CURRICULUM-OPD: EFFICIENT MULTI-TEACHER MULTI-DOMAIN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) transfers knowledge through token-level teacher supervision on student-generated responses. In practical applications, a student model often needs to handle tasks such as mathematical reasoning and code generation, and each task may have several potential teachers. Training in this setting requires repeated response generation and teacher evaluation, incurring substantial computational cost. To improve distillation efficiency, we examine how teacher feedback and task gains evolve during training and find that continued distribution alignment does not necessarily yield continued task improvement. We distinguish contextual and background feedback by comparing teacher–student log-probability differences with and without the input question. Background feedback gradually becomes dominant as task gains diminish, suggesting that contextual feedback may provide more direct guidance for learning task-specific knowledge. Based on this analysis, we propose Curriculum-OPD, an efficient framework for multi-teacher multi-domain OPD. The framework selects training examples using contextual feedback on short student responses. On-Policy Hybrid Distillation (OPHD) assigns external teacher supervision or reference-conditioned self-distillation to each example. Domain-balanced update projection further addresses differences in gradient scale and conflicting update directions during joint training. Experiments on seven benchmarks across mathematics, code, and medicine demonstrate that Curriculum-OPD outperforms standard OPD in all three domains. Learning-curve comparisons further show that Curriculum-OPD reaches comparable performance in fewer training updates. Code is available at: https://anonymous.4open.science/r/Curriculum-OPD-574D/.

1 INTRODUCTION

In the post-training phase of large language models (LLMs), on-policy distillation (OPD) is a training framework in which a student is trained on its own generated responses with token-level supervision from a teacher (Agarwal et al., 2024). OPD facilitates capability transfer from a larger teacher as well as the integration of capabilities from different models into a single student model (Yang et al., 2026; Hou et al., 2026). In particular, multi-teacher distillation is important for multi-domain post-training, where experts in mathematical reasoning, code generation, and other domains can contribute their capabilities to the same student model.

Multi-teacher distillation requires loading multiple models and repeatedly generating student responses and evaluating teacher feedback. Such costs lead to the need for proper allocation of training resources. Difficulty-based data selection employs student success rates and difference in distributions respectively in order to allocate the training resources (Xu et al., 2026a; Zhao et al., 2026b). However, estimating success rates involves repeated experiments while difference in distributions might not measure how relevant the feedback is to the current question. In the case of multi-domain distillation, current approaches tend to emphasize the extension of training to a greater number of domains, whereby there is one teacher per domain (Ma et al., 2026). However, such an approach fails to take into account the fact that different problems within a certain domain might require different teachers. Dynamic domain balancing involves the adjustment of sampling probabilities or loss weights based on the learning process of each domain (Gao et al., 2026; Sun et al., 2026b). These limitations motivate the following question: **how can a student model acquire new knowledge more effectively from multiple teachers across different domains?**

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.

To understand which feedback is relevant to task improvement, we compare the training dynamics of OPD under different teacher LLMs. Across three domains, the student LLM generally improves rapidly early in training. Later gains diminish or fluctuate, even as teacher alignment continues. We further compare teacher–student log-probability differences on the same student-generated continuations with and without the input question. We refer to these feedback mechanisms as contextual feedback and background feedback, respectively. As gains in task performance diminish, background feedback gradually becomes dominant. These observations suggest that the contextual feedback related to the questions may be more useful for improving downstream performance. We therefore use contextual feedback to guide sample selection.

Based on this analysis, we propose Curriculum-OPD, an efficient framework for multi-teacher multi-domain OPD. It adopts three strategies for efficient training example selection, supervision allocation, and joint training coordination. First, we use contextual feedback to assess the learning value of candidate questions and select a compact training set. This assessment relies on short student responses, making it efficient to screen a large question pool. Second, different questions within a domain may benefit from different teachers. On-policy hybrid distillation (OPHD) therefore selects supervision for each question by comparing feedback from external teachers and reference-conditioned self-distillation. Third, domain-balanced update projection copes with problems caused by the different gradient scales and conflicting update directions during joint training.

Our contributions are threefold:

- We distinguish contextual and background feedback through teacher–student log-probability differences with and without the input question. Background feedback dominates as task gains diminish. This suggests that contextual feedback provides more direct task-specific knowledge.
- We propose Curriculum-OPD, an efficient framework for multi-teacher multi-domain OPD. It combines sample selection via contextual feedback with teacher selection through on-policy hybrid distillation (OPHD). Domain-balanced update projection coordinates joint training.
- Experiments on seven benchmarks across mathematics, code, and medicine demonstrate that Curriculum-OPD outperforms standard OPD in all three domains. Learning-curve comparisons further show that Curriculum-OPD reaches comparable performance in fewer training updates.

2 PRELIMINARIES

Let π_θ be a student, $\pi_0 = \pi_{\theta_0}$ its initialization, and $\{q_m\}_{m=1}^M$ external teachers sharing its vocabulary. At update k, the student generates $y = (y_1, \ldots, y_T) \sim \pi_{\theta_k}(\cdot \mid x)$ for question x. Write $\xi_t = (x, y_{<t})$ and $p_k(\cdot \mid \xi_t) = \pi_{\theta_k}(\cdot \mid \xi_t)$; this rollout policy stays fixed during batch optimization.

On-policy distillation. OPD scores student-generated prefixes with a teacher (Agarwal et al., 2024). At a fixed prefix ξ_t, its signal compares teacher and rollout log-probabilities:

$$\begin{aligned} A_{m,t}^{\mathrm{OPD}} &= \log q_m(y_t \mid \xi_t) - \log p_k(y_t \mid \xi_t), \\ \mathbb{E}_{y_t \sim p_k(\cdot \mid \xi_t)}[A_{m,t}^{\mathrm{OPD}}] &= -D_{\mathrm{KL}}(p_k(\cdot \mid \xi_t) \| q_m(\cdot \mid \xi_t)). \end{aligned} \tag{1}$$

A positive signal favors tokens the teacher supports more than the rollout policy. With teacher and prefix fixed, its detached-advantage surrogate matches the reverse-KL gradient at θ_k (Appendix G). ExOPD is a reward-extrapolated variant (Yang et al., 2026), defined in Section 4.2.

On-policy self-distillation. Reference-conditioned OPSD scores student-generated prefixes using additional reference information s (Zhao et al., 2026c). Generation sees only x, while the detached scoring branch with parameters ϕ also sees s. The two supervision targets are

$$q_{m,t}^{\mathrm{ext}}(\cdot) = q_m(\cdot \mid \xi_t), \qquad q_{\phi,t}^{\mathrm{self}}(\cdot) = \mathrm{stopgrad}[\pi_\phi(\cdot \mid x, s, y_{<t})]. \tag{2}$$

Substituting $q_{\phi,t}^{\mathrm{self}}$ for q_m in Eq. (1) gives the OPSD signal. Reference conditioning can change the target even when $\phi = \theta_k$. Section 4.2 combines these supervision sources in OPHD.

3 UNDERSTANDING ON-POLICY DISTILLATION

Does continued teacher alignment imply continued task improvement? We examine both alongside feedback with and without the question to understand what remains to be learned during distillation.

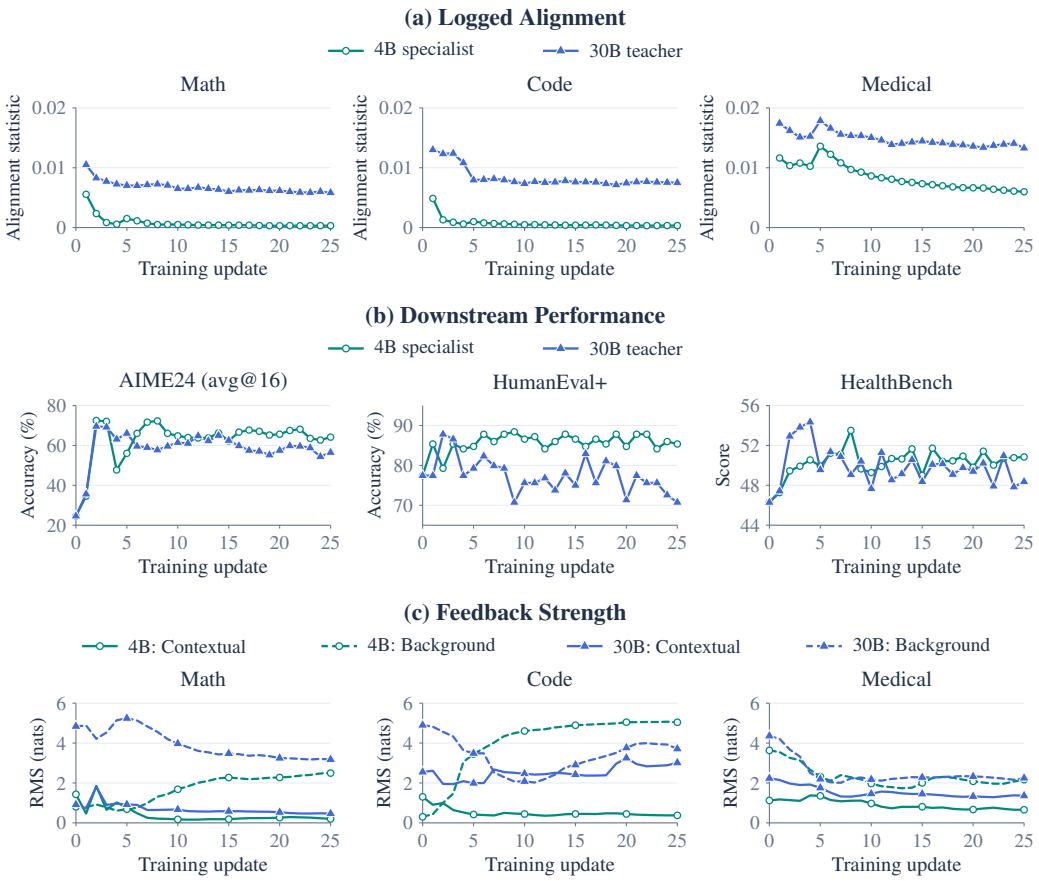


Figure 1: Specialist and 30B supervision over 25 ExOPD updates: (a) logged alignment (Appendix G), (b) downstream performance, and (c) contextual and background feedback.

3.1 Two Learning Phases in On-Policy Distillation

Setup. Starting from Qwen3-4B, we compare domain specialists and a 30B generalist over 25 ExOPD updates in mathematics, code, and medicine. Downstream evaluation uses AIME24, HumanEval+, and HealthBench. Feedback diagnostics use fresh 16-token previews on 256 fixed questions per domain. Appendix D.2 gives the evaluation and feedback-diagnostic protocols.

Results. Figure 1(a–b) shows rapid early gains and domain-dependent later trajectories. Mathematics declines from its early peaks, medicine fluctuates, and the code specialist largely retains its improvement. The logged alignment statistic continues to decrease as task gains plateau or fluctuate.

To examine the feedback behind these trends, Figure 1(c) compares teacher–student log-probability differences with the question present (r, contextual feedback) and removed (b, background feedback), summarized by per-question RMS magnitudes. With specialists, r falls rapidly in mathematics and code while b grows later; in medicine, both decline overall. Using background feedback as a proxy for generic response preferences motivates discounting the disagreement that persists without the question during sample selection (Section 4.1).

Takeaway. A decreasing alignment statistic does not reliably imply continued task improvement. This motivates screening for question-dependent feedback.

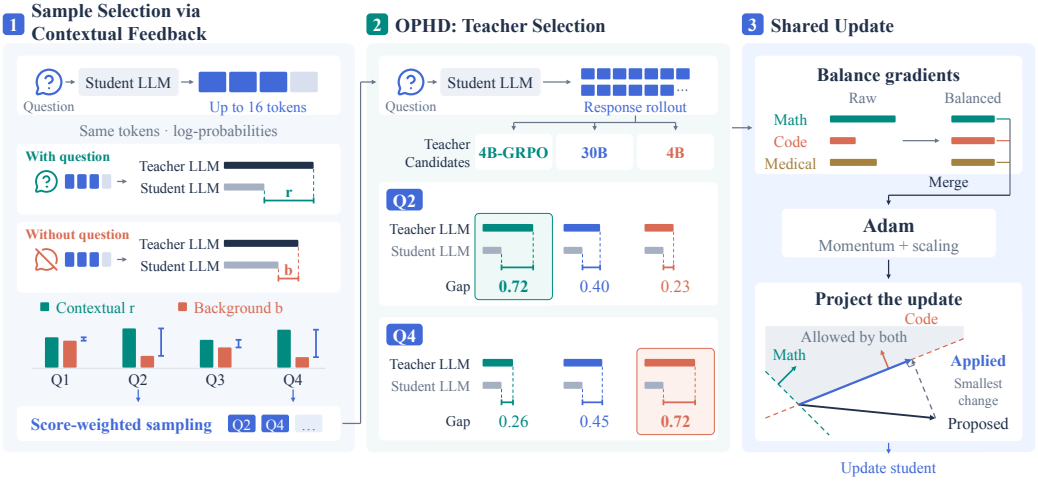


Figure 2: Curriculum-OPD combines sample selection via contextual feedback, Gap-based teacher selection, and domain-balanced update projection. The 4B candidate in the middle panel is the initial student conditioned on verified reference information s; its gray comparison bar uses the same student without s. Gap values are illustrative.

4 Curriculum-OPD

In this paper, we propose Curriculum-OPD, an efficient framework for multi-teacher multi-domain on-policy distillation (Figure 2). It combines sample selection via contextual feedback to prioritize question-relevant learning signals; On-Policy Hybrid Distillation (OPHD) to select external supervision or reference-conditioned self-distillation for each question; and domain-balanced update projection to address gradient-scale differences and update conflicts in joint training.

4.1 Sample Selection via Contextual Feedback

Samples differ in their learning value, but correctness-based screening requires repeated rollouts and measures current task success rather than the learning signal itself (Xu et al., 2026a). Building on our preceding analysis (Section 3), we use contextual feedback from short student responses to estimate the question-dependent component of teacher supervision and guide efficient sample selection.

For question x, the initial student π_0 generates $w_{1:L_x}$ up to EOS or L tokens. We score each preview token with the specialist q_s and student π_0, both with and without the question, and subtract student from teacher log-probabilities to obtain contextual feedback r_t and background feedback b_t:

$$\begin{aligned} r_t &= \log q_s(w_t \mid x, w_{<t}) - \log \pi_0(w_t \mid x, w_{<t}), \\ b_t &= \log q_s(w_t \mid \varnothing, w_{<t}) - \log \pi_0(w_t \mid \varnothing, w_{<t}), \\ \widetilde{r}_t &= \mathrm{clip}(r_t - b_t, \min(0, r_t), \max(0, r_t)) \,. \end{aligned} \tag{3}$$

All four scores reuse the same token w_t and preceding tokens $w_{<t}$; empty-question scoring does not generate another response. Here $\varnothing$ preserves the prompt template with an empty user question. Subtracting b_t discounts background feedback, while clipping prevents amplification or reversal of r_t. With $p_t = \pi_0(w_t \mid x, w_{<t})$, we compute the root-mean-square sensitivity score:

$$S(x) = \left[\frac{1}{L_x} \sum_{t=1}^{L_x} \left(\widetilde{r}_t (1 - p_t) \right)^2 \right]^{1/2} . \tag{4}$$

The factor $1 - p_t$ is the sensitivity of $\log p_t$ to its token logit; squaring avoids cancellation across signs (Appendix C). We sample N examples without replacement using $S(x)$ and independently shuffle their training order. Selection requires no reference answers; weights affect sampling only.

4.2 On-Policy Hybrid Distillation (OPHD)

Not every teacher provides useful supervision for a given sample, and its feedback may even reinforce incorrect responses (Hou et al., 2026). OPHD therefore uses the teacher–student log-probability Gap to select a suitable supervision source for each sample, while also accommodating reference-conditioned self-distillation alongside external teachers.

Selecting a supervision source. The initial student generates $z \sim \pi_0(\cdot \mid x)$ once per retained question. All candidates score the same tokens. Eligible sources $\mathcal{E}(x)$ comprise the specialist, 30B, and the reference-conditioned initial student when verified reference information s is available. Using ℓ_q for the mean token log-probability of z, we assign

$$\begin{aligned} G_m(x,z) &= \ell_{q_m}(z \mid x) - \ell_{\pi_0}(z \mid x), \quad m \in \{\text{spec}, 30\text{B}\}, \\ G_{\text{self}}(x,z) &= \ell_{\pi_0}(z \mid x, s) - \ell_{\pi_0}(z \mid x), \\ m(x) &= \arg\max_{m \in \mathcal{E}(x)} G_m(x,z). \end{aligned} \tag{5}$$

All three Gaps subtract the same question-only initial-student score. The self-distillation Gap compares that student with and without reference information s. This shared baseline makes source selection a likelihood ranking on z, even when all Gaps are nonpositive. Assignments and tie-breaking are fixed before training.

Learning from the selected source. Source selection precedes training, while learning uses fresh responses $y \sim \pi_{\theta_k}(\cdot \mid x)$. External supervision follows ExOPD (Yang et al., 2026), recovering OPD at $\lambda = 1$. With $\ell_{k,t} = \log p_k(y_t \mid \xi_t)$ and $\ell_{0,t} = \log \pi_0(y_t \mid \xi_t)$,

$$A^{\text{ext}}_{m,t} = \lambda[\log q_m(y_t \mid \xi_t) - \ell_{0,t}] - [\ell_{k,t} - \ell_{0,t}]. \tag{6}$$

For self-distillation, reference conditioning provides a detached target refreshed at each update:

$$q^+_{k,t} = \text{stopgrad}[\pi_{\theta_k}(\cdot \mid x, s, y_{<t})], \qquad A^{\text{self}}_t = \log q^+_{k,t}(y_t) - \ell_{k,t}. \tag{7}$$

Using the clipped surrogate $\mathcal{L}_{\text{clip}}$ (Appendix G), each response receives one source's loss:

$$\mathcal{L}_x = \begin{cases} \mathcal{L}_{\text{clip}}(A^{\text{ext}}_{m(x)}), & m(x) \in \{\text{spec}, 30\text{B}\}, \\ \mathcal{L}_{\text{clip}}(A^{\text{self}}), & m(x) = \text{self}. \end{cases} \tag{8}$$

4.3 Domain-Balanced Update Projection

Multi-domain training can suffer from conflicting objectives, where improving one domain comes at the expense of another (Sener & Koltun, 2018). We propose domain-balanced update projection to balance the gradient contributions of each domain and constrain conflicting updates, enabling coordinated learning across domains within a shared student.

Balancing domain contributions. Let g_r be domain r's mean gradient, preserving its teacher proportions, and p_r its sample fraction. Across D domains, we normalize gradients to control their relative influence, then restore the norm of the ordinary mixture $\bar{g} = \sum_r p_r g_r$:

$$u = \frac{1}{D} \sum_{r=1}^{D} \frac{g_r}{\|g_r\|}, \qquad g = \frac{\|\bar{g}\|}{\|u\|} u. \tag{9}$$

A shared Adam optimizer uses g, falling back to $\bar{g}$ in degenerate cases.

Constraining the shared update. Domain means can conceal group conflicts, and Adam can alter the balanced direction (Liu et al., 2026). We therefore project the proposed displacement $d_0 = \theta_k - \widetilde{\theta}_{k+1}$ using normalized nonzero loss gradients h_j of nonempty domain–supervisor groups:

$$d^* = \arg\min_d \frac{1}{2}\|d - d_0\|^2 \quad \text{subject to} \quad h_j^\top d \geq 0 \quad \forall j, \qquad \theta_{k+1} = \theta_k - d^*. \tag{10}$$

The projection makes the smallest Euclidean correction satisfying the group constraints (Lopez-Paz & Ranzato, 2017). At differentiable points, the applied step $-d^*$ has a nonpositive linearized loss change for each group on the frozen batch. Adam's moment states remain unchanged. Appendices F and H give finite-step conditions and the algorithm.

5 EXPERIMENTS

We evaluate Curriculum-OPD through five research questions, covering capability transfer, training efficiency, and the contribution of its design choices:

- RQ1: How does Curriculum-OPD perform in single-domain distillation?
- RQ2: How many updates are needed to reach comparable task performance?
- RQ3: Can Curriculum-OPD improve mathematics and code in a shared student?
- RQ4: How does teacher assignment affect response correctness and question coverage?
- RQ5: What does the ablation study reveal about the contribution of each component?

5.1 EXPERIMENTAL SETUP

Models. Students are initialized from Qwen3-4B. External teachers comprise the domain specialists 4B-Math, 4B-Code, and 4B-Medical, alongside Qwen3-30B-A3B-Instruct-2507. We compare individual teachers and the two-teacher pool within each domain.

Datasets. External distillation uses 1,792 questions per domain from DeepMath, Eurus code, and MedCaseReasoning, respectively. Evaluation covers AIME 2024/2025 and HMMT 25 (Feb./Nov.) for mathematics, HumanEval+ and MBPP+ for code, and 250 fixed HealthBench questions.

Baselines. We compare with standard OPD (Agarwal et al., 2024), reward-extrapolated ExOPD (Yang et al., 2026), and reference-conditioned OPSD (Zhao et al., 2026c).

Implementation Details. Single-domain experiments train separate students; joint experiments train one student on mathematics and code for 14 updates. OPD, ExOPD, and Curriculum-OPD share questions and data order in the joint comparison; OPSD follows its reference-conditioned training recipe. Training uses batch size 256 and learning rate 10^{-5} for OPD, ExOPD, and Curriculum-OPD. OPHD assigns one eligible source to each retained question, and joint training adds domain balancing and update projection. Appendix B provides the full protocol.

Evaluation Metrics. Mathematics and code scores average correctness over 32 and four responses per question, respectively, with equal weight across benchmarks within each domain. Code uses EvalPlus tests. HealthBench averages weighted rubric scores over four responses per question, using a local Qwen3-32B judge. Scores measure rubric satisfaction (Appendix E).

5.2 SINGLE-DOMAIN DISTILLATION (RQ1)

In the single-domain setting, we compare single-teacher, multi-teacher, and Curriculum-OPD distillation. Table 1 shows that distilled performance does not simply follow original teacher strength: the domain specialist outperforms the larger general teacher in mathematics and code because its task-focused distribution better matches the student's domain-specific trajectories. Fixed or uniformly averaged supervision can therefore introduce signals that are irrelevant to a given sample. Supervision should be matched to each sample rather than chosen by teacher size alone.

Curriculum-OPD selects informative samples using contextual feedback and assigns a suitable supervision source to each sample using Gap. OPHD includes reference-conditioned self-distillation alongside external teachers, allowing either source to supervise each selected sample.

5.3 TRAINING EFFICIENCY (RQ2)

To analyze training efficiency, we compare task gains across training updates. Figure 3 shows that standard OPD and ExOPD improve more slowly, whereas Curriculum-OPD reaches comparable mathematics performance in fewer updates. The baselines continue to supervise the full question pool, so as the student approaches the teacher, additional alignment increasingly reflects task-irrelevant preferences and yields smaller task gains per update. Curriculum-OPD uses the contextual-feedback score to construct a compact question pool and concentrates the limited updates on question-relevant supervision, producing earlier task improvement. After a question is retained, the teacher–student log-probability Gap is used to choose its supervision source rather than to select the question itself. Thus, training efficiency depends on the task relevance of supervision in each update, not only on the number of updates. Because the curves use different training protocols,

Table 1: Single-domain distillation of Qwen3-4B. The 4B Teacher is the domain specialist; 30B Teacher is Qwen3-30B-A3B-Instruct-2507. Bold and underline mark the best and second-best student scores.

Method	Math Reasoning					Code Generation			Medical
	AIME 2024	AIME 2025	HMMT 25 (Feb.)	HMMT 25 (Nov.)	Overall	Human Eval+	MBPP+	Overall	Health Bench
Original Models									
4B	22.70	19.70	10.80	9.40	15.70	75.20	65.90	70.55	45.06
4B-Math	58.60	56.20	32.30	37.10	46.10	84.30	69.70	77.00	47.06
4B-Code	36.80	32.10	17.30	20.30	26.60	88.00	70.80	79.40	46.85
4B-Medical	61.25	45.52	29.90	39.17	43.96	78.81	74.07	76.44	49.74
30B	75.80	61.60	43.00	55.40	59.00	81.20	78.00	79.60	54.58
Single-Teacher Distillation									
OPD (4B Teacher)	62.81	55.10	32.81	37.81	47.14	83.69	69.11	76.40	49.25
OPD (30B Teacher)	55.52	43.44	29.06	34.90	40.73	80.49	69.97	75.23	49.50
ExOPD (4B Teacher)	<u>63.75</u>	54.27	<u>36.15</u>	**39.17**	<u>48.33</u>	84.30	69.78	77.04	50.09
ExOPD (30B Teacher)	54.90	48.23	31.04	35.00	42.29	75.30	66.73	71.02	49.41
Multi-Teacher Distillation									
OPD	58.65	51.77	30.31	37.92	44.66	84.30	70.17	77.24	49.28
ExOPD	62.40	<u>56.88</u>	33.33	38.54	47.79	<u>84.76</u>	<u>70.97</u>	<u>77.86</u>	**50.46**
Curriculum-OPD	**66.04**	**58.33**	**36.56**	<u>38.85</u>	**49.95**	**86.81**	**71.03**	**78.92**	<u>50.27</u>

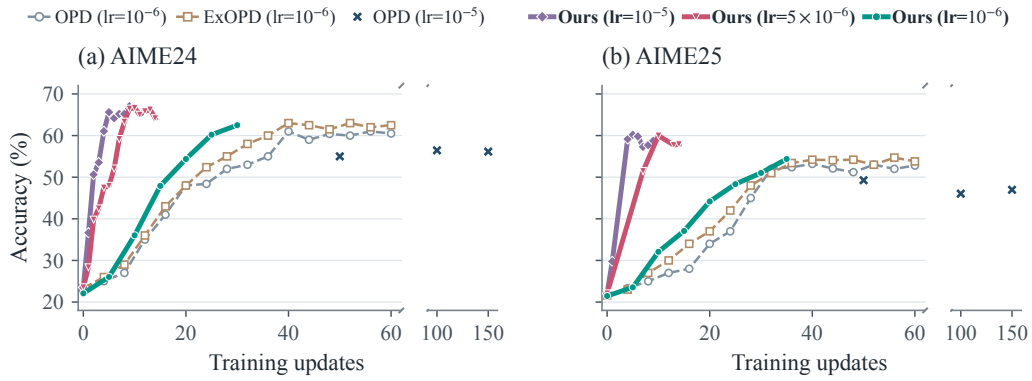


Figure 3: AIME24 and AIME25 accuracy versus training updates at different learning rates.

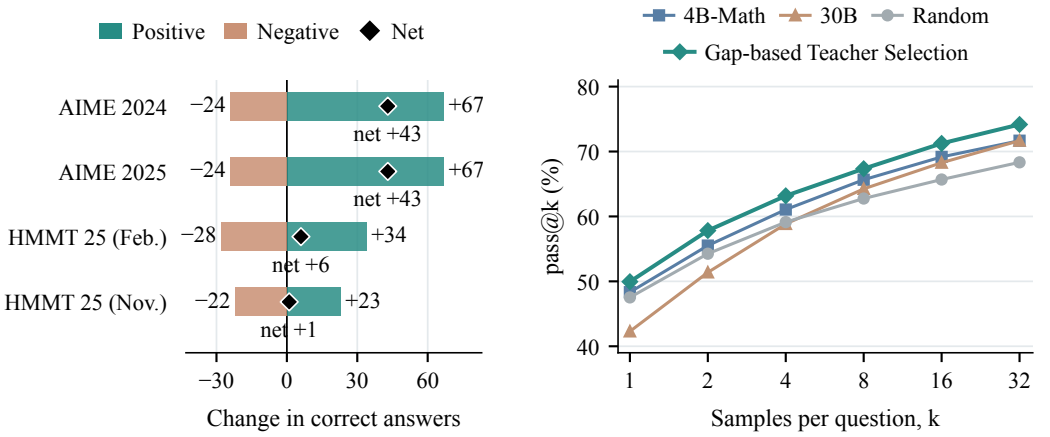
they support an update-efficiency comparison rather than a claim of strict computational speedup or wall-clock savings without matched hardware and implementation measurements.

5.4 Multi-Domain Distillation (RQ3)

To evaluate joint mathematics and code learning in one student, we compare the distillation configurations in Table 2. These results suggest that external teacher supervision provides knowledge that reference-conditioned self-distillation alone does not fully capture. Integrating supervision from multiple domains further calls for coordinating gradient scales and update directions. Curriculum-OPD uses Gap to select more useful supervision and domain-balanced update projection to coordinate the joint update, achieving higher average performance in both domains. These results suggest that coordinating cross-domain supervision signals and parameter updates contributes to the gains in joint distillation, beyond simply combining domain losses or mixing training examples.

Table 2: Performance of a shared Qwen3-4B student after 14 updates of joint math and code distillation. Original models are shown for reference. Bold and underline mark the best and second-best scores among the joint-training methods.

Method	Math Reasoning					Code Generation		
	AIME 2024	AIME 2025	HMMT 25 (Feb.)	HMMT 25 (Nov.)	Overall	Human Eval+	MBPP+	Overall
Original Models								
4B	22.70	19.70	10.80	9.40	15.70	75.20	65.90	70.55
4B-Math	58.60	56.20	32.30	37.10	46.10	84.30	69.70	77.00
4B-Code	36.80	32.10	17.30	20.30	26.60	88.00	70.80	79.40
30B	75.80	61.60	43.00	55.40	59.00	81.20	78.00	79.60
Multi-Teacher Multi-Domain Distillation								
OPD	60.21	<u>56.25</u>	32.71	36.98	46.54	<u>86.43</u>	67.53	<u>76.98</u>
ExOPD	<u>60.52</u>	**56.56**	<u>33.54</u>	<u>38.65</u>	<u>47.32</u>	84.30	<u>68.65</u>	76.47
OPSD	23.96	21.35	11.15	8.65	16.28	77.29	64.55	70.92
Curriculum-OPD	**63.33**	**56.56**	**35.10**	**40.73**	**48.93**	**88.57**	**71.63**	**80.10**



(a) Assignment effects (b) Solution coverage

Figure 4: Teacher selection and solution coverage. (a) Correct-answer count differences relative to Random. (b) Coverage across four mathematics benchmarks.

5.5 Analysis of Performance Gains (RQ4)

To analyze gains from teacher assignment, we compare Random with Gap-based Teacher Selection while fixing the training questions, teacher counts, and budget. Figure 4 shows that Gap-based assignment yields a positive net gain on all four mathematics benchmarks, both increasing correctness on questions already solved and expanding coverage of solvable questions. Random preserves teacher-usage proportions and does not account for the match between each question and teacher ability, whereas Gap uses question-dependent teacher–student feedback to select a suitable teacher for each sample. These results suggest that question-level teacher selection helps the student produce correct answers more consistently across repeated samples.

5.6 Ablation Study (RQ5)

Table 3 shows that the components address different failure modes. Without sample selection, uniform sampling spends updates on mastered or poorly matched questions, so the limited training budget is used on examples with weaker learning signal and both domain averages decline. Removing teacher selection also degrades both domains because non-adaptive sources can mismatch the question and the student's state, leaving no targeted corrective guidance. Removing balanced up-

Table 3: Ablation study of Curriculum-OPD components. Checkmarks denote retained modules, dashes denote removed modules; bold values mark the best score in each column.

Method	Modules			Math Reasoning					Code Generation		
	Sample selection	Teacher selection	Balanced updates	AIME 2024	AIME 2025	HMMT 25 (Feb.)	HMMT 25 (Nov.)	Overall	Human Eval+	MBPP+	Overall
ExOPD	✓	–	–	60.52	56.56	33.54	38.65	47.32	84.30	68.65	76.47
w/o sample selection	–	✓	✓	62.29	55.00	32.60	36.88	46.69	86.13	67.92	77.03
w/o teacher selection	✓	–	✓	62.08	55.83	32.40	38.44	47.19	87.80	70.90	79.35
w/o balanced updates	✓	✓	–	63.13	56.04	**35.73**	38.02	48.23	87.20	70.30	78.75
Curriculum-OPD	✓	✓	✓	**63.33**	56.56	35.10	**40.73**	**48.93**	**88.57**	**71.63**	**80.10**

dates lowers scores even after selection because directly combining mathematics and code gradients lets scale differences and conflicting directions cancel useful updates. Curriculum-OPD coordinates these three stages by identifying useful questions, matching supervision, and coordinating cross-domain updates, so its gains come from their coordination rather than simple module addition.

6 Related Work

On-policy distillation and self-distillation. OPD supervises student-generated trajectories (Gu et al., 2024; Agarwal et al., 2024). External distillation studies divergence objectives, intermediate distributions, reward extrapolation, and advantage estimation (Ko et al., 2024; Shing et al., 2025; Yang et al., 2026; Zhao et al., 2026a). Self- and context distillation construct conditional targets from reference solutions, demonstrations, or environmental feedback (Zhao et al., 2026c; Shenfeld et al., 2026; Hübotter et al., 2026; Ye et al., 2026). Analyses examine teacher–student compatibility, support on student prefixes, and gradient estimation (Li et al., 2026b; Fu et al., 2026a; Zhu et al., 2026), and reference-conditioned self-correction in thinking models (Kaur et al., 2026).

Data selection and distillation curricula. Data-efficient OPD explores repeated rollouts on few questions (Fu et al., 2026b) and teacher-generated training questions (Li et al., 2026a). Sample selection uses instruction-conditioned reference-answer losses or task-specific gradient influence (Li et al., 2024; Xia et al., 2024). Distillation curricula allocate training and token-level supervision through data balancing, competence, entropy, and disagreement (Hou et al., 2026; Xu et al., 2026a; Lee et al., 2026; Xu et al., 2026b). PG-OPD uses early-prefix compatibility to select candidates for longer rollouts (Zhao et al., 2026b). Our screening instead constructs a fixed training pool from short student previews, contrasting feedback with and without the question.

Multi-teacher and multi-domain distillation. Collaborative decoding and teacher debate construct supervision from multiple teachers (Yun et al., 2026; Wang et al., 2026). MOPD integrates domain RL teachers on student trajectories (Ma et al., 2026). Open-MOPD and D^3-MOPD address capability imbalance through domain budgets and sampling (Gao et al., 2026; Sun et al., 2026b), while TrustMOPD combines teachers using token-level reliability (Sun et al., 2026a). We assign external or self-distillation supervision per question and coordinate updates across domains.

7 Conclusion

We study how distillation feedback translates into task improvement under a limited training budget. The observed separation between rapid alignment and less consistent later gains motivates Curriculum-OPD: selecting examples via contextual feedback, selecting supervision through OPHD, and coordinating multi-domain learning through domain-balanced update projection. Under matched 14-update joint training, Curriculum-OPD improves mathematics and code in the same final student. Learning curves show earlier attainment of comparable AIME24 accuracy relative to published trajectories, connecting the limited-budget results to progress during training. Assignment controls further distinguish gains in success frequency from broader question coverage.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval, code development, experiment analysis, and manuscript editing. HealthBench responses are scored by a local Qwen3-32B judge using the protocol in Appendix E. The authors are responsible for validating AI-assisted outputs and for the final text, implementation, and scientific claims of this work.

REPRODUCIBILITY STATEMENT

Section 5.1 summarizes the models, datasets, baselines, and evaluation metrics. Appendices A–C provide model identifiers, training settings, and screening details; Appendix E specifies the medical evaluation protocol. Source code and experimental configurations are available at https://anonymous.4open.science/r/Curriculum-OPD-574D/.

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. doi: 10.48550/ARXIV.2306.13649. URL https://arxiv.org/abs/2306.13649.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting On-Policy Distillation: Empirical Failure Modes and Simple Fixes, 2026a. URL https://arxiv.org/abs/2603.25562.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan-ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking On-Policy Distillation of Large Language Models II: One Training Example, 2026b. URL https://arxiv.org/abs/2609.04172.

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, 2026. URL https://arxiv.org/abs/2608.19098.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, Chengquan Zhang, Zhuotao Tian, Han Hu, Yi Yang, Fei Wu, and Hehe Fan. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe, 2026. URL https://arxiv.org/abs/2605.03677.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement Learning via Self-Distillation, 2026. URL https://arxiv.org/abs/2601.20802.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking On-Policy Self-Distillation for Thinking Models, 2026. URL https://arxiv.org/abs/2607.05184.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards Streamlined Distillation for Large Language Models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.03898.

Chia-Hsuan Lee, Zelei Cheng, Yu Wang, Renkun Ni, Sambit Sahu, Shi-Xiong Zhang, and William Campbell. SEAD: Competence-Aware On-Policy Distillation via Entropy-Guided Supervision, 2026. URL https://arxiv.org/abs/2606.28562.

Gengsheng Li, Mao Zheng, Mingyang Song, Jie Sun, Zeyuan Liu, Ruiqi Liu, Qiyong Zhong, Haiyun Guo, Junfeng Fang, and Jinqiao Wang. Data-free On-policy Distillation, 2026a. URL https://arxiv.org/abs/2609.14193.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.naacl-long.421/.

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, 2026b. URL https://arxiv.org/abs/2604.13016.

Fengbei Liu, Rachit Saluja, Sunwoo Kwak, Ruibo Wang, Ruining Deng, Heejong Kim, Johannes C. Paetzold, and Mert R. Sabuncu. Madam: Metric-aware multi-objective adam, 2026. URL https://arxiv.org/abs/2606.03904.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08840.

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, 2026. URL https://arxiv.org/abs/2606.30406.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1810.04650.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-Distillation Enables Continual Learning, 2026. URL https://arxiv.org/abs/2601.19897.

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. TAID: Temporally Adaptive Interpolated Distillation for Efficient Knowledge Transfer in Language Models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cqsw28DuMW.

Jie Sun, Mao Zheng, Mingyang Song, Zeyuan Liu, Gengsheng Li, Houcheng Jiang, Yilin Cheng, Bichuan Feng, Yuchen Cai, Junfeng Fang, and Xiang Wang. Distill What You Trust: Reliability-Aware Multi-Teacher On-Policy Distillation, 2026a. URL https://arxiv.org/abs/2609.23697.

Zechen Sun, Zhiwei Zhang, Fei Zhao, Juntao Li, Mu Chuan, Huayu Deng, Guojian Zhan, Wenliang Chen, Yao Hu, and Min Zhang. D^3-MOPD: Dynamic Domain ScheDuling for Efficient Multi-Teacher Distillation, 2026b. URL https://arxiv.org/abs/2608.24987.

Jianze Wang, Ying Liu, Jinlong Chen, Xuchun Hu, Qilong Zhang, Yu Cao, Jun Wang, Hua Yang, Yong Xie, and Qianglong Chen. MAD-OPD: Breaking the Ceiling in On-Policy Distillation via Multi-Agent Debate, 2026. URL https://arxiv.org/abs/2605.01347.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting Influential Data for Targeted Instruction Tuning. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, and Zhipeng Wang. Paced: Distillation and on-policy self-distillation at the frontier of student competence, 2026a. URL https://arxiv.org/abs/2603.11178.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token Importance in On-Policy Distillation, 2026b. URL https://arxiv.org/abs/2604.14084.

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

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-Policy Context Distillation for Language Models, 2026. URL https://arxiv.org/abs/2602.12275.

Taewon Yun, Jisu Shin, Jeonghwan Choi, Seunghwan Bang, and Hwanjun Song. Distilling Long-CoT Reasoning through Collaborative Step-wise Multi-Teacher Decoding, 2026. URL https://arxiv.org/abs/2605.02290.

Hanyang Zhao, Haoxian Chen, Han Lin, Genta Indra Winata, David Yao, and Wenpin Tang. OPD+: Rethinking the Advantage Design for On-Policy Distillation, 2026a. URL https://arxiv.org/abs/2606.01039.

Qingfei Zhao, Huan Song, Shuyu Tian, Jiawei Shao, and Xuelong Li. Prefix-Guided On-Policy Distillation: Mining Golden Trajectories from Rollouts, 2026b. URL https://arxiv.org/abs/2606.21994.

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, 2026c. URL https://arxiv.org/abs/2601.18734.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The Many Faces of On-Policy Distillation: Pitfalls, Mechanisms, and Fixes, 2026. URL https://arxiv.org/abs/2605.11182.

Table 4: Model names and full repository identifiers used in this study.

Label	Model
4B	Qwen/Qwen3-4B
4B-Math	Keven16/Qwen3-4B-Non-Thinking-RL-Math-Step500
4B-Code	Keven16/Qwen3-4B-Non-Thinking-RL-Code-Step300
4B-Medical	nsk7153/Qwen3-4B-MedCaseReasoning-RL
30B	Qwen/Qwen3-30B-A3B-Instruct-2507

A Model identities

Table 4 lists the external models. The medical specialist is based on the 2507 model and shares compatible vocabulary indices and added-token mappings with the student.

B Detailed experimental settings

Models and training data. We initialize students from Qwen3-4B and use three domain specialists, 4B-Math, 4B-Code, and 4B-Medical, together with Qwen3-30B-A3B-Instruct-2507. Full model identifiers appear in Appendix A. Training pools contain 1,792 questions per domain. Appendix C specifies contextual-feedback screening and the code training configuration. Joint training combines the DeepMath and code pools; its OPD and ExOPD comparisons use the same questions and order to compare supervision and update rules.

Teacher settings. Table 1 separates original models, single-teacher students, and multi-teacher students. Single-teacher OPD and ExOPD use each external teacher separately: 4B Teacher denotes the domain specialist, and 30B Teacher denotes Qwen3-30B-A3B-Instruct-2507. Multi-teacher OPD averages both teacher-specific losses on each response. Mathematics and code ExOPD draw questions uniformly and assign half of each batch to each external teacher.

Implementation details. Joint training runs continuously for 14 updates with batch size 256, evenly split between mathematics and code, learning rate 10^{-5}, one rollout per question, and a 16,384-token response limit. OPHD compares the eligible sources on a common initial-student response and fixes assignments before training (Section 4.2). External supervision uses $\lambda = 1.25$; self-distillation targets refresh from the reference-conditioned student at each update. Domain balancing and projection coordinate the shared update (Section 4.3).

For the main tables, single-domain OPD, ExOPD, and Curriculum-OPD train for seven updates in mathematics and code and 14 in medicine. Training uses one nominal seed. The code Curriculum-OPD score averages two seven-update runs with identical questions, order, and seeds.

Evaluation metrics. Benchmark evaluations informed development. We evaluate mathematics on AIME24, AIME25, HMMT-Feb, and HMMT-Nov, and code on HumanEval+ and MBPP+ through EvalPlus. Scores average response correctness over 32 samples per mathematics question and four per code question. Generation uses non-thinking mode, temperature 1, top-p 1, and a 16,384-token limit. Domain scores are unweighted benchmark means.

Medical evaluation uses 250 fixed HealthBench questions with four responses each and a 4,096-token limit, keeping the same sampling settings. A local Qwen3-32B judge scores the benchmark's rubrics; these values measure rubric satisfaction under this judging protocol. Figure 1 uses the separate protocol in Appendix D.2. Single-domain rows combine independently trained checkpoints, whereas multi-domain rows evaluate one shared checkpoint.

C Sample selection via contextual feedback

Local score interpretation. For a fixed prefix, let v_t be the student's vocabulary logits, $p_t^{\mathrm{vec}} = \mathrm{softmax}(v_t)$, and e_{w_t} the one-hot vector for the sampled token. Then

$$\nabla_{v_t} \log \pi_0(w_t \mid x, w_{<t}) = e_{w_t} - p_t^{\mathrm{vec}}, \qquad \|e_{w_t} - p_t^{\mathrm{vec}}\|_1 = 2(1 - p_t). \tag{11}$$

Holding the bounded coefficient $\widetilde{r}_t$ fixed, its local feedback vector is $c_t = \widetilde{r}_t(e_{w_t} - p_t^{\text{vec}})$. The sample score in Eq. (4) is therefore half the root mean square of $\|c_t\|_1$ across the preview. For nonzero r_t, Eq. (3) gives $\widetilde{r}_t = \alpha_t r_t$ with $0 \le \alpha_t \le 1$; when $r_t = 0$, the bounded coefficient is also zero. Thus the contrast contracts the magnitude of frozen token-level feedback while preserving its direction. The score summarizes local logit sensitivity at the sampled preview prefixes.

Scoring protocol. We use one preview of at most 16 tokens per candidate, sampled from the initial student in non-thinking mode with temperature 1 and top-p 1. EOS can shorten the preview and, when returned, is included in scoring. All four probability arrays use teacher forcing on identical saved token IDs, including the first token. The empty-question context removes only the user content while retaining the system message, role structure, and assistant boundary. The tested inputs contain one user turn, optionally preceded by a system turn. No reference solution is required, and neither the empty context nor the preview is appended to a training prompt.

All four views use teacher-forced prefill scoring with batch-invariant execution and prefix caching disabled. We store the shared token IDs and their scalar log-probabilities. Each scoring view includes question prefill and vocabulary normalization, with identical scoring settings across views.

Weighted sampling. For $S(x) > 0$, draw seeded $u_x \sim U(0, 1)$ and retain the N smallest keys

$$\kappa_x = \frac{-\log u_x}{S(x)}. \tag{12}$$

This exponential-race construction implements sequential sampling without replacement, with probabilities proportional to the remaining weights. The implementation compares logarithms of the keys for numerical stability and breaks ties by sample identifier. Selection requires at least N positive-score examples and excludes zero-score examples. A separate seeded hash fixes the training order independently of the sampling keys. Sampling weights never enter the training loss.

Training configuration. Mathematics and code screening retain 1,792 questions from 56,608 and 25,276 eligible candidates, respectively, scored by the corresponding 4B specialist. Code uses seven updates, batch size 256, learning rate 10^{-5}, and one rollout per question capped at 16,384 tokens.

Code extraction. Code evaluation retains the extracted program when it contains the exact required entry point. Otherwise, it takes the last syntactically valid Python code block defining that entry point. The rule prevents standalone lists or assertions from displacing the required function. Extraction uses the code syntax and entry-point name, preserving the generated implementation and benchmark tests without consulting test outcomes.

D Additional experimental protocols

D.1 Teacher-assignment controls under ExOPD

The ExOPD teacher-assignment controls use the same initial Qwen3-4B, 1,792 questions, batch size 256, seven updates, learning rate 10^{-5}, and one assigned supervisor per trajectory. Each checkpoint is evaluated with 32 responses per question on the four mathematics benchmarks.

Teacher assignment and solution coverage. Figure 4 uses four ExOPD teacher-assignment configurations: fixed 4B-Math, fixed 30B, random assignment (Random), and Gap-based Teacher Selection. All use $\lambda = 1.25$ and the training budget above. The two routing runs assign 1,132 questions to 4B-Math and 660 to 30B. Their random control preserves these counts and is distinct from the equal-half assignment in Table 1. These controls hold the training pool fixed and restrict supervision to the external teachers, isolating the effect of teacher assignment.

For each of a benchmark's 30 questions, c_i^G and c_i^R count the correct responses in 32 samples from Gap-based Teacher Selection and Random. Panel (a) aggregates their differences:

$$\Delta_i = c_i^G - c_i^R, \qquad P = \sum_i \max(\Delta_i, 0), \qquad N = \sum_i \min(\Delta_i, 0), \qquad \Delta = P + N. \tag{13}$$

The signed bars aggregate question-level count differences between independently sampled response sets. The net count divided by 960 responses gives the change in mean correctness.

Panel (b) uses the pass@k estimator of Chen et al. (2021) on each checkpoint:

$$\widehat{\text{pass@}k} = \frac{100}{120}\sum_{i=1}^{120}\left[1 - \frac{\binom{32-c_i}{k}}{\binom{32}{k}}\right], \qquad k \in \{1, 2, 4, 8, 16, 32\}, \tag{14}$$

where c_i is the correct-response count and $\binom{32-c_i}{k} = 0$ when $32 - c_i < k$. Equal benchmark sizes make the pooled question average identical to the equal-weight benchmark average. At $k = 1$, the estimator equals mean response correctness; at $k = 32$, it equals the percentage of questions with at least one observed correct response. Correctness uses benchmark answers.

D.2 Figure 1 protocol

Figure 1 evaluates initialization and every update through 25, using non-thinking generation, temperature 1, top-p 1, and a 32,768-token response limit. Mathematics evaluates all 30 AIME24 questions with 16 responses each and reports mean correctness. Code evaluates all 164 HumanEval+ questions with one response each and a 36,864-token context limit, using the main-table prompting, extraction, and tests. Where four responses are saved, the curve uses the first, independently of its test outcome; other checkpoints generate one response. Code uses avg@1; the main table uses avg@4.

Medical trajectories use the same 250 HealthBench questions, one response per question, seed 42, and a 40,960-token context window. The local Qwen3-32B judge evaluates all 2,700 rubric criteria at each checkpoint with temperature 0 and a 1,024-token output limit. Scores follow Appendix E. Initialization is reevaluated here; the main table uses four 4,096-token responses per question.

Short-preview diagnostics. Panel (c) uses 256 fixed training questions per domain, uniformly selected by a seeded hash from the corresponding 1,792-question pool. At initialization and each update through 25, the current student generates a new preview of at most 16 tokens with temperature 1, top-p 1, and a fixed per-question seed. All observed previews contain 16 tokens. We evaluate r_t and b_t in Eq. (3) using the current student π_k in place of π_0 and the teacher used in that trajectory. Both models score identical tokens with and without the question; the empty view preserves the system message, chat template, and generated prefix. Each plotted value is the mean of the 256 per-question RMS magnitudes, in nats. These quantities summarize short-preview feedback, whereas panel (a) uses the logged alignment statistic. All four views use teacher-forced prefill with batch-invariant execution and prefix caching disabled. Sample selection uses only the initial student; later checkpoints serve this diagnostic and never change the selected pool.

The mathematics specialist trajectory has 59, 35, and one truncated responses out of 480 at updates 2, 3, and 7, respectively. All responses complete within the limit at the remaining checkpoints, including the update-4 decline in correctness. The larger-teacher trajectory has 74, 94, and 130 truncated responses at updates 2–4 (15.4%, 19.6%, and 27.1%); its final checkpoint has none.

E Local HealthBench evaluation protocol

The main medical evaluation uses 250 fixed HealthBench questions and four responses per question, generated with non-thinking prompts, temperature 1, top-p 1, and a 4,096-token limit. Responses are scored against the benchmark rubrics by a local Qwen3-32B judge.

For each question, the judge evaluates criteria with their signed weights. The question score is the sum of satisfied weights divided by the sum of positive weights. We average these raw scores, clip the aggregate to $[0, 1]$, and multiply by 100. Individual question scores remain unclipped.

F Local update analysis

We characterize the constraint in Eq. (10) for a fixed batch and detached supervision targets. Let $\mathcal{L}_j$ be a differentiable supervision-group loss, $g_j = \nabla\mathcal{L}_j(\theta_k)$, and $h_j = g_j/\|g_j\|$ for nonzero g_j. Feasibility gives $g_j^\top d^* \geq 0$, so the linearized loss change along the applied step $-d^*$ is nonpositive.

The feasible set is a closed convex cone containing zero, so its Euclidean projection satisfies $\|d^*\| \le \|d_0\|$. If $\mathcal{L}_j$ has an L_j-Lipschitz gradient on a neighborhood of the update segment, then

$$\mathcal{L}_j(\theta_k - d^*) - \mathcal{L}_j(\theta_k) \le -g_j^\top d^* + \frac{L_j}{2}\|d^*\|^2. \tag{15}$$

A non-increasing finite-step loss follows when $g_j^\top d^* \ge (L_j/2)\|d^*\|^2$. For the piecewise-smooth clipped surrogate, these statements apply at differentiable starting points and, for the finite-step bound, along smooth update segments. Adam's moment states remain unprojected, and feasibility is checked on the stored displacement within numerical tolerance.

G Policy surrogate and training diagnostic

Local OPD gradient. At a fixed prefix ξ_t, the detached signal in Eq. (1) gives the surrogate $-\mathbb{E}_{y_t\sim p_k}[A^{\mathrm{OPD}}_{m,t}\log \pi_\theta(y_t \mid \xi_t)]$. At $\theta = \theta_k$, its gradient equals that of $D_{\mathrm{KL}}(\pi_\theta(\cdot \mid \xi_t)\|q_m(\cdot \mid \xi_t))$ because the extra constant has zero score-function expectation. This equality holds at fixed prefixes; sequence-level optimization also depends on their policy-induced distribution.

Clipped optimization. For $\rho_t(\theta) = \pi_\theta(y_t \mid \xi_t)/p_k(y_t \mid \xi_t)$, set $\bar{A}_t = \mathrm{stopgrad}(A_t)$. The sampled-token OPD branches use the raw log-ratio signal as the advantage, retaining its original scale. We stabilize the likelihood ratio as $\widehat{\rho}_t = \exp[\mathrm{clip}(\log \rho_t, -20, 20)]$ and use the surrogate:

$$c_t(\theta, A) = \min\left\{\widehat{\rho}_t(\theta)\bar{A}_t,\ \mathrm{clip}(\widehat{\rho}_t(\theta), 1-\epsilon_-, 1+\epsilon_+)\bar{A}_t\right\}. \tag{16}$$

The implementation applies dual clipping for negative advantages. With $c_{\mathrm{dual}} > 1$, detached rollout-correction weight w_t, and valid-response aggregation $\mathrm{Agg}_{\mathcal{M}}$, the loss is

$$\mathcal{L}_{\mathrm{clip}}(A) = -\,\mathrm{Agg}_{\mathcal{M}}\left[w_t \begin{cases} \max\{c_t(\theta, A), c_{\mathrm{dual}}\bar{A}_t\}, & \bar{A}_t < 0, \\ c_t(\theta, A), & \bar{A}_t \ge 0. \end{cases}\right]. \tag{17}$$

Only the likelihood-ratio numerator carries gradients. OPHD applies this surrogate to the advantage from the selected source in Eq. (8). External-teacher and reference-conditioned examples share the same likelihood-ratio, clipping, masking, and rollout-correction rules.

Masking and aggregation. For a batch B, let $M_{i,t}$ be its response-token mask and $T_i = \sum_t M_{i,t}$. External OPD and Curriculum-OPD use per-response averaging:

$$\mathrm{Agg}_{\mathcal{M}}(f) = \frac{1}{|B|}\sum_{i\in B}\frac{\sum_t M_{i,t}f_{i,t}}{T_i + 10^{-8}}. \tag{18}$$

The mask excludes prompts and padding, includes the first generated EOS, and retains generated tokens up to the response cap when EOS is absent. All nonempty responses ($T_i > 0$) enter training regardless of answer correctness or format. Each microbatch contains one response, and token means are accumulated with equal sample weights, implementing the per-response aggregation in Eq. (18). Domain and group losses use sample means, with $p_r = |B_r|/|B|$.

Rollout correction and constants. For each sampled token, let $\widehat{p}^{\mathrm{train}}_{k,i,t}$ and $\widehat{p}^{\mathrm{roll}}_{k,i,t}$ denote frozen pre-update probabilities from the training and rollout backends. The token-level correction is

$$w_{i,t} = \mathrm{stopgrad}\big[M_{i,t}\min\big\{5, \exp\big[\mathrm{clip}\big(\log \widehat{p}^{\mathrm{train}}_{k,i,t} - \log \widehat{p}^{\mathrm{roll}}_{k,i,t}, -20, 20\big)\big]\big\}\big]. \tag{19}$$

These weights are applied directly to every valid response token. Both PPO clipping bounds are $\epsilon_- = \epsilon_+ = 0.2$ and $c_{\mathrm{dual}} = 3$. Each batch receives one optimization epoch; entropy and auxiliary KL coefficients are zero. Single-domain and joint Curriculum-OPD runs share these settings.

Logged alignment. The training field actor/kl_loss uses the low_var_kl transformation. Define $\delta_t = \log q_m(y_t \mid \xi_t) - \log \pi_\theta(y_t \mid \xi_t)$ and $\widetilde{\delta}_t = \mathrm{clip}(\delta_t, -20, 20)$. The implementation aggregates $\mathrm{clip}(e^{\widetilde{\delta}_t} - \widetilde{\delta}_t - 1, -10, 10)$ over masked response tokens, with microbatch scaling before trainer-level logging. For an unclipped ratio with sampling and scoring under the same student distribution, the conditional expectation is reverse KL. Figure 1 plots this statistic.

The configured auxiliary KL coefficient is zero; the external advantage drives optimization. Screening scores short initial-student previews with question and empty-question contexts; teacher-selection Gap scores the fixed answer z; logged alignment transforms per-token ratios on training responses. Each diagnostic uses its own token source and aggregation rule.

H TRAINING ALGORITHM

Algorithm 1 Curriculum-OPD with OPHD teacher selection

Require: Initial student θ_0; specialist and larger teachers; candidate pool $\mathcal{D}$
screening teacher q_s; budget N per domain; preview cap L; updates K; λ
Ensure: Trained student θ_K

Freeze $\pi_0 \leftarrow \pi_{\theta_0}$; initialize shared Adam state M_0
for each domain's candidate x **do**
  Sample $w \sim \pi_0(\cdot \mid x)$, stopping at EOS or L tokens
  Score the same w with π_0, q_s under x and $\varnothing$
  Compute bounded feedback $\widetilde{r}_t$ and weight $S(x)$ Eqs. (3)–(4)
end for
Sample N questions per domain without replacement using $S(x)$ Eq. (12)
Fix an independent random training order; discard previews from training inputs
for each retained example x **do**
  Sample and cache $z \sim \pi_0(\cdot \mid x)$ for source comparison
  Form $\mathcal{E}(x)$: external teachers and reference-conditioned π_0 if eligible
  Score the same z with all $m \in \mathcal{E}(x)$; fix $m(x) \leftarrow \arg\max_m G_m(x, z)$
end for
for $k = 0, \ldots, K-1$ **do**
  Take the next batch B_k from the fixed retained-pool order
  for each $x \in B_k$ **do**
    Generate $y \sim \pi_{\theta_k}(\cdot \mid x)$ without s or z
    if $m(x) = \text{self}$ **then**
      Score y with detached $\pi_{\theta_k}(\cdot \mid x, s)$; form A^{self} Eq. (7)
    else
      Score y with $q_{m(x)}$ and π_0; form $A^{\text{ext}}_{m(x)}$ Eq. (6)
    end if
    Set $\mathcal{L}_x$ to the selected source's loss Eq. (8)
  end for
  if joint training **then**
    Compute domain means $\{g_r\}$ and domain–supervisor group gradients $\{h_j\}$
    Balance domain gradients to obtain g Eq. (9)
  else
    $g \leftarrow \nabla_\theta \, \text{mean}_{x \in B_k} \mathcal{L}_x$
  end if
  $(\widetilde{\theta}_{k+1}, M_{k+1}) \leftarrow \text{Adam}(\theta_k, g, M_k)$
  $d_0 \leftarrow \theta_k - \widetilde{\theta}_{k+1}$
  $d^* \leftarrow \text{Project}(d_0, \{h_j\})$ for joint training, else $d^* \leftarrow d_0$
  $\theta_{k+1} \leftarrow \theta_k - d^*$, retaining M_{k+1}
end for
return θ_K