap-score mass target covers a broader set of informative samples, while the recovery scale amplifies the gradient signal of the selected high-demand samples. Other combinations can improve individual metrics, but their gains are less consistent.

Trade-off Across General and Domain Performance. Based on these results, we use the 3-1

schedule with $r_g = 2$ and $\rho_g = 0.8$ in the main experiments. This configuration balances general recovery and role-play preservation across the evaluated benchmark families.

D Detailed Ablation Results

Tables 8 and 9 report the per-metric Role-Play and Medical Reasoning QA results, respectively, for the four matched variants summarized in Table 3.

Table 8: Detailed ablation results for the Role-Play dialogue instantiation. The four variants follow the matched ablation chain in Table 3. Higher is better for all metrics, and bold marks the best result among the compared variants.

Variant	General capability									Role-Play domain capability					
	GPQA Diamond	Zebra Logic	HMMT 25	LCB v5	LCB v6	IF Eval	Arena HP	Arena CW	LiveBench 1125	Gen. Avg.	Story Cons.	Anthro.	Char. Fid.	Story Qual.	Dom. Avg.
CaMOPD	62.63	76.30	31.04	37.28	24.00	80.41	33.20	33.30	63.50	49.07	47.81	33.96	44.65	53.47	45.00
w/o selection	60.98	74.10	28.13	34.41	23.43	77.63	33.20	31.10	61.90	47.21	44.78	32.53	43.63	52.88	43.46
w/o selection and scaling	61.74	74.70	27.29	35.48	24.57	78.00	30.20	25.50	61.10	46.51	44.75	31.92	41.22	52.50	42.60
w/o selection, scaling, and decoupling	61.49	75.20	31.04	35.13	21.71	78.19	29.50	22.10	60.50	46.10	43.36	34.49	39.36	50.79	42.00

Table 9: Detailed ablation results for the Medical Reasoning QA instantiation. The four variants follow the matched ablation chain in Table 3. Higher is better for all metrics, and bold marks the best result among the compared variants.

Variant	General capability										Medical domain capability		
	GPQA Diamond	Zebra Logic	HMMT 25	LCB v5	LCB v6	IF Eval	Arena HP	Arena CW	LiveBench 1125	Gen. Avg.	MedQA USMLE	MedXpertQA Text	Med. Avg.
CaMOPD	61.99	72.60	34.58	56.63	29.71	59.89	16.40	31.80	49.00	45.84	85.78	23.63	54.71
w/o selection	60.61	73.10	31.88	56.63	28.57	58.04	15.90	31.10	48.60	44.94	85.62	23.59	54.61
w/o selection and scaling	61.36	71.70	31.46	55.56	28.00	58.23	15.90	31.40	48.10	44.63	85.62	23.55	54.59
w/o selection, scaling, and decoupling	58.84	69.90	29.79	54.84	28.00	58.04	12.10	29.00	48.20	43.19	85.39	23.71	54.55