EFFICIENT ON-POLICY DISTILLATION OF LONG REASONING VIA OFF-POLICY PREFIX DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a student on its own reasoning with dense teacher feedback. However, the teacher must supervise long reasoning traces that it would not have written itself, which makes the student learn slowly. We observe that the accuracy of the student greatly increases when it continues from the beginning of the reasoning generated by the teachers, and vice-versa the performance of the teacher is degraded when it has to start from a partial reasoning traces written by the student. This large effect of the reasoning prefix motivates our proposed OPD variant, **Prefix-Supervised On-Policy Distillation (PS-OPD)**. At each training step, the student continues a reasoning prefix generated offline by the teacher, and it is trained by supervised fine-tuning on the prefix and OPD on its own continuation. In this way, the student will generate higher quality answers and receive more informative supervision from the teacher, as demonstrated by faster convergence and better final performance of PS-OPD compared to OPD. For example, on AIME24, AIME25 and AIME26 with a 7B math teacher and 1.5B student, PS-OPD reaches in 7 training steps the accuracy OPD attains after 50 steps, ends 2.1 points higher, and needs about 40% fewer GPU-hours for training, including the generation of teacher traces. The gains carry over to a teacher fine-tuned by reinforcement learning from the student itself: PS-OPD leads OPD over the first ten steps and ends 5.3 points higher. PS-OPD is also more robust to the choice of distillation loss: across four forward- and reverse-KL losses, its final accuracy spans 4.0 and 1.3 points with the two teachers, against 8.0 and 4.4 for OPD. It also curbs OPD's response-length explosion: with the RL-finetuned teacher under the main loss, up to 79% of OPD's training rollouts reach the length cap, against at most 12% of PS-OPD's.

1 INTRODUCTION

On-policy distillation (OPD) trains a student on its own responses with dense token-level teacher feedback, so that it learns on the reasoning states it actually visits (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025). The effectiveness of OPD varies depending on the student-teacher pair, and on how useful the teacher feedback on student-generated contexts is (Li et al., 2026; Zhu et al., 2026). In fact, OPD works well from a *matched* teacher, finetuned by reinforcement learning (RL) from the student checkpoint itself (Li et al., 2026; Zhu et al., 2026), while stronger models are not necessarily effective teachers.

However, when applying on-policy distillation to reasoning models, which the most relevant setup with modern LLMs, the teacher feedback grows less confident and less discriminative along the student rollout (Liu et al., 2026a;b; Fu et al., 2026a). Thus, much of the dense feedback carries little signal and dilutes the useful part (Ko et al., 2026), and OPD absorbs it slowly (Fu et al., 2026b), so dense feedback alone does not make learning over long reasoning paths efficient. Moreover, OPD often leads to response length explosion, where the student generates longer and longer reasoning, often with unnecessary thoughts and repetitions, over training steps (He et al., 2026; Luo et al., 2026; Zhu et al., 2026; Liu et al., 2026a). This degeneration can lead to answer truncation and thus degradation in the performance. In this work, we aim to mitigate these issues of OPD in long-horizon reasoning LLMs by a simple and effective approach to better leverage the teacher capabilities beyond token-level supervision. Prior works observed that a weaker model (student) can generate correct solutions, which otherwise would not found, when it starts from openings written by a stronger

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.

model (teacher) (Zhang et al., 2026). Interestingly, the opposite, i.e. the teacher's accuracy keeps falling when it continues a student-written prefix of growing length, also holds (Li et al., 2026; Zhu et al., 2026; Xie et al., 2026; Zhou et al., 2026). We reproduce these effects in Figure 2 for the models and datasets used in our main experiments. Then, we propose **prefix-supervised on-policy distillation (PS-OPD)** to leverage this phenomenon for effective OPD with reasoning LLMs.

First, before fine-tuning, we collect reasoning traces generated by the teacher on the training questions. Second, at each distillation step, the student answer is prefilled with 50% of the teacher reasoning (*prefix*) for the same question, which it has to continue and complete with the final answer (unlike in pure OPD where the student generates the entire reasoning and answer). Finally, the student is trained with a supervised loss on the prefix and teacher's OPD feedback on its own continuation. Using the prefix as *context*, i.e. the starting point of the rollout, enables (1) seeing better answers during training than in OPD and (2) having the OPD loss on shorter rollouts, which preserve more informative teacher feedback. At the same time using the prefix as *target* in supervised fine-tuning (SFT) trains the student to write high-quality reasoning starts itself. Unlike off-policy distillation on the full trace, PS-OPD keeps feedback on student-generated reasoning, and unlike OPD, it supervises the better opening the student often fails to generate, thus combining the advantages of both approaches.

We validate our approach via extensive experiments, repeated over multiple random seeds, on math tasks, using DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025) as student model, two teacher models with different capabilities (Skywork-OR1-Math-7B (He et al., 2025b) and JustRL-DeepSeek-1.5B (He et al., 2025a)), and DAPO-Math-17k (Yu et al., 2025) as training dataset. When evaluated on the AIME benchmarks (Mathematical Association of America, 2024; 2025; 2026), the models distilled with our PS-OPD consistently outperform those obtained via the original OPD formulation across benchmarks and teachers, both in terms of final accuracy (see Table 2) and convergence speed (see Figures 3 and 7). PS-OPD achieve similar gains over recent variants of OPD (Zhang et al., 2026; Xing et al., 2026) (Table 1), and matches or outperforms off-policy distillation. In particular, PS-OPD outperforms TrOPD (Xing et al., 2026), which has some similar elements to our method, while maintaining a significantly simpler formulation.

Besides final performance, PS-OPD has additional benefits. First, thanks to inserting the teacher prefixes, the rollouts generated by the student during distillation are significantly shorter than in vanilla OPD, which in turns reduces training cost even when accounting for the time of generating the prefixes (Figure 5). Second, the stronger teacher supervision and higher-quality student rollouts stabilize training and prevent, or largely mitigate, response length explosion. Third, unlike OPD, PS-OPD performance remains high, and at times even improve, when replacing the reverse KL loss approximated by the top-k student tokens, a common OPD objective (Li et al., 2026; Yang et al., 2026), with teacher tokens or, with the matched teacher, with forward KL, which are popular choices in other distillation frameworks (Hinton et al., 2015; Kim & Rush, 2016) (Table 2). This highlights the flexibility of our approach and its potential to be applied to other use-cases.

In summary, **PS-OPD is a simple variant of on-policy distillation** that

- combines supervised learning on teacher-written reasoning prefixes with on-policy distillation on student-generated continuations for **better teacher supervision**,
- learns faster than OPD and achieves higher final performance, while reducing training cost, i.e. is **more efficient and effective** than the original OPD
- prevents response length explosion commonly observed in OPD and is robust to variations of the training objective, indicating a **more stable and general** algorithm.

2 Background: Off-Policy and On-Policy Distillation

Let x be a prompt, π_θ the student and π_T a frozen teacher, both autoregressive language models over a vocabulary $\mathcal{V}$ with an end-of-sequence token. A response $y = (y_1, \dots, y_L)$ has contexts $c_t = (x, y_{<t})$ and probability $\pi(y \mid x) = \prod_{t=1}^{L} \pi(y_t \mid c_t)$, so, denoting D_{KL} the KL-divergence, for any two such models P, Q we have

$$D_{\mathrm{KL}}\big(P(\cdot \mid x) \| Q(\cdot \mid x)\big) = \mathbb{E}_{y \sim P(\cdot|x)}\Big[\textstyle\sum_{t=1}^{L} D_{\mathrm{KL}}\big(P(\cdot \mid c_t) \| Q(\cdot \mid c_t)\big)\Big]. \tag{1}$$

SFT and off-policy distillation. Supervised fine-tuning (SFT) trains the student by the mean token cross-entropy of a response $z = (z_1, \ldots, z_M)$ to x:

$$\mathcal{L}_{\mathrm{SFT}}(z) = -\frac{1}{M}\sum_{t=1}^{M} \log \pi_\theta(z_t \mid x, z_{<t}) = -\frac{1}{M} \log \pi_\theta(z \mid x). \tag{2}$$

Off-policy distillation applies SFT to responses z sampled from π_T. The token sum $M\mathcal{L}_{\mathrm{SFT}}(z)$ is then, in expectation, the sequence-level objective of Kim & Rush (2016):

$$\mathbb{E}_{z\sim\pi_T(\cdot|x)}\big[M\mathcal{L}_{\mathrm{SFT}}(z)\big] = \mathcal{H}\big(\pi_T(\cdot \mid x)\big) + D_{\mathrm{KL}}\big(\pi_T(\cdot \mid x)\|\pi_\theta(\cdot \mid x)\big), \tag{3}$$

where $\mathcal{H}$ denotes the distribution entropy, and only the KL term depends on θ. With the forward per-token KL $F_\theta(c) = D_{\mathrm{KL}}(\pi_T(\cdot \mid c)\|\pi_\theta(\cdot \mid c))$, Equation 1 makes it the expected sum of F_θ over the contexts $c = (x, z_{<t})$ the teacher writes, not those the student writes at test time (Bengio et al., 2015; Ross et al., 2011), hence the distillation is off-policy.

On-policy distillation (OPD). OPD (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025) moves the contexts to the student. It freezes the current student as $\pi_{\bar\theta}$, samples $y \sim \pi_{\bar\theta}(\cdot \mid x)$, and has the teacher score it, with one forward pass returning $\pi_T(\cdot \mid c_t)$ at every position. For each context c_t, we can measure the divergence on the next-token distribution of teacher and student, i.e. $\pi_T(\cdot \mid c_t)$ and $\pi_\theta(\cdot \mid c_t)$, with either the forward or the reverse KL-divergence, which are defined as

$$F_\theta(c_t) = \sum_{v\in\mathcal{V}} \pi_T(v \mid c_t) \log \frac{\pi_T(v \mid c_t)}{\pi_\theta(v \mid c_t)}, \quad \text{and} \quad R_\theta(c_t) = \sum_{v\in\mathcal{V}} \pi_\theta(v \mid c_t) \log \frac{\pi_\theta(v \mid c_t)}{\pi_T(v \mid c_t)} \tag{4}$$

respectively. Then, the OPD loss minimizes the average divergence with either formulation, i.e.

$$\mathcal{L}_{\mathrm{OPD}}(y) = \frac{1}{L}\sum_{t=1}^{L} d_t, \quad \text{with} \quad d_t \in \big\{F_\theta(c_t),\, R_\theta(c_t)\big\}. \tag{5}$$

If $d_t = R_\theta(c_t)$, Equation 1 makes the expected token sum $L\mathcal{L}_{\mathrm{OPD}}(y)$ at $\theta = \bar\theta$ the sequence-level reverse KL, its left side with $P = \pi_{\bar\theta}$ and $Q = \pi_T$. If $d_t = F_\theta(c_t)$, the forward-KL case of generalized knowledge distillation (GKD; Agarwal et al., 2024), its gradient is the per-token SFT gradient of a token v averaged over $v \sim \pi_T(\cdot \mid c_t)$:

$$\nabla_\theta F_\theta(c_t) = -\sum_{v\in\mathcal{V}} \pi_T(v \mid c_t)\, \nabla_\theta \log \pi_\theta(v \mid c_t).$$

Forward-KL OPD thus differs from off-policy distillation in two respects: the student, not the teacher, writes the contexts, and at each one the teacher supplies its next-token distribution, not one token sampled from it.

OPD objective functions. Therefore, two choices are possible as training objective in OPD, i.e. forward KL, which makes the student cover every token the teacher finds plausible, and reverse KL which let it concentrate on a few the teacher rates highly. GKD (Agarwal et al., 2024) compares both and finds the better choice task-dependent. Computing the exact forward and reverse KL in Equation equation 4, with sums over all vocabulary entries (about 152,000 in our case), at each of thousands of positions of the rollouts would be prohibitive. Then, in practice, the KL divergence is approximated by limiting the sum to the k vocabulary tokens with highest probability (Li et al., 2026; Xing et al., 2026; Fu et al., 2026a; Zhu et al., 2026). The top-k most likely vocabulary tokens can be chosen according to either the teacher or student predicted distributions, i.e. $\pi_T(\cdot \mid c_t)$ or $\pi_\theta(\cdot \mid c_t)$. A natural choice is to use teacher top-k when optimizing forward KL (Xing et al., 2026) and, vice-versa, the student top-k for the reverse KL (Li et al., 2026), though other combinations are also used, such as reverse KL over the teacher's top-k (Fu et al., 2026a) or over the intersection of both top-k sets (Zhu et al., 2026). However, in our experiments we noticed that other configurations can be equally, or at times even more, effective, especially for PS-OPD. Therefore we explore all four possible combinations to provide an exhaustive analysis of the impact of this training choice (see Table 2).

3 PS-OPD: Prefix-Supervised On-Policy Distillation

In the following we present our proposed approach, PS-OPD, to integrate an off-policy distillation element into the common on-policy distillation framework. First we present a simple experiment about the importance of a high-quality reasoning prefix for the correctness of the final answer, which motivates the design of our algorithm, then provide the details of PS-OPD.

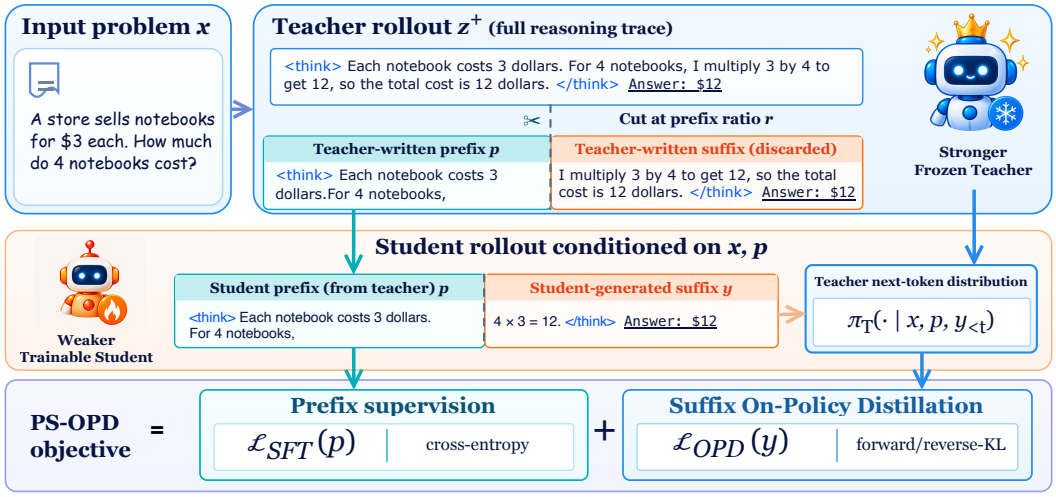


Figure 1: **Overview of PS-OPD.** A teacher rollout is cut at a fixed fraction r of its reasoning span; the student continues from the prefix p and is trained by cross-entropy on p plus distillation on its own suffix y. The example is illustrative; real traces run to thousands of tokens.

3.1 Effect of reasoning prefix on accuracy

To test the impact of a reasoning prefix, we generate answers on questions from AIME24, AIME25 and AIME26 (Mathematical Association of America, 2024; 2025; 2026) with Skywork-OR1-Math-7B (teacher) and DeepSeek-R1-Distill-Qwen-1.5B (student). For each question, we then cut the teacher's reasoning trace at 25%, 50% and 75% of its reasoning span, give the same question and the resulting reasoning prefix to the student, and let it continue until the answer. We repeat the same procedure switching the models, that is, letting the teacher conclude the reasoning started by the student. Figure 2 shows the resulting accuracy of the final answers for each configuration (error bars: 95% bootstrap intervals over questions). The student's accuracy steadily grows with the length of the teacher prefix, from 19.6% without a prefix, nearly doubling when seeing half of the teacher's reasoning, and reaching 45.3% with three quarters of it. Conversely, the teacher's accuracy drops approximately linearly with the length of the student prefix, from 59.9% without a prefix to 31.8%. This shows that the start of reasoning has a crucial impact on the final accuracy of the answers. A weak model can greatly benefit from good prefixes, and a powerful one cannot recover from a poor beginning.

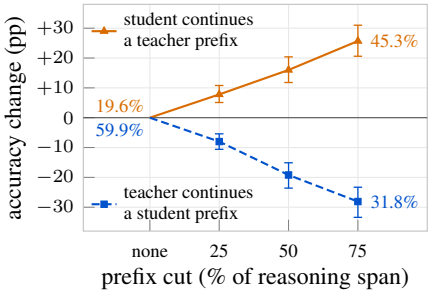


Figure 2: **Effect of prefix quality.** Each model continues the other's response, cut at 25, 50 or 75% of reasoning: the student gains from good prefixes, the teacher cannot overcome bad prefixes.

3.2 Improving OPD via off-policy prefix distillation

Given such a large impact of the teacher prefix on the weaker student's performance, we want to integrate it into the original OPD framework. We propose to achieve this with the following paradigm. First, before distillation, we use the teacher to write high-quality reasoning traces, from which we extract a prefix p for a question x, i.e., $p_t \sim \pi_T(\cdot \mid x, p_{<t})$. Then, during fine-tuning, the student writes answers y continuing from the teacher prefixes, i.e., $y_t \sim \pi_{\bar{\theta}}(\cdot \mid x, p, y_{<t})$. Finally, the student is trained by SFT to write p and by OPD on the continuation y it samples from (x, p) (see illustration in Figure 1). The same prefix is thus both *context*, the state the rollout starts from, and *target*, text the student learns to write (Section 4.3 and the context-only and loss-only controls of Appendix D analyze these two roles). At evaluation the student receives only the question and writes its opening itself.

We name our approach **Prefix-Supervised On-Policy Distillation (PS-OPD)**. This can be seen as a combination of off-policy distillation, which trains the student only on the teacher rollouts, and on-policy distillation, which relies solely on the student generations (Section 2). Below, we provide a more detailed description of the different components of PS-OPD.

Prefix construction. A frozen teacher rollout z^+ on x is cut at a fixed fraction r of its reasoning span, capped at 8,192 tokens. This gives us the prefix $p = z^+_{1:k}$ with $k = \min(\lfloor r\, M_{\text{think}} \rfloor, 8{,}192)$, where M_{think} is its length up to, not including, the closing `</think>` tag. We use $r = 50\%$ except in the length ablation (Appendix G), giving a mean prefix of 3,751 tokens. Rollouts are used regardless of final-answer correctness, so that PS-OPD does not need any verifier and could be readily extended to domains other than math. Questions with no usable trace, whose trace never closes `</think>`, or whose cut is empty get no prefix (around 15% of the cases; Appendix A gives the share for each teacher).

Loss. For a training question x and corresponding prefix p, the student, with current parameters $\bar{\theta}$, generates continuations (end of reasoning and final answers) autoregressively via $y_t \sim \pi_{\bar{\theta}}(\cdot \mid x, p, y_{<t})$. Then, on-policy distillation is carried out on the tokens in y, i.e., denoting $c_t = (x, p, y_{<t})$, the objective function becomes

$$\mathcal{L}_{\text{OPD}}(y) = \frac{1}{L}\sum_{t=1}^{L} \max\{0, d_t\}, \quad \text{with} \quad d_t \in \big\{F_\theta(c_t),\, R_\theta(c_t)\big\}, \tag{6}$$

which follows the definitions in Equation 5. Additionally, we approximate the forward ($F_\theta(c_t)$) or reverse KL ($R_\theta(c_t)$) of Equation 4 using the top-128 tokens (see the discussion of OPD objectives in Section 2). The final training loss $\mathcal{L}_{\text{PS-OPD}}$ then combines the OPD loss on the rollout with the SFT loss on the prefix, i.e.

$$\mathcal{L}_{\text{SFT}}(p) = -\frac{1}{k}\sum_{t=1}^{k} \log \pi_\theta(p_t \mid x, p_{<t})\,, \qquad \mathcal{L}_{\text{PS-OPD}} = \mathcal{L}_{\text{OPD}}(y) + \lambda\, \mathcal{L}_{\text{SFT}}(p), \tag{7}$$

where $\lambda > 0$ controls the trade-off between the two losses (in the experiments we fix $\lambda = 0.3$ except for the ablation of Section 4.3).

4 Experimental Results

4.1 Experimental Setup

Models and training data. We fine-tune `DeepSeek-R1-Distill-Qwen-1.5B` (DeepSeek-AI, 2025) (the student model) to learns single-turn mathematical reasoning with long traces from two teachers that share its Qwen2 tokenizer: `Skywork-OR1-Math-7B` (He et al., 2025b), which is RL-fine-tuned from `DeepSeek-R1-Distill-Qwen-7B`, and `JustRL-DeepSeek-1.5B` (He et al., 2025a), RL-fine-tuned from the student checkpoint itself for 4,380 RL steps. The first teacher is a stronger math model, while the second is more similar to the student, which allows us to study how these factors affect the success of our PS-OPD and the baselines. Every run uses the full DAPO-Math-17k pool (Yu et al., 2025) of 17,917 problems, unfiltered by solvability. The teachers prefixes are constructed as in Section 3.2 from one trace per problem sampled from each teacher (Appendix C.2 compares the two prefix pools).

Baselines. We compare PS-OPD to four baselines. First, **OPD** (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025) represents the original on-policy distillation approach, and trains the OPD loss in Equation 5 on student rollouts without any prefix. Second, **Forward-Horizon (FH)** adapts the method of Zhang et al. (2026), who stop student rollouts early and distill only their first generated tokens with the sampled-token reverse KL (Equation 11), with the intuition that training only on the prefix is sufficient for the success of OPD. In our experiment, FH minimizes (forward or reverse) KL (Equation 4) on the first 4,096 generated tokens of student rollouts (without any prefix), approximated with (teacher or student) top-k tokens, for direct comparison to our objective. Third, **TrOPD** (Xing et al., 2026) is the published OPD variant closest to our approach, also combining on- and off-policy distillation. For the on-policy component, student-generated tokens are divided into the trust region and outliers, which receive different objective functions. For the off-policy component, its code adds each teacher response as a separate training sequence, imitated on a share of its

Table 1: **Main comparison, both teacher regimes, two objectives.** avg@16 accuracy (%), three-seed mean at step 50. *Loss*: forward KL over the teacher's top-128 tokens (the main objective) or reverse KL over the student's. Bold marks each group's best Avg. (mean of the three benchmarks). †Our objective with FH's horizon; Zhang et al. (2026) use sampled-token reverse KL.

		AIME24		AIME25		AIME26		Avg.	
Method	**Loss**	Skywork	JustRL	Skywork	JustRL	Skywork	JustRL	Skywork	JustRL
Student		28.8		23.3		17.9		23.3	
Off-policy distillation		31.9	49.7	25.3	32.9	20.3	31.3	25.9	38.0
OPD	FKL	33.1	43.5	26.4	31.5	19.8	28.0	26.4	34.3
FH (Zhang et al., 2026)	FKL†	32.6	51.3	26.6	35.1	20.7	31.7	26.6	39.4
PS-OPD (ours)	FKL	36.0	49.1	27.5	35.9	22.0	33.8	**28.5**	**39.6**
OPD	RKL	37.2	48.9	27.5	35.1	24.3	32.2	29.7	38.7
FH (Zhang et al., 2026)	RKL	39.9	50.8	28.3	32.2	26.4	31.6	31.5	38.2
TrOPD (Xing et al., 2026)	RKL	32.6	49.1	26.0	33.7	18.8	32.8	25.8	38.5
PS-OPD (ours)	RKL	41.1	50.3	29.5	35.0	25.6	32.4	**32.1**	**39.2**
Teacher		68.1	56.3	51.7	35.6	60.6	39.2	60.1	43.7

tokens that is phased out over training (from 100% to 85% within our 50 training steps), while the student's rollouts start from the question alone. Therefore, TrOPD shares some similarity with our PS-OPD but involves a significantly more complex algorithm. We run the official implementation of TrOPD, including training objective and rollouts construction, on our models and data (settings, code checks and repository link in Appendix K.4). Finally, **off-policy distillation** trains the student with supervised fine-tuning, minimizing the cross-entropy loss, on the complete teacher traces. This represents the results on not using any student rollout, and thus most direct teacher supervision.

Training. In the main comparison we train every model for 50 training steps at a constant learning rate of 10^{-5}, checkpointed after each of the first 10 training steps and every 10 thereafter. We repeat each run with three random seeds and report the average results of the three models. A *training step* is one batch, taken in one optimizer step (two for TrOPD, following their implementation). The OPD-family arms use 32 prompts $\times$ 8 rollouts per training step and share the data, optimizer and token budgets: prompts are capped at 1,536 tokens and rollouts at 27,648. In PS-OPD rollout's teacher prefix counts toward this cap, so a prefix of h tokens leaves the student $27{,}648 - h$ tokens for the rollout, at least 19,456 because prefixes are capped at 8,192. Off-policy distillation matches the 256-sequence count and caps each sequence, prompt and trace, at 32,768 tokens. Training runs on an NVIDIA cluster, where an OPD-family run uses two nodes with four GH200 GPUs each, and on an AMD cluster, where it uses three nodes with four MI250X graphics compute dies each. More training details can be found in Appendix A.

Evaluation. We evaluate the performance of each model on AIME24 (Mathematical Association of America, 2024), AIME25 (Mathematical Association of America, 2025) and AIME26, the latter as released by MathArena (Dekoninck et al., 2026). We report avg@16 (accuracy averaged over 16 samples per problem) with a fixed 32,768-token budget (31,744 response tokens, decoding settings in Appendix A), and the unweighted average of the three (*Avg.*). Decontamination removes one AIME26 problem which appears almost identical in the training set, leaving 89 evaluation problems (contamination analysis in Appendix J). Since using different clusters lead slight fluctuations in the results for the same model, every comparison is made within one cluster (Appendix A.1 lists which experiment ran where).

4.2 Main results: performance comparison to baselines

Final accuracy. Table 1 reports the accuracy, averaged over three runs, of PS-OPD and of the baselines after 50 training steps on the AIME benchmarks, when distilling from two teachers (`Skywork`, `JustRL`). For OPD, FH and PS-OPD, we test both forward KL (FKL) over the teacher's top-128 tokens and reverse KL (RKL) over the student's top-128 as on-policy distillation losses. For both teachers and distillation losses, PS-OPD achieves the best average performance, outperforming both on- and off-policy baselines. When using FKL as loss, PS-OPD reaches 2 points higher accuracy

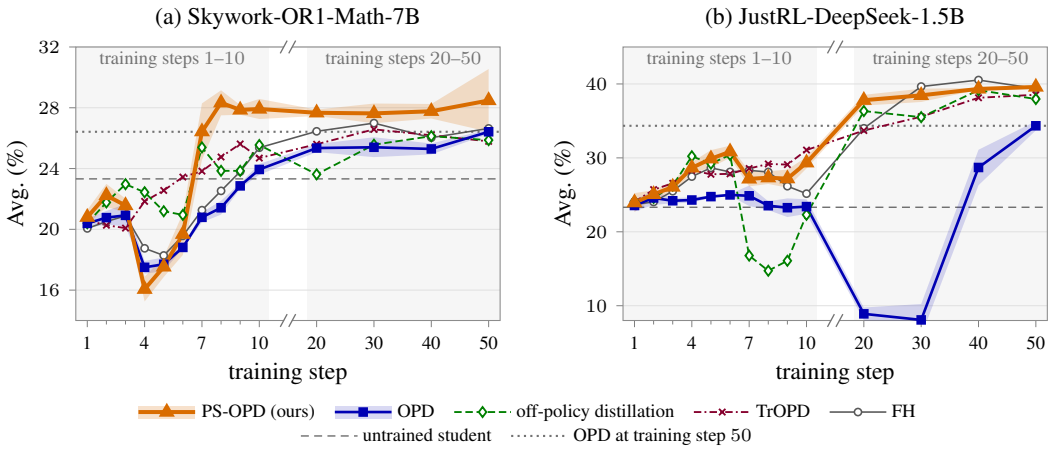


Figure 3: **Convergence under both teacher regimes.** Avg. against training step, forward KL over the teacher's top-128, three seeds mean and standard deviation.

than OPD and FH with `Skywork` as teacher, and $+5.3$ and $+0.2$ points over OPD and FH respectively with `JustRL` teacher. With RKL, PS-OPD achieves 39.2% average accuracy with the `JustRL` teacher with all baselines remaining 0.5–1.2 points lower. For the `Skywork` teacher, PS-OPD (32.1%) has slightly better accuracy than FH (31.5%) and significantly better than TrOPD (25.8%). These results show the importance of training on the prefix and limiting the number of tokens used in the OPD loss, since PS-OPD and FH have, respectively, the best and second-best results in most cases. Moreover, despite its more complex algorithm, TrOPD is consistently worse than PS-OPD. Interestingly, the stronger teacher `Skywork` leads to worse final accuracy than the weaker `JustRL`, confirming that compatibility between teacher and student is more important to the success of distillation than absolute performance of the teacher. Overall, these results show the effectiveness across benchmarks, teachers, loss functions and seeds of our PS-OPD approach.

Convergence speed. Figure 3 shows how the evaluation accuracy (average over the three benchmarks) evolves over training steps for the different methods (forward KL as distillation loss is used here), for both teachers. With the `Skywork` teacher (left plot in Figure 3), PS-OPD reaches OPD's final (step 50) accuracy after 7 training steps and stays above it afterwords, a level OPD itself reaches only at training step 50. FH follows a similar trend as OPD, thus being outperformed by PS-OPD. With the `JustRL` teacher (right plot in Figure 3), PS-OPD accuracy remains above OPD across all training steps, ending 5.3 points higher at training step 50, and above FH at most steps. In particular, OPD accuracy remains close to student initial accuracy for most steps, increasing only after many training updates. Thus, PS-OPD generally converges faster than OPD and the other baselines. We show in Figure 9 similar curves for the reverse KL objective: while the advantage in convergence speed of PS-OPD over OPD is less pronounced in that it remains, it remains noticeable.

4.3 Additional Analyses

Detailed comparison of distillation objectives. In Table 2 we provide a detailed comparison of OPD and PS-OPD when varying the distillation loss (forward vs reverse KL) and its approximation (top-k tokens sampled from the teacher or the student model). We report the results on the three benchmarks and their average for both teachers. First, we see that PS-OPD achieves higher average accuracy for both teachers across all configurations, with gains between 2% and 10%. Thus, the off-policy guidance provided by the teacher prefixes is beneficial across distillation objectives, regardless of their exact formulations. Moreover, the performance of PS-OPD remain relatively stable when varying the loss and top-k sampling, while for OPD it suffers a large drop for RKL and teacher top-k with `Skywork`. Finally, while OPD attains the best results with reverse KL and student top-k, the default setup, for both teachers, PS-OPD gets the best accuracy with forward KL for the `JustRL` teacher and with reverse KL for the `Skywork` teacher, with minimal impact of the top-k sampling choice. Thus, PS-OPD is both more stable than OPD and opens up the possibility of achieving better results by varying the formulation of the objective.

Length explosion and truncation rate. It has been observed that OPD may lead to increased response length and potentially degeneration of the quality of the answers (He et al., 2026; Luo et al.,

Table 2: **Effect of the distillation loss, OPD against PS-OPD.** Final avg@16 accuracy shown. Rollouts are student-generated throughout, so only the divergence and the scored support change.

			AIME24		AIME25		AIME26		Avg.	
Method	**Loss**	**Top-**128	Skywork	JustRL	Skywork	JustRL	Skywork	JustRL	Skywork	JustRL
Student			28.8		23.3		17.9		23.3	
OPD	FKL	teacher	33.1	43.5	26.4	31.5	19.8	28.0	26.4	34.3
PS-OPD (ours)	FKL	teacher	36.0	49.1	27.5	35.9	22.0	33.8	**28.5**	**39.6**
OPD	FKL	student	33.4	45.7	26.0	32.2	20.9	27.4	26.8	35.1
PS-OPD (ours)	FKL	student	34.6	50.9	28.8	36.0	22.0	33.3	**28.4**	**40.1**
OPD	RKL	teacher	24.0	48.7	23.5	31.5	17.5	30.3	21.7	36.8
PS-OPD (ours)	RKL	teacher	40.7	50.8	30.1	35.0	26.3	30.5	**32.4**	**38.8**
OPD	RKL	student	37.2	48.9	27.5	35.1	24.3	32.2	29.7	38.7
PS-OPD (ours)	RKL	student	41.1	50.3	29.5	35.0	25.6	32.4	**32.1**	**39.2**
Teacher			68.1	56.3	51.7	35.6	60.6	39.2	60.1	43.7

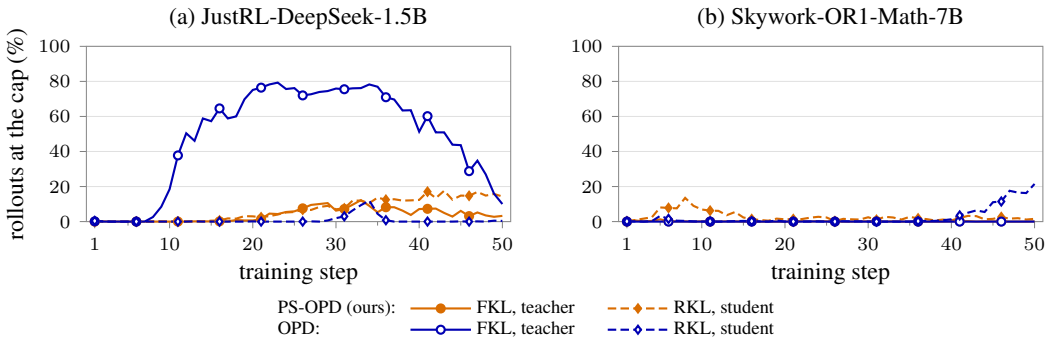


Figure 4: **Truncation of training rollouts.** Share of each training step's rollouts that reach the 27,648-token training cap, for OPD and PS-OPD under the two losses of Table 1; three-seed means.

2026; Zhu et al., 2026; Liu et al., 2026a). We analyze this by monitoring the response length and truncation rate (i.e. the percentage of generate rollouts which hit the tokens limit and thus, likely, do not contain the final answer) over training steps for OPD and PS-OPD with both distillation losses and teachers. As shown in Figure 4 we observe a steep increase in truncation rate, up to around 80%, for OPD on FKL loss with the `JustRL` teacher in the central part of the training run, while truncation rate remain significantly lower for PS-OPD. With the `Skywork` teacher, truncation rate increases to above 20% at step 50 for OPD with RKL loss while PS-OPD's remain next to 0%. The corresponding curves for the response length are shown in Figure 8. Therefore, the prefix supervision appears to prevent this pitfall of on-policy distillation, and the more stable training dynamics may in turns lead to better final performance.

Comparison of PS-OPD vs OPD computational cost. PS-OPD slightly reduces the cost of generating the student rollouts (since these start from the teacher's prefixes) and has faster convergence than in OPD as discussed above. However, it requires to generate teacher reasoning traces for the entire training datasets before fine-tuning starts. Then, in Figure 5, we compare average accuracy over GPU-hours for OPD and PS-OPD (`Skywork` as teacher) as a proxy of the computational cost of the two methods. The shaded gray area corresponds to the one-time cost of generating the teacher traces which PS-OPD has to sustain before training. Even with this initial extra cost, PS-OPD reaches OPD's step-50 accuracy after 18.9 GPU-hours, 12.74 of them for the teacher traces, while OPD first checkpoint at that accuracy is after 34.9 GPU-hours. At equal budgets of 20 and 30 GPU-hours, PS-OPD is 2.4 and 1.1 points higher.

Effect of weight of the prefix loss. We analyze the effect of varying the value of λ in the PS-OPD loss (Equation 7), i.e. the weight of the prefix loss compared to the distillation loss. Figure 6 shows average accuracy over three seeds (with standard deviation) over training steps for $\lambda = 0.3, 1, 3$.

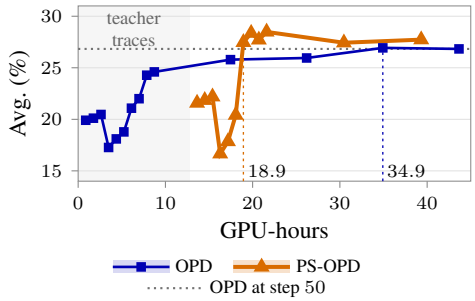


Figure 5: **Comparison of PS-OPD vs OPD computational cost.** We show average accuracy over GPU-hours. The shaded gray area corresponds to the one-time cost of generating the teacher traces. Even with this initial cost, PS-OPD outperforms OPD with less compute effort.

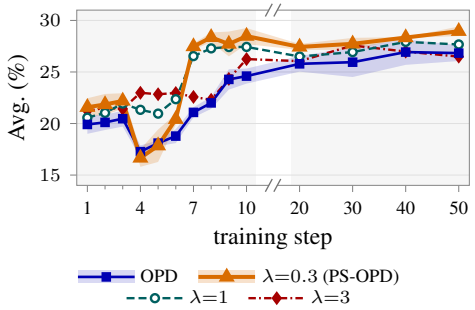


Figure 6: **Effect of the weight on the prefix loss.** We show average accuracy over three seeds over training step for different values of λ in the PS-OPD loss (Equation 7). The chosen $\lambda = 0.3$ attains the best results, but other values are also better that, or on par with, OPD.

While $\lambda = 0.3$, the default value in our experiments, attains the best results, other values outperform or match OPD, demonstrating the stability of PS-OPD over hyperparameter variations.

Other analyses. We separate the roles of the teacher prefix as context and as target with dedicated controls in Appendix D, and study the prefix length (Appendix G) and a ten times smaller learning rate (Appendix H). Finally, we confirm the advantage of PS-OPD with a preregistered test on three fresh seeds (Appendix K.2), at equal GPU-hours (Appendix B.4), and on a second cluster (Table 13).

5 Related Work

OPD and teacher choice. GKD (Agarwal et al., 2024) and sampled-token reverse-KL OPD (Lu & Thinking Machines Lab, 2025) train on student rollouts. REOPOLD (Ko et al., 2026) reallocates credit in student rollouts on our available-teacher pair, orthogonally to prefix supervision. Li et al. (2026) reported gains from SFT before OPD, which our SFT-plus-OPD controls separate from prefix supervision (Appendix D). A matched teacher transfers more than larger, comparably accurate ones (Li et al., 2026; Zhu et al., 2026). DeepSeek-V4 (DeepSeek-AI, 2026) and Kimi K3 (Kimi Team, 2026) merge RL-finetuned specialists by multi-teacher OPD.

Teacher prefixes. Teacher-written reasoning can help a weaker model continue (Zhang et al., 2026; Yang et al., 2025). Conversely, a stronger teacher continuing student-written prefixes can lose accuracy (Zhu et al., 2026), more so as they grow (Li et al., 2026; Xie et al., 2026; Zhou et al., 2026). Relay-OPD (Xu et al., 2026) showed both effects (Appendix L.1 reviews the evidence in both directions). Relay-OPD, Prefix-GRPO (Wang et al., 2026) and ReOPD (Liao et al., 2026) use teacher text as rollout context, and UFT (Liu et al., 2025) and Prefix-RFT (Huang et al., 2026), the closest RL analogues, use a verifier reward instead of teacher feedback. TrOPD (Xing et al., 2026) also uses teacher prefix as discussed in the baselines above.

6 Conclusion

PS-OPD is a simple variant of on-policy distillation, in which the student continues a teacher prefix and is trained by supervised fine-tuning on the prefix and by OPD on its own continuation. With the available teacher, PS-OPD reaches in 7 training steps the accuracy OPD attains after 50, and keeps its lead until the end of training. Thus, it needs about 40% fewer GPU-hours than OPD to reach this accuracy, including the generation of teacher traces. With the matched teacher, PS-OPD leads OPD over the first ten training steps, ends 5.3 points higher and remains stable, whereas OPD suffers from response length explosion. PS-OPD is also more robust than OPD to the choice of distillation loss, which highlights the flexibility of our approach. These results show that the teacher prefix makes on-policy distillation faster and more effective with both teachers, including a stronger teacher that is not RL-fine-tuned from the student.

AI USE STATEMENT

We used generative AI tools, namely an LLM-based coding and writing assistant and independent LLM judges, for these tasks whose disclosure the ICLR policy requires: implementing methods (training, evaluation, scoring and analysis code, including the loss implementations of the controls, and the orchestration of experiment and cluster jobs); designing and giving feedback on experiments (suggesting controls, statistical analyses and verification steps, under the authors' direction); cleaning and reformatting datasets (evaluation manifests, and the decontamination scan, whose candidate matches were adjudicated by independent LLM judges rather than human annotators; Appendix J); interpreting results (analysing the runs and drafting descriptions of results, which the authors reviewed); and proposing and refining hypotheses (candidate explanations of the early accuracy dip, and the wording of the preregistered fresh-seed hypothesis). We did not use them to generate synthetic training data, to develop theoretical models or conceptual frameworks, to formulate or prove mathematical claims, or for translation; qualitative or thematic data analysis does not apply to this work. We also used them for these recommended-disclosure tasks: creating and editing software code, suggesting experimental parameters, creating figures and tables (plot and table generators), drafting and editing parts of the paper for readability and structure, identifying and summarizing relevant literature, formatting references, suggesting a paper structure, and creating artifacts (the provenance manifest). We did not use them for transcription or survey questions.

We have reviewed all AI-assisted work. A script recomputes every value listed in the provenance manifest (Appendix K.2) from checksummed evaluation artifacts; code changes were checked with unit tests and with bit-identity checks against unchanged training paths; each contamination verdict was reached independently by two LLM judges; and the authors reviewed all AI-assisted text, claims and code. The authors take full responsibility for the content of this paper, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly released models and datasets under their own licences and involves no human subjects and no personally identifying data. The contamination audit (Appendix J) is reported to improve the interpretation of a public benchmark, not to impugn it; we identify the single excluded problem so that others can reproduce or contest the exclusion. Table 10 reports the cost per training step in GPU-hours and the one-time cost of the teacher traces, which bear on the work's environmental cost.

REPRODUCIBILITY STATEMENT

Code, configurations, prefix pools and the provenance manifest are available at the anonymous repository https://anonymous.4open.science/r/psopd-code-8845/. Appendix A documents the handoff construction, the mixed objective, training configuration, and the evaluation harness, and Appendix A.1 the hardware and software of the two clusters and where each experiment and dataset was produced. Appendix J describes statistical uncertainty and the contamination audit, including the one excluded problem. Appendix B reports the complete step-by-step tables underlying Section 4.1, including per-seed breakdowns, and Appendix C covers the objective grid, both prefix pools and the support-overlap, length and truncation diagnostics. Appendix K documents the controls and baselines, the statistical and cost analyses, and a provenance manifest, released with the code, that links the Avg. and per-benchmark values it lists to their evaluation artifacts, runs, seeds and checkpoints. The rollout diagnostics of Appendix F are computed from the saved generations, and the optimizer statistics from the training logs.

REFERENCES

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

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1171–1179, 2015.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/abs/2606.19348.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: MathArena as an evaluation platform for mathematics with LLMs. *arXiv preprint arXiv:2605.00674*, 2026. URL https://arxiv.org/abs/2605.00674.

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. *arXiv preprint arXiv:2603.25562*, 2026a.

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. *arXiv preprint arXiv:2609.04172*, 2026b. URL https://arxiv.org/abs/2609.04172.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, Ning Ding, and Zhiyuan Liu. JustRL: Scaling a 1.5B LLM with a simple RL recipe. *arXiv preprint arXiv:2512.16649*, 2025a.

Haonan He, Haodi Lei, Yun Luo, Haoran Zhang, Shunkai Zhang, Yizhuo Li, Shengji Tang, Zhilin Wang, Runzhe Zhan, Lei Bai, Ganqu Cui, Fangchen Yu, Yafu Li, Peng Ye, Ning Ding, and Yu Cheng. SimpleOPD: Simple tokenizer-agnostic on-policy distillation for long-context reasoning. *arXiv preprint arXiv:2608.14277*, 2026. URL https://arxiv.org/abs/2608.14277.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025b.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M. Ponti, and Ivan Titov. Blending supervised and reinforcement fine-tuning with prefix sampling. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2507.01679.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. URL https://aclanthology.org/D16-1139.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026. URL https://arxiv.org/abs/2603.11137.

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

Baohao Liao, Hanze Dong, Christof Monz, Xinxing Xu, Li Dong, and Furu Wei. Multi-turn on-policy distillation with prefix replay. *arXiv preprint arXiv:2607.04763*, 2026. URL https://arxiv.org/abs/2607.04763.

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

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. UFT: Unifying supervised and reinforcement fine-tuning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.16984.

Yanjiang Liu, Jie Lou, Xinyan Guan, Yuqiu Ji, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Xing Yu, and Yaojie Lu. Your teacher can't help you here: Combating supervision fidelity decay in on-policy distillation. *arXiv preprint arXiv:2605.30833*, 2026b. URL https://arxiv.org/abs/2605.30833.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying OPD: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026. URL https://arxiv.org/abs/2604.08527.

Mathematical Association of America. American Invitational Mathematics Examination (AIME) 2024, 2024. URL https://maa.org/maa-invitational-competitions/.

Mathematical Association of America. American Invitational Mathematics Examination (AIME) 2025, 2025. URL https://maa.org/maa-invitational-competitions/.

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

Zikun Qu, Min Zhang, Mingze Kong, Zhiwei Shang, Zhengyu Chen, Yikun Ban, Shuang Qiu, and Zhongxiang Dai. SPOT: Sparse probing and outcome calibration for on-policy distillation. *arXiv preprint arXiv:2608.04419*, 2026. URL https://arxiv.org/abs/2608.04419.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Yihan Wang, Zhong Guan, Haoran Sun, Jiale Huang, Likang Wu, and Hongke Zhao. From trajectories to prefixes: Reusing teacher trajectories via replayed prefixes and online continuation. *arXiv preprint arXiv:2607.19395*, 2026. URL https://arxiv.org/abs/2607.19395.

Yan Xie, Sijie Zhu, Tiansheng Wen, Bo Chen, and Yifei Wang. On the position bias of on-policy distillation. *arXiv preprint arXiv:2606.22600*, 2026. URL https://arxiv.org/abs/2606.22600.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026. URL https://arxiv.org/abs/2606.01249.

Haolei Xu, Xiaowen Xu, Haiwen Hong, Zixuan Ni, Hongxing Li, Yiwen Qiu, Weiming Lu, and Yongliang Shen. Pass the baton: Trajectory-relayed on-policy distillation. *arXiv preprint arXiv:2607.26057*, 2026. URL https://arxiv.org/abs/2607.26057.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, and Gang Chen. OPRD: On-policy representation distillation. *arXiv preprint arXiv:2606.06021*, 2026. URL https://arxiv.org/abs/2606.06021.

Wang Yang, Xiang Yue, Vipin Chaudhary, and Xiaotian Han. Speculative thinking: Enhancing small-model reasoning with large model guidance at inference time. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.12329.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler, II, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1276. URL https://aclanthology.org/2026.findings-acl.1276/.

Qingfei Zhao, Huan Song, Shuyu Tian, Jiawei Shao, and Xuelong Li. Prefix-guided on-policy distillation: Mining golden trajectories from rollouts. *arXiv preprint arXiv:2606.21994*, 2026. URL https://arxiv.org/abs/2606.21994.

Ziheng Zhou, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation. *arXiv preprint arXiv:2605.27028*, 2026. URL https://arxiv.org/abs/2605.27028.

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

A Implementation Details

Prefix placement. The prefix p (Section 3) is placed as the first response tokens with response mask 0. Thus, the student generates with p in context, but only its own tokens enter $\mathcal{L}_{\mathrm{OPD}}(y)$ (Equation 6). A k-token prefix leaves the student $27{,}648 - k$ tokens of the rollout cap, never below $19{,}456$ since $k \leq 8{,}192$. With the available teacher, prefixed examples have on average a $3{,}751$-token smaller budget than unprefixed ones (Table 12). No reported run equalizes it, for either teacher, so it is a confound, not a control. We call RH the arm (method variant) that uses the prefix as context only (λ=0). Therefore, OPD and RH differ in both prefix presence and training-time generation budget. Conversely, all λ arms share exactly the same prefixed budget.

Prefix coverage. Of our $17{,}917$ training problems, 86.5% receive a prefix with the available teacher (mean length $3{,}751$ tokens), and the rest train as vanilla OPD. A batch thus mixes prefixed and unprefixed examples, and at most 1.4% of micro-batches have no prefix. With the matched teacher, 83.0% receive a prefix and 17.0% do not (mean length $3{,}196$ tokens; Table 12). We admit traces with a wrong final answer. For each problem, a PS-OPD arm and its RH counterpart use the same teacher response, so their difference never reflects a different trace. The prefix is cut before the first `</think>`, and none of the available teacher's $15{,}492$ prefixes or the matched teacher's $14{,}875$ contains the tag. At evaluation, every trained model receives only the prompt x.

Top-k distillation loss. Each objective approximates the forward or reverse KL of Section 2 on a support $S_t \subseteq \mathcal{V}$, the teacher's ($S_t^T$) or the student's ($S_t^\theta$) 128 most likely tokens at c_t. With the ordered pair (ν, μ) fixed by the direction ($\nu = \pi_T$ forward, $\nu = \pi_\theta$ reverse), $\nu_v = \nu(v \mid c_t)$ and $\mu_v = \mu(v \mid c_t)$, the per-token loss is

$$d(\nu \| \mu; S_t) = \max\big(0, \textstyle\sum_{v \in S_t} \nu_v \log(\nu_v / \mu_v)\big). \tag{8}$$

Unless stated otherwise, OPD here is the forward-KL loss $d_{\mathrm{FKL}}(c_t; \theta) = d(\pi_T \| \pi_\theta; S_t^T)$. For d_{FKL}, one teacher forward pass returns the log-probabilities of its 128 most likely tokens at every position. The student's log-probabilities at these tokens come from a log-softmax over its whole output layer ($151{,}936$ entries). Neither side is renormalized, and the only clamp is at zero (Equation 8). Both distributions and the rollouts use temperature 1. With the available teacher, the 128 tokens hold 99.6% (OPD) and 99.7% (PS-OPD) of the teacher's probability at student-generated positions, averaged over the 50 training steps of the seed-1 runs.

Loss normalization. Equations 6 and 7 are written per example, but in training each term is normalized separately inside each gradient-accumulation micro-batch of about three rollouts. A training step's 256 rollouts are split over D=2 data-parallel ranks (Appendix A.1). A workload-balancing partition assigns each rank's rollouts to $M = \lceil T / 32{,}768 \rceil$ micro-batches, where T is the larger rank's count of prompt, prefix and suffix tokens. For micro-batch b with generated suffix tokens S_b and prefixed rollouts P_b, the loss is

$$\ell_b = \frac{1}{|S_b|} \sum_{(i,j) \in S_b} d_{ij} + \frac{\lambda}{|P_b|} \sum_{i \in P_b} \mathcal{L}_{\mathrm{SFT}}(p_i),$$

with $\mathcal{L}_{\mathrm{SFT}}(p_i)$ averaged over the tokens of p_i and the second term 0 when $P_b = \emptyset$. Gradients are summed over micro-batches, averaged over ranks, and clipped. On every logged training step with no prefix-free micro-batch, the logged loss equals M times the sum of the two logged terms (maximum error 4.4×10^{-7} over 479 training steps). Since 0.24–1.36% of micro-batches per arm have no prefix, the realized ratio of the two terms' total weights is $\lambda(1-e) = 0.986$–$0.998\,\lambda$, where e is that fraction. This is not a batch-level mean, as an average prefix token carries about 2λ the weight of an average suffix token. Our trainer supports global token-mean normalization, but no reported arm uses it.

Optimization. We use AdamW with $\beta = (0.9, 0.999)$, weight decay 0.01, global gradient-norm clipping at 1.0, and a constant learning rate without warmup. It is 10^{-5} in the main study, and 10^{-6} and 2×10^{-5} in the single-seed checks of Appendices H and B.3. Clipping was active at every logged training step of every available-teacher run (978 steps, pre-clip norms 3.4–115), which removes the summed gradient's scaling with M. Clipping fixes the norm of the gradient passed to AdamW, not of AdamW's parameter update, whose moment history also depends on λ.

Evaluation settings. Every checkpoint is scored at avg@16 with a $32{,}768$-token budget ($31{,}744$ response tokens). The available teacher and its students are decoded at temperature 1.0 and top-

Table 3: **Where each experiment and dataset was produced.** GPUs per run: student + teacher for OPD-family runs, then (after the semicolon) off-policy distillation. On the AMD cluster one GPU is one MI250X GCD, and the Skywork data are byte-identical copies or same-builder rebuilds. Two-way prefixes are cut from NVIDIA evaluation responses, and continuations and unprefixed references ran on AMD.

experiment or artifact	**cluster**	**GPUs per run**
Data		
Skywork traces (T=1.0, top-p 1.0) and prefix pool (Table 12)	NVIDIA	1 per job, 16 jobs
JustRL traces (T=1.0, top-p 1.0) and prefix pool (Table 12)	AMD	1 per shard, 16 shards in 2 jobs
Skywork off-policy distillation and TrOPD data	NVIDIA	built from its traces
JustRL off-policy distillation and TrOPD data	AMD	built from its traces
Training		
Five-arm comparison (Table 4), its per-seed and paired tables; fresh seeds 4–6	NVIDIA	$4 + 4$; 4
Controls and ablations (λ, prefix length, learning rate, long off-policy distillation), cost and speed; Figures 9 and 10	NVIDIA	$4 + 4$; 4
Replication of the available pair (Table 13; its forward-KL, teacher-support row is the five-arm runs)	NVIDIA	$4 + 4$
Table 1; objective grid (Table 2, Appendix C); Figures 3, 7 and 8	AMD	$8 + 4$; 4
Evaluation		
Every checkpoint, on its training cluster; the untrained student and both teachers on both	both	4 (NVIDIA), 8 (AMD), one shard per GPU
Two-way handoff (Figure 2, Table 26)	both	8 (continuations)

p 1.0. The matched teacher and its students use temperature 0.6 and top-p 0.95. The untrained student's rows in Tables 1 and 2 use 0.6 and 0.95 for both teacher-student pairs (23.32% Avg.). So does its dashed line in both panels of Figures 3 and 7, even beside the available teacher's arms decoded at 1.0 and 1.0. Training rollouts always use temperature 1.0 and top-p 1.0.

Checkpoints and truncation logging. Main-study arms (10^{-5}) save a checkpoint after each of the first 10 training steps and every 10 thereafter, and the 10^{-6} arms of Appendix H every 10 steps. Figure 4 counts a training rollout as truncated when it reaches the 27,648-token cap, prefix included for PS-OPD. The logged clip ratio (Table 18) equals this share only at training steps whose longest rollout reaches the cap. Otherwise it logs $1/256$, which the figure records as zero.

A.1 Compute and where each experiment ran

Every comparison, including every cost comparison, is made within one cluster, and tables from different clusters are not compared entry by entry (see Table 3). The NVIDIA cluster has four GH200 GPUs (96 GB) per node on a Grace (Arm) host. The AMD cluster has four MI250X GPUs per node, each with two graphics compute dies (GCDs) of 64 GB, and we count one GCD as one GPU. Software versions (NVIDIA / AMD) are vLLM 0.19.1 / 0.20.1, PyTorch 2.10.0 on CUDA 12.9 / ROCm 7.0 (NCCL 2.27.5 / RCCL 2.26.6), transformers 5.8.0 / 4.57.6, Ray 2.55.1 / 2.56.1, and flash-attn 2.8.3 / 2.8.4 (ROCm fork). Both use Python 3.12 and verl 0.9.0.dev with our changes, with vLLM pull request #54335 backported with verl pull request #7806 for the student-top-128 objectives. In an OPD-family run, the student uses two data-parallel ranks of 32,768 tokens each under FSDP, with Ulysses sequence parallelism 2 on four GPUs (NVIDIA) or 4 on eight GCDs (AMD). The teacher runs as four vLLM replicas at tensor parallelism 1 on four GPUs, with batched-token limit 8,192 at memory utilization 0.80 on NVIDIA and 4,096 at 0.75 on AMD. DAPO-Math-17k is byte-identical across clusters, and the scorer and problem exclusion are shared. Each teacher's data were built on the cluster that generated its traces. The untrained student and both teachers are scored on both.

Table 4: Accuracy (avg@16, %) of the main comparison with the available teacher at every checkpoint, three-seed mean (NVIDIA cluster, learning rate 10^{-5}, temperature 1.0, top-p 1.0). Each block starts with the untrained student (step 0) and ends with the teacher (Skywork), each evaluated once with the same decoding. Bold: best arm per training step within a block (ties jointly). Off-policy distillation uses batches of 256 sequences; per-seed Avg. is in Table 5. Not comparable with Table 1, which is from the AMD cluster and decodes its student row at temperature 0.6, top-p 0.95.

training step	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**20**	**30**	**40**	**50**
AIME24 *Student (before training): 26.5*														
PS-OPD (λ=0.3)	**27.8**	**27.6**	**27.8**	19.4	21.0	25.1	**33.8**	**35.6**	**36.3**	**34.4**	**34.9**	34.2	**35.6**	**35.7**
OPD	24.9	25.4	24.0	21.5	21.1	22.3	26.5	27.4	29.1	31.2	33.1	33.3	33.6	33.2
Off-policy distillation	25.1	26.1	26.6	**27.7**	26.4	25.8	32.0	29.9	29.6	31.8	30.8	31.3	30.8	34.2
FH (H=4,096)	24.8	26.3	25.8	23.1	22.6	23.3	26.8	27.7	28.6	32.1	33.7	**34.8**	34.4	34.3
TrOPD	26.2	24.4	25.1	26.9	**27.4**	**29.4**	30.3	30.4	33.1	32.1	33.0	31.7	33.3	32.6
Teacher (Skywork): 69.0														
AIME25 *Student (before training): 19.4*														
PS-OPD (λ=0.3)	**22.1**	21.8	22.7	19.0	20.1	22.2	**26.7**	**26.5**	**25.6**	**27.4**	**26.5**	26.1	**28.0**	**27.3**
OPD	20.3	20.7	22.4	20.0	20.9	21.2	22.5	23.1	25.3	24.5	25.1	25.6	25.7	26.5
Off-policy distillation	21.0	**23.0**	22.4	22.5	21.9	22.4	25.1	24.0	24.5	25.3	25.1	25.8	24.2	26.3
FH (H=4,096)	21.5	21.0	21.2	21.0	22.0	22.0	23.9	25.6	24.6	24.4	26.4	**26.3**	25.9	24.9
TrOPD	21.2	21.3	**23.0**	**24.2**	**24.3**	**23.8**	24.4	24.7	24.4	25.2	26.1	26.2	26.6	26.9
Teacher (Skywork): 53.3														
AIME26 *Student (before training): 11.4*														
PS-OPD (λ=0.3)	**14.8**	**16.2**	16.1	11.6	12.4	13.9	**22.0**	**23.0**	**21.3**	**23.6**	**20.9**	**22.8**	21.5	**23.9**
OPD	14.6	14.2	15.0	10.3	12.3	12.9	14.3	15.6	18.5	18.1	19.2	19.0	21.5	20.8
Off-policy distillation	13.0	**16.2**	**16.3**	15.2	13.5	14.4	19.5	16.6	17.4	19.4	17.9	20.0	19.5	20.0
FH (H=4,096)	13.9	15.7	13.2	11.9	12.9	12.5	14.7	15.1	17.0	19.0	19.5	21.6	**22.2**	20.4
TrOPD	14.2	14.0	14.2	**15.9**	**15.2**	**16.6**	17.2	18.2	17.6	19.3	20.5	20.5	18.2	19.4
Teacher (Skywork): 60.8														
Avg. *Student (before training): 19.1*														
PS-OPD (λ=0.3)	**21.6**	**21.8**	**22.2**	16.7	17.8	20.4	**27.5**	**28.4**	**27.7**	**28.5**	**27.4**	**27.7**	**28.3**	**28.9**
OPD	19.9	20.1	20.5	17.3	18.1	18.8	21.1	22.0	24.3	24.6	25.8	26.0	26.9	26.8
Off-policy distillation	19.7	**21.8**	21.8	21.8	20.6	20.9	25.5	23.5	23.8	25.5	24.6	25.7	24.8	26.8
FH (H=4,096)	20.1	21.0	20.1	18.7	19.1	19.3	21.8	22.8	23.4	25.2	26.5	27.5	27.5	26.5
TrOPD	20.5	19.9	20.8	**22.4**	**22.3**	**23.2**	24.0	24.5	25.1	25.5	26.5	26.1	26.0	26.3
Teacher (Skywork): 61.0														

B Additional Experiments

All experiments here use the available teacher (Skywork) on the NVIDIA cluster. Each trained method, control or prefix-loss weight λ is an *arm*. A difference is *unresolved* when its seed-paired confidence interval includes zero.

B.1 Accuracy at every checkpoint

PS-OPD has a higher Avg. accuracy than all four baselines at every checkpoint from training step 7 on. Table 4 gives three-seed means per benchmark, with the untrained student and teacher, and Table 5 each seed's Avg. score. Several gaps are unresolved, e.g., those to FH at training steps 20, 30 and 40 (0.9, 0.2 and 0.8 points).

B.2 Seed-paired window comparisons

Table 6 tests which leads of PS-OPD (λ=0.3) exceed seed variation, pairing seeds after averaging each over a window. The lead over OPD is resolved over the first ten training steps, the full later window and the last checkpoint. Three-seed Avg. window means over steps 1–10 are 23.25% (PS-OPD), 22.81% (TrOPD), 22.49% (off-policy distillation), 21.13% (FH) and 20.66% (OPD). Over steps 20–50, they are 28.11% (PS-OPD), 27.03% (FH), 26.38% (OPD), 26.23% (TrOPD) and 25.49% (off-policy distillation). At step 50, PS-OPD reaches 28.95%. The windows were chosen

Table 5: Avg. accuracy (%) per seed and training step (learning rate 10^{-5}, decoding as in Table 4). Off-policy distillation uses batches of 256 sequences.

arm	seed	1	2	3	4	5	6	7	8	9	10	20	30	40	50
Off-policy distillation	1	20.8	20.9	20.9	21.4	20.4	20.1	25.1	22.3	23.5	25.4	24.4	25.8	24.4	26.1
	2	19.1	22.0	22.1	23.1	20.6	20.4	26.1	23.5	24.3	25.9	25.3	25.8	24.6	26.1
	3	19.2	22.4	22.3	20.9	20.8	22.0	25.4	24.7	23.7	25.1	24.2	25.5	25.4	28.1
FH (H=4,096)	1	19.7	20.1	19.1	17.9	19.0	19.1	22.1	23.0	23.4	25.9	26.9	28.6	27.3	27.0
	2	20.9	20.7	19.5	19.3	19.5	19.8	22.4	22.1	24.1	24.4	27.5	26.7	26.8	26.9
	3	19.6	22.2	21.5	18.7	18.9	19.0	21.0	23.3	22.7	25.3	25.3	27.3	28.5	25.7
TrOPD	1	20.4	19.7	20.1	22.9	21.4	23.8	24.9	24.3	24.7	25.7	26.5	24.8	25.2	25.5
	2	20.3	19.3	20.5	22.0	22.5	23.4	23.9	24.7	25.9	25.4	27.2	27.4	26.2	26.2
	3	20.8	20.6	21.8	22.1	23.0	22.5	23.1	24.4	24.6	25.5	25.9	26.1	26.6	27.2
OPD	1	19.3	20.9	20.3	17.8	18.3	19.5	21.6	21.7	24.4	25.3	26.2	25.0	27.7	27.6
	2	19.5	19.9	21.3	16.6	18.3	18.5	20.6	22.7	25.2	24.6	26.3	27.6	25.5	26.1
	3	20.9	19.5	19.8	17.3	17.8	18.3	21.1	21.6	23.2	23.9	24.9	25.3	27.6	26.8
RH (λ=0)	1	20.7	21.2	20.6	13.0	15.8	16.8	26.3	26.3	28.8	26.8	27.8	27.0	27.7	26.6
	2	20.5	21.6	21.1	12.6	14.8	16.8	25.3	26.7	28.4	28.4	27.5	27.2	26.5	26.5
	3	20.5	19.7	22.3	12.6	13.4	16.1	23.7	27.1	26.6	27.5	26.8	28.5	28.1	28.3
λ=0.3	1	20.6	22.3	23.0	17.1	18.6	21.0	27.1	28.6	27.5	27.7	27.1	27.5	28.4	29.0
	2	22.1	22.6	21.6	17.2	18.9	20.1	27.6	28.7	26.6	29.2	27.7	27.3	28.3	28.4
	3	22.1	20.7	21.9	15.7	16.0	20.2	27.7	27.8	29.0	28.5	27.5	28.4	28.4	29.4
λ=1	1	21.2	21.3	21.7	22.0	21.3	23.6	27.7	27.2	27.7	28.7	27.0	27.3	28.5	27.5
	2	20.3	20.7	22.1	21.0	20.9	22.0	25.6	27.3	27.1	26.8	26.5	27.7	28.0	27.3
	3	20.2	21.1	22.0	21.0	20.6	21.5	26.3	27.3	27.5	26.8	25.9	25.8	27.2	28.2
λ=3	1	21.7	22.3	21.5	22.8	23.2	22.6	23.6	22.6	24.5	26.0	24.9	27.4	26.3	26.1
	2	20.2	21.6	21.1	23.1	22.3	23.8	21.7	22.0	23.7	26.3	27.1	28.4	26.7	27.1
	3	21.8	20.6	21.7	23.0	23.0	22.5	22.5	22.3	24.8	26.5	26.1	26.8	28.0	26.3

Table 6: PS-OPD minus each arm in Avg., paired within seeds 1–3: mean difference in points and 95% t interval (two degrees of freedom). **Bold**: the interval excludes zero ($|t| > 4.3$, two-sided $p < 0.05$, unadjusted). FH uses H=4,096, RH is λ=0, Off-pol. is off-policy distillation. Per-benchmark differences with OPD: Table 16.

minus	steps 1–10	steps 20–30	steps 20–50	step 50
OPD	**+2.59** [+2.22, +2.97]	+1.71 [−1.14, +4.56]	**+1.74** [+0.58, +2.89]	**+2.12** [+0.59, +3.64]
Off-pol.	+0.76 [−0.43, +1.95]	**+2.42** [+0.86, +3.97]	**+2.63** [+2.24, +3.01]	**+2.15** [+0.07, +4.23]
FH	**+2.12** [+1.25, +2.99]	+0.55 [−2.08, +3.17]	+1.08 [−0.42, +2.58]	+2.40 [−0.40, +5.19]
TrOPD	+0.45 [−0.26, +1.15]	+1.27 [−1.01, +3.54]	**+1.88** [+0.28, +3.48]	**+2.67** [+0.85, +4.49]
RH	**+1.85** [+1.49, +2.21]	+0.11 [−0.37, +0.59]	**+0.73** [+0.05, +1.41]	**+1.81** [+0.12, +3.49]
λ=1	−0.43 [−1.66, +0.79]	+0.87 [−1.79, +3.52]	+0.86 [−0.76, +2.48]	**+1.27** [+0.67, +1.88]
λ=3	+0.41 [−0.60, +1.43]	+0.77 [−1.53, +3.08]	+1.34 [−0.26, +2.94]	**+2.45** [+0.07, +4.84]

after scoring and $\lambda = 0.3$ was selected by its step-50 Avg. mean on these seeds, so every entry is exploratory (multiple-testing correction in Appendix J).

B.3 Longer off-policy distillation and a doubled learning rate

Off-policy distillation for 200 training steps. Four times the training steps bring off-policy distillation within an unresolved 0.8 Avg. points of PS-OPD at step 50, a checkpoint whose weight was selected on these seeds, but do not show whether the two converge. We run the same recipe (batches of 256 sequences, constant learning rate 10^{-5}) to 200 training steps, roughly three epochs, scoring seeds 1–3 every 40 steps, plus one seed-1 run at 2×10^{-5} (Table 7). In Avg., the three-seed mean reaches 28.14% at training step 200, an unresolved 0.80 points below PS-OPD at step 50 (28.95%, $t = -2.2$) and 1.31 points above OPD at step 50. This last difference is nearly equal across seeds, so its $t = 45.3$ reflects a standard deviation of 0.05 points that three seeds cannot estimate reliably, not a precise effect. Equal training steps are not equal cost, since an off-policy distillation step generates no rollouts and needs no teacher scoring (Appendix B.4).

Variation between same-seed runs. Single checkpoints of nominally identical runs can differ by about 2 Avg. points. The longer runs' first 40 training steps repeat the off-policy distillation arm of Table 4, with the same recipe and seeds. The sequential control's warm start (Appendix K.1) even matches their training loss to within 4×10^{-5} over the first 10 steps. Yet at step 40 they score 1.12

Table 7: Off-policy distillation (batches of 256 sequences) to 200 training steps: accuracy (%), three-seed mean at 10^{-5} with across-seed standard deviation in points. Last row: a single seed-1 run at 2×10^{-5}.

training step	**40**	**80**	**120**	**160**	**200**
AIME24	32.43 (0.67)	34.65 (1.18)	32.57 (0.32)	34.72 (1.73)	34.86 (1.26)
AIME25	25.42 (1.44)	26.18 (1.39)	27.50 (0.91)	25.35 (1.20)	27.29 (1.46)
AIME26	19.97 (1.08)	21.48 (0.87)	20.91 (1.94)	24.57 (0.65)	22.27 (1.99)
Avg.	25.94 (0.95)	27.44 (0.72)	26.99 (0.99)	28.21 (0.35)	28.14 (0.82)
Avg., seed 1, 2×10^{-5}	24.9	28.2	27.4	28.8	29.8

Table 8: Learning-rate screen (seed 1): accuracy (%); PS-OPD is $\lambda = 0.3$. Steps 1–10 come from runs saved at every training step, steps 20–50 from separate runs saved every ten; at 2×10^{-5} OPD's is a rerun, because the original run's step-50 checkpoint was not kept. Step 10: the run saved every step, and in parentheses the run saved every ten.

		Avg.				**step 50**		
arm	**rate**	**steps 1–10**	**step 10**	**steps 20–50**	**step 50**	**AIME24**	**AIME25**	**AIME26**
OPD	10^{-5}	20.91	25.3	26.62	27.6	33.5	27.9	21.3
	$2{\times}10^{-5}$	22.55	24.5 (25.1)	24.33	24.7	31.0	22.5	20.7
PS-OPD	10^{-5}	23.35	27.7	27.99	29.0	36.0	27.3	23.7
	$2{\times}10^{-5}$	24.84	25.2 (28.2)	28.59	29.6	36.7	27.9	24.4

Avg. points above that arm on average, 2.45 in seed 2 ($t = 1.6$), and up to 3.1 points on a single benchmark (AIME25, seed 2). This is consistent with evaluation sampling noise, although drift after training step 10 is not excluded.

OPD and PS-OPD at 2×10^{-5} (seed 1). At twice the main learning rate, PS-OPD ($\lambda = 0.3$) stays ahead of OPD in Avg. for seed 1 (Table 8). The higher rate raises the steps-1–10 average of OPD by 1.64 points and of PS-OPD by 1.49. The lead of PS-OPD in this window is thus similar at both rates (+2.29 points at 2×10^{-5}, +2.44 at 10^{-5}). At step 50, the higher rate leaves PS-OPD at 29.6% against 29.0% and OPD at 24.7% against 27.6%. The larger steps-20–50 lead at 2×10^{-5} (4.26 against 1.37 points) therefore comes mostly from lower OPD scores. With one seed, none of these differences is resolved, and two runs of one configuration can disagree by more than several of them. For example, two same-seed runs of PS-OPD at 2×10^{-5} score 25.2% and 28.2% at step 10, a gap of 3.0 points.

B.4 Convergence speed and cost

PS-OPD reaches OPD's final accuracy in fewer training steps and GPU-hours than OPD. Table 9 gives each arm's *crossing*, the first checkpoint at which its accuracy reaches a threshold, and Table 10 the cost of a training step. In Avg., the lower threshold $T_a = 26.8\%$ is the step-50 three-seed mean of both OPD (26.83%) and off-policy distillation (26.80%). Only PS-OPD's three-seed mean reaches the higher $T_c = 28.0\%$ within 50 training steps. On each benchmark, T_a is OPD's step-50 three-seed mean rounded down to 0.1 point (33.1%, 26.4% and 20.8% on AIME24, AIME25 and AIME26). Single-benchmark scores vary more across runs than Avg. (see Caveats), so per-benchmark crossings are descriptive and more fragile.

Training steps to OPD's step-50 accuracy. In Avg., PS-OPD's three-seed mean first reaches T_a at training step 7, as does each seed, and no later checkpoint falls back below T_a minus 1 point. OPD's mean first reaches T_a at the step-40 checkpoint, so somewhere in training steps 31–40 (27–35 GPU-h). Over all 729 resamplings of the three seeds of both arms, PS-OPD's crossing is 7 in 94.5% of cases and 8 otherwise. On fresh seeds 4–6, whose first checkpoint is at step 10, PS-OPD is above T_a there in every seed, while OPD's mean reaches T_a only at 50. A single speed-up ratio would depend on the target, so we report crossing steps and curves instead. For example, the Avg. mean of PS-OPD reaches every target from 23% to 27% at training step 7, while OPD's reaches 24% at step 9, 25% at 20 and 26% at 40.

Table 9: Training steps to the thresholds: first checkpoint at which the three-seed mean reaches T_a (per benchmark and Avg.) or T_c (Avg.), and in parentheses, where different, the *sustained* crossing, after which no checkpoint falls below the threshold minus 1 point. "–": not reached; *: sustained only at the last checkpoint, so unconfirmed. GPU-h: the Avg. T_a crossing times the arm's GPU-h per training step (Table 10); *total* adds the 12.74 GPU-h of teacher traces where used.

	T_a				T_c	GPU-h to Avg. T_a	
arm	**AIME24**	**AIME25**	**AIME26**	**Avg.**	**Avg.**	**training**	**total**
Seeds 1–3							
OPD	30	50*	40	40	–	34.9	34.9
Off-policy distillation	50*	–	–	–§	–	–	–
to 200 training steps†	80	120 (200*)	80	80	160	2.9	15.6
PS-OPD (λ=0.3)	7	7	7	7	8	6.2	18.9
RH (λ=0)	8	9	9	9	–	7.9	20.6
λ=1	7	7	10 (40)	8	–	7.1	19.8
λ=3	10	30	–	30	–	28.4	41.1
FH (H=4,096)	20	–	30	30	–	n/a	n/a
TrOPD	9 (40)	40	–	–	–	–	–
Decoupled control	20 (–)	20	40	40	–	35.4	48.1
Fresh seeds 4–6							
OPD	40	–	50*	50*	–	43.7	43.7
Off-policy distillation	30 (–)	30 (–)	30	30	–	1.1	13.9
PS-OPD (λ=0.3)	10	10	10	10	10	≤8.9	≤21.6

Checkpoints are at training steps 1–10, 20, 30, 40 and 50, so 40 means somewhere in 31–40. Seeds 4–6 have no checkpoint before 10, so 10 there means at or before 10. n/a: FH's measured cost is that of our loss-mask implementation (Table 10, note b). The sequential SFT→OPD control reaches neither threshold, on any benchmark or in Avg. (row omitted). §Avg. mean 26.796% at step 50, 0.004 points below T_a. †Separate 200-step runs scored at steps $40, 80, \ldots, 200$, so 40 means at or before 40.

Which arms cross early. In Avg., early mean crossings occur only with the prefix in the rollout. RH, which conditions on the prefix without imitating it, reaches T_a at training step 9 and $\lambda = 1$ at 8. Conversely, $\lambda = 3$ needs 30, so the prefix is not sufficient at every weight. The decoupled control, which imitates the same prefixes but rolls out without them, reaches T_a only at training step 40 (its mean is 26.52% at step 10 and 26.77% at 20). Its steps-1–10 average is within an unresolved 0.42 points of PS-OPD's (Appendix K.1). However, its mean is above PS-OPD's over training steps 4–6 (by 2.5 Avg. points on average) and below it over 7–10 (by 2.4), in Avg. and on each benchmark. We did not test these window gaps.

Cost. PS-OPD costs 1.5% more per training step than OPD (0.886 against 0.873 GPU-h), although it decodes 33% fewer tokens. The reason is that its training tokens, prefix included and all scored by the teacher, are nearly the same (2.47M against 2.45M). Identical configurations vary in cost by up to about 9% across seeds. At the mean crossings, PS-OPD reaches T_a after $7 \times 0.886 = 6.2$ GPU-h of training and OPD after $40 \times 0.873 = 34.9$, or 18.9 against 34.9 with the 12.74 GPU-h of teacher traces. Off-policy distillation continued to 200 training steps reaches both Avg. thresholds with fewer GPU-hours: T_a at step 80 (2.9 GPU-h, 15.6 with traces) and T_c at 160 (5.8 and 18.5, against 7.1 and 19.8 for PS-OPD). However, reaching T_c takes it 20 times the sequences of PS-OPD's 8 steps.

Accuracy at a common budget. Table 11 compares every arm except FH at fixed GPU-hour budgets, using the per-step means of Table 10 (SFT and OPD segment means for the sequential control). PS-OPD is above OPD in Avg. at every tabulated budget. At 20 GPU-h, PS-OPD (8.2 training steps) scores 28.2% against OPD's 25.8% (22.9 steps). At 30 GPU-h the lead is 1.1 points ($t = 11.8$), and at 40 GPU-h an unresolved 0.9 ($t = 1.8$). At 10 GPU-h, below the trace cost, only OPD has a value (24.8%). Off-policy distillation's 50- and 200-step runs end at 14.6 and 19.9 GPU-h, traces included. On fresh seeds 4–6, PS-OPD scores 28.3% and 27.9% in Avg. at 30 and 40 GPU-h, against OPD's 25.8% and 26.6%. Values between checkpoints are interpolated, and per-step cost is treated as constant, although start-up overhead concentrates in the first training steps.

TrOPD at equal cost. At equal training cost, PS-OPD is above TrOPD, while OPD is unresolved against it. TrOPD's 50 steps cost 29.8 training GPU-h (42.6 with the traces it also uses).

Table 10: Cost per training step (GH200 GPUs). GPU-h per step: Slurm elapsed time times allocated GPUs over n complete 50- or 200-step runs, including start-up, checkpointing and teardown but not evaluation. *Generated*: decoded by the student; *training*: prompt plus response over all training sequences, which OPD-family teachers also score.

		GPU-h per step		tokens per step (M)	
arm	**GPUs**	**mean [range]**	n	**generated**	**training**
OPD	8	0.873 [0.843, 0.900]	6	2.41	2.45
Off-policy distillation	4	0.0371 [0.0368, 0.0374]	6	0	2.17
Off-policy distillation, 200 steps	4	0.0360 [0.0357, 0.0363]	3	0	2.17
PS-OPD (λ=0.3)[a]	8	0.886 [0.850, 0.912]	6	1.61	2.47
RH (λ=0)[a]	8	0.874 [0.827, 0.897]	3	1.59	2.45
λ=1[a]	8	0.886 [0.880, 0.893]	3	1.64	2.50
λ=3[a]	8	0.946 [0.932, 0.954]	3	1.69	2.55
FH (H=4,096)[b]	8	0.935 [0.933, 0.938]	3	2.59	2.63
TrOPD[c]	8	0.596 [0.585, 0.611]	3	2.33	2.65
Decoupled control[d]	8	0.885 [0.855, 0.928]	3	2.42	2.55
Sequential SFT→OPD[e]	4, 8	0.695 [0.665, 0.718]	3	1.94	2.40
Teacher traces, one-time[f]	1 per job	12.74 GPU-h in total		∼145M in total	

[a] About 0.83M offline prefix tokens per training step, counted in training but not in generated tokens. [b] Loss on the first 4,096 generated tokens only; generation and teacher scoring still cover the full response. This is the cost of our loss-mask implementation, not of Zhang et al. (2026)'s method, which stops generation at H and so decodes and scores at most H tokens per rollout. We did not measure that cost, so we make no GPU-hour comparison with FH (Tables 9 and 11). [c] Adds 32 teacher-response sequences per training step (288 in total, $\approx$ 0.27M tokens). Its lower cost should not be read as an algorithmic saving. This arm requests the teacher's top-64 log-probabilities (other OPD-family arms: top-128) plus that of the sampled token, and runs from a separate code base. We did not isolate which difference lowers its cost. [d] Adds one cross-entropy sequence per prefixed prompt (at most 32, $\approx$ 0.11M tokens). [e] Ten SFT training steps on 4 GPUs, then 40 OPD steps on 8, averaged over all 50. [f] One Skywork-OR1-Math-7B trace per problem (16,972 traces), one GPU per job; 81.3M of the ∼145M tokens counted exactly, the rest estimated from character counts. Used by every arm above except OPD and FH.

Table 11: Accuracy (%) at a common cumulative GPU-hour budget (seeds 1–3), charging the 12.74 GPU-h of teacher traces to every arm that uses them (FH omitted, Table 10, note b). Each seed's curve is interpolated linearly between scored checkpoints, never extrapolated. Upper block: Avg., three-seed means, with the per-seed range for PS-OPD and OPD. Lower block: seed-paired PS-OPD minus OPD in points, t in parentheses (two degrees of freedom); per-benchmark values are descriptive.

arm	**20 GPU-h**	**30 GPU-h**	**40 GPU-h**
PS-OPD (λ=0.3)	28.2 [28.0, 28.4]	27.5 [27.1, 27.8]	27.8 [27.4, 28.4]
OPD	25.8 [25.0, 26.6]	26.4 [26.2, 26.7]	26.9 [25.8, 27.6]
RH (λ=0)	27.1	27.4	27.5
λ=1	27.3	26.5	27.0
λ=3	22.4	26.1	27.4
TrOPD	25.8	26.2	26.2
Decoupled control	25.5	26.8	26.0
Sequential SFT→OPD	22.4	24.7	25.1
PS-OPD minus OPD			
Avg.	+2.4 (6.0)	+1.1 (11.8)	+0.9 (1.8)
AIME24	+2.6 (2.4)	+1.4 (4.0)	+1.0 (1.8)
AIME25	+1.1 (16.0)	+0.9 (2.4)	+0.1 (0.4)
AIME26	+3.6 (6.8)	+1.0 (7.3)	+1.6 (1.4)

At 29.8 training GPU-h, PS-OPD (33.6 training steps) scores 28.0% in Avg., 1.7 points above TrOPD's 26.3% at step 50 ($t = 5.2$, 95% CI $[+0.3, +3.0]$). OPD (34.1 steps, 26.4%; $+0.1$ points, $[-2.1, +2.2]$) stays unresolved against it, also with the traces charged (42.6 GPU-h, 48.7 OPD training steps, $t = 0.7$). We did not isolate whether TrOPD's lower per-step cost is an algorithmic saving (Table 10, note c).

Uncertainty in GPU-hours to T_a. OPD needs 15.99 GPU-h more than PS-OPD to reach the Avg. T_a, traces included. This advantage keeps its sign at the 2.5% quantile of the resamplings below, but three seeds and this checkpoint grid do not pin down its size. First, resampling each arm's per-run GPU-h per training step (20,000 resamples) gives $[15.4, 16.6]$. Second, resampling the three seeds of both arms (729 combinations) at fixed cost leaves the median at 16.0 but lowers the 2.5% quantile to 7.3. In 11% of these resamples, OPD's mean never reaches T_a within 50 training steps. Finally, also drawing OPD's crossing uniformly within training steps 31–40, where it can lie anywhere, gives a median of 11.2 and a 95% range of $[0.2, 16.2]$.

Caveats. First, for seeds 1–3, PS-OPD, RH, the other λ arms and off-policy distillation join two runs, one saved at every step over steps 1–10 and a separate 50-step run for steps 20–50. OPD, FH, TrOPD and the decoupled control are single runs. At training step 10, same-seed Avg. scores from the two runs differ by at most 1.5 points (mean absolute difference 0.8 points over the 15 arm–seed pairs), and single-benchmark scores by up to 2.5, 2.9 and 3.9 points on AIME24, AIME25 and AIME26. The implied standard deviation of a single checkpoint's Avg. score is roughly 0.6 points from these pairs, and 0.8 points when six further same-seed off-policy distillation reruns are added (0.9–1.3 points for a single benchmark). Therefore, per-seed crossings of T_c, and per-benchmark crossings generally, are fragile. Second, checkpoints skip training steps 11–19, 21–29 and so on, so OPD's Avg. crossing at 40 could lie up to nine training steps (7.9 GPU-h) earlier.

C THE TWO TEACHERS AND THE FOUR OBJECTIVES

This appendix details the grid of Table 2: OPD and PS-OPD with the available teacher (Skywork), the matched teacher (JustRL, RL-finetuned from the student) and four objectives. An arm is one method under one objective. We cover objectives (Appendix C.1), prefix pools (Appendix C.2), the available-teacher grid and its NVIDIA replication (Appendix C.3), support overlap (Appendix C.4), and length and truncation (Appendix C.5). Figure 7 repeats Figure 3 under reverse KL over the student's top-128. Everything here is exploratory.

Each block of results comes from one cluster and is never compared across clusters, since scores do not transfer between them. The matched- and available-teacher grids ran on the AMD cluster, and the available-teacher replication and the five-arm comparison of Appendix B.1 on the NVIDIA cluster. We score two frozen checkpoints on both clusters at temperature 1.0, top-p 1.0. From NVIDIA to AMD, the untrained student goes from 19.09 to 20.09% Avg. ($+1.00$ points) and the available teacher from 61.02 to 60.12% (-0.91) for identical weights at identical decoding. One rises and one falls, so the difference is not a constant offset. (Table 1's 23.32% is the same untrained student on the AMD cluster at temperature 0.6, top-p 0.95.)

C.1 THE FOUR OBJECTIVES

The objectives differ only in KL direction and scored support. Each is Equation 8, without renormalization and clamped at zero, over a support S_t that is the teacher's 128 most likely tokens at c_t (S_t^T) or the student's (S_t^θ). The *forward* objectives sum $\pi_T(v)\log(\pi_T(v)/\pi_\theta(v))$ over S_t and the *reverse* ones $\pi_\theta(v)\log(\pi_\theta(v)/\pi_T(v))$, both distributions being full-vocabulary softmaxes at c_t. All four sample rollouts from the current student, so the grid never changes who writes the text that carries the loss.

Truncation and the zero clamp. Let ν be the distribution that supplies the weights (π_T forward, π_θ reverse) and μ the other, with masses $\nu(S_t)$, $\mu(S_t)$ on S_t and renormalizations $\tilde{\nu}, \tilde{\mu}$ there. The

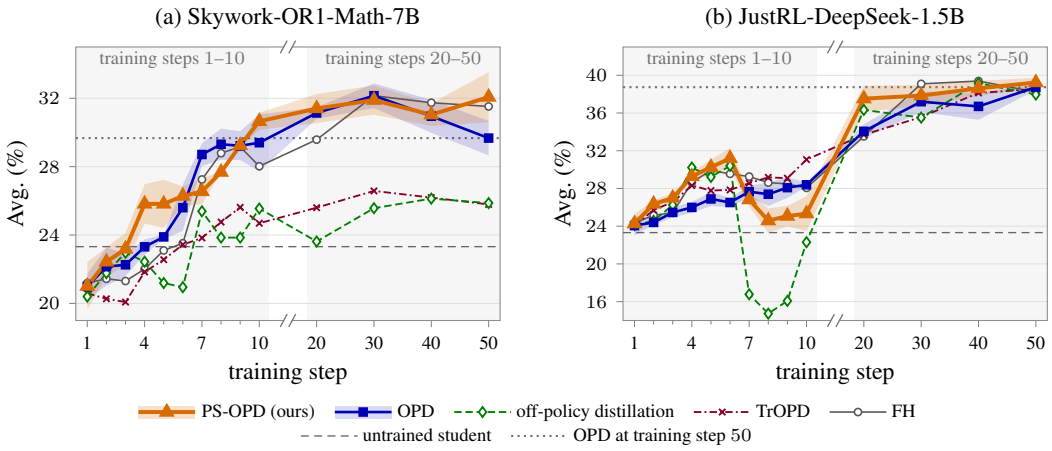


Figure 7: **Convergence under reverse KL over the student's top-**128**.** Figure 3 with the OPD arms (PS-OPD, OPD, FH) trained under reverse KL over the student's top-128; off-policy distillation and TrOPD are Figure 3's runs. Avg. against training step; three-seed means, bands ± 1 standard deviation; dashed, untrained student; dotted, OPD at step 50. Evaluation as in Figure 3: (a) temperature 1.0, top-p 1.0; (b) temperature 0.6, top-p 0.95; the untrained student (23.3%) is at temperature 0.6, top-p 0.95 in both panels, as in Table 1.

sum before the clamp splits exactly, and is negative exactly when the second line holds:

$$\sum_{v \in S_t} \nu_v \log \frac{\nu_v}{\mu_v} = \nu(S_t)\, D_{\mathrm{KL}}(\tilde{\nu} \| \tilde{\mu}) + \nu(S_t) \log \frac{\nu(S_t)}{\mu(S_t)}, \tag{9}$$

$$\sum_{v \in S_t} \nu_v \log \frac{\nu_v}{\mu_v} < 0 \iff \frac{\mu(S_t)}{\nu(S_t)} > \exp\big(D_{\mathrm{KL}}(\tilde{\nu} \| \tilde{\mu})\big). \tag{10}$$

Since ν flips with the direction, the clamp binds asymmetrically. Since the summands take either sign, the truncated sum is neither a lower nor an upper bound on $D_{\mathrm{KL}}(\nu \| \mu)$. With S_t fixed within a training step and the clamp inactive, the derivative of the per-position value d with respect to the logit of a token $u \notin S_t$ is $\pi_\theta(u)\pi_T(S_t) > 0$ (forward) or $-\pi_\theta(u)(d + \pi_\theta(S_t)) < 0$ (reverse). Thus, a descent step lowers the out-of-support logits in the forward objectives and raises them in the reverse ones. Renormalizing the student on S_t would give them zero gradient, and the clamp passes none where the sum is negative, so both choices change the objective. Forward KL over the teacher's top-128 does not log how often the clamp binds. With the available teacher (AMD cluster, three-seed means), the other three objectives clamp under 1% of loss-carrying positions at training step 1 in every arm. At step 50, they clamp 62.5% (OPD) and 26.6% (PS-OPD) under reverse KL on the teacher's support, 31.9% and 5.6% on the student's, and 0.1% and 0.7% under forward KL on the student's.

C.2 The two prefix pools

The two prefix pools differ materially only in the teacher, the resulting coverage and the mean prefix length (Table 12). This is a limitation of the two-pair comparison, not a controlled variable (Appendix I). Both are built as in Section 3.2 on the full DAPO-Math-17k set, in one format, with DAPO's ground truth for every problem. Only 1,445 problems differ in prompt text, each because the builder's answer-line rule applies to it in the one pool that prefixes it. This rule drops DAPO's leading instruction and appends a fixed request to reason step by step and box the final answer.

C.3 The available-teacher grid: replication, seeds and benchmarks

With the available teacher, PS-OPD's Avg. is above OPD's on both clusters for every objective we ran, over training steps 1–10 and 20–50. The seed-paired interval excludes zero everywhere except under reverse KL over the student's top-128 (Tables 13 and 16). Table 13 repeats the grid on the NVIDIA cluster with the same teacher, student, prefix pool, budget, objectives, seeds and

Table 12: **The two prefix pools.** Same procedure for both: one trace per problem on the same 17,917 problems, cut at $r = 0.5$ of the reasoning span, cap 8,192 tokens, coverage 1.0, shared coverage seed. Traces are used regardless of final-answer correctness, and prefix lengths are re-tokenized counts. Traces were generated on the AMD (matched teacher) and NVIDIA (available teacher) clusters (Appendix A.1).

	matched teacher	**available teacher**
trace source	`JustRL-DeepSeek-1.5B`	`Skywork-OR1-Math-7B`
trace sampling	T=1.0, top-p 1.0	T=1.0, top-p 1.0
traces generated (one per problem)	16,972	16,972
correct / wrong / too short / truncated	10,447 / 3,798 / 2,077 / 595	11,530 / 3,202 / 2,196 / 44
prefixed rows (of 17,917)	14,875 (83.0%)	15,492 (86.5%)
prefix length, mean / median	3,195.6 / 2,951	3,751.2 / 3,169
prefix length, min / max	57 / 8,192	233 / 8,192
prefix source: correct / wrong / ungraded	70.1% / 23.8% / 6.1%	74.3% / 19.3% / 6.4%
prefixes containing `</think>`	0	0

Table 13: **Replication of the available-teacher grid on the NVIDIA cluster.** Avg. (avg@16, %) for seeds 1–3 at learning rate 10^{-5}, temperature 1.0, top-p 1.0. "OPD, min": OPD's lowest scored checkpoint over training steps 20–50 (step in parentheses). Δ: PS-OPD minus OPD paired within seed, **bold** if its seed-paired 95% interval excludes zero (per benchmark: Table 16). Reverse KL over the student's top-128 was not run. †Later lead preregistered and tested on three fresh seeds (Appendix K.2).

	Avg. at step 50			**Δ (PS-OPD − OPD)**		
objective (KL direction, scored top-128)	PS-OPD	OPD	OPD, min	1–10	20–50	50
Available teacher, NVIDIA cluster	*student* 19.1		*teacher* 61.0			
forward, teacher†	28.9	26.8	25.8 (20)	**+2.59**	**+1.74**	**+2.12**
reverse, teacher	32.9	21.6	21.6 (50)	**+0.70**	**+5.46**	**+11.35**
forward, student	28.3	27.2	25.0 (20)	**+2.32**	**+2.13**	+1.09

evaluation protocol, read against its own reference scores (Section 4.1). Its row for forward KL over the teacher's support is the five-arm comparison's own runs (Table 4) and the paper's one preregistered objective (Appendix K.2). The other two rows are separate runs. For the AMD cluster, Tables 14, 15 and 16 give every seed and checkpoint in Avg., per-benchmark means over both step ranges, and seed-paired differences per benchmark (the last also for NVIDIA). Every score here is at temperature 1.0, top-p 1.0, except the untrained student's AMD reference.

C.4 Top-128 support overlap

Support overlap is only a descriptive diagnostic. Table 17 gives this overlap, logged at every training step. At each student-sampled position that carries loss, it is the fraction of the 128 scored token ids that also lie in the student's own top-128 over the full vocabulary, averaged over the positions of each micro-batch and then over micro-batches. It is defined only for the teacher's support and differs from the top-128 probability *mass* of Appendix A.

First, overlap does not order the teachers as transfer does. At the first training step, from the same initial student, the available teacher's top-128 overlaps the student's *more* (0.651) than the matched teacher's (0.606; seed-paired difference $+0.044$, $[+0.031, +0.058]$). Second, it is not a quality measure. With the available teacher, the prefix lowers it from 0.651 to 0.628 at training step 1 (-0.023, $[-0.028, -0.018]$). Under forward KL, PS-OPD stays below OPD at all 50 training steps, by 0.087 over steps 20–50 ($[-0.095, -0.080]$), yet its accuracy is higher over steps 1–10 and 20–50. Third, it moves in different directions when lengths explode. Under reverse KL on the available teacher's support, OPD's overlap goes $0.651 \rightarrow 0.286$ over 50 training steps as its accuracy falls to 21.7%, while PS-OPD holds 0.508 at 32.4%. There the fall tracks truncation, with per-seed correlations of -0.99 to -1.00 with the evaluation truncation rate over the 14 scored checkpoints on both clusters (Table 17 shows the AMD cluster only). Conversely, under forward KL on the matched teacher's support, OPD's overlap dips from 0.655 at training step 5 to 0.557 at step 10

Table 14: **Available-teacher grid, Avg. per seed.** AMD cluster, avg@16 accuracy (%) at learning rate 10^{-5}; columns are training steps. Scores are at temperature 1.0, top-p 1.0. The available teacher scores 60.12% and the untrained student 23.32% (at temperature 0.6, top-p 0.95, as in Table 1).

arm	seed	1	2	3	4	5	6	7	8	9	10	20	30	40	50
forward, teacher															
PS-OPD	1	20.9	21.8	21.3	16.7	17.5	19.7	27.7	28.7	27.9	27.9	27.5	27.0	27.3	30.2
	2	21.2	21.8	22.1	16.4	18.2	21.1	27.3	28.9	28.2	28.6	27.9	28.3	27.8	26.2
	3	20.3	23.1	21.3	15.1	16.9	18.3	24.3	27.4	27.5	27.2	27.7	27.6	28.2	29.1
OPD	1	20.7	21.3	20.1	17.8	17.8	18.9	20.7	21.9	22.8	24.3	25.4	24.9	25.1	26.6
	2	20.4	20.5	21.5	17.0	18.1	18.8	20.5	21.0	22.7	23.9	25.1	26.1	25.1	26.3
	3	20.1	20.5	21.2	17.7	17.2	18.8	21.2	21.3	23.1	23.6	25.6	25.2	25.7	26.5
reverse, teacher															
PS-OPD	1	20.3	21.9	24.8	25.5	24.4	26.0	26.7	27.5	30.9	30.9	32.3	32.2	32.7	31.7
	2	21.3	22.6	23.5	26.2	25.4	25.2	27.6	27.2	29.8	29.9	31.3	31.7	31.4	32.6
	3	21.1	23.6	22.8	27.2	24.8	27.0	27.1	27.4	29.4	30.7	31.2	31.2	34.1	32.7
OPD	1	20.2	21.0	22.6	22.2	22.5	24.7	27.3	29.0	29.1	28.4	30.1	29.5	26.1	20.8
	2	20.9	21.9	22.3	22.7	22.9	25.3	29.0	28.4	29.1	28.9	29.9	30.6	28.0	22.6
	3	20.4	20.4	22.9	22.7	24.5	24.4	29.8	28.9	28.4	28.6	28.7	30.0	24.3	21.6
forward, student															
PS-OPD	1	20.0	21.9	21.2	17.4	17.8	20.1	27.3	27.7	28.0	29.6	27.8	29.6	26.3	29.2
	2	20.5	23.4	22.5	16.9	17.1	19.7	27.0	28.0	27.0	28.2	26.8	27.3	28.3	27.5
	3	21.5	21.6	23.0	15.3	17.6	18.8	27.7	26.8	28.9	29.3	27.5	28.3	27.7	28.7
OPD	1	20.4	21.2	21.0	18.2	18.0	18.7	21.6	22.0	23.0	24.4	25.7	26.0	27.0	26.3
	2	20.5	20.7	22.0	18.2	18.1	18.9	20.9	21.8	22.6	23.8	23.3	26.1	24.5	27.1
	3	19.9	19.4	20.2	18.2	16.7	18.3	20.8	20.9	23.5	23.0	24.4	26.4	25.5	26.9
reverse, student															
PS-OPD	1	22.1	23.3	24.2	26.4	27.2	26.8	26.5	27.7	29.1	30.9	30.7	31.3	31.2	31.6
	2	19.5	21.6	22.9	26.5	24.4	26.3	27.3	27.7	29.3	30.1	32.4	31.5	30.3	30.9
	3	21.5	22.3	22.4	24.5	25.8	25.7	25.8	27.7	29.3	31.0	31.1	32.9	31.6	33.7
OPD	1	20.4	22.6	22.7	22.8	24.0	26.8	29.2	29.6	29.9	31.2	31.3	31.5	31.4	30.6
	2	20.8	20.9	21.5	23.5	23.7	25.8	29.0	28.3	29.5	28.0	31.3	32.9	31.6	29.9
	3	21.8	23.0	22.6	23.7	23.9	24.2	28.0	30.1	28.3	28.9	30.7	32.0	29.8	28.6

Table 15: **Available-teacher grid per benchmark.** AMD cluster, avg@16 accuracy (%), three-seed means over training steps 1–10 and over steps 20, 30, 40 and 50 ("20–50"); step 50 is in Table 2. The untrained student scores 28.8, 23.3 and 17.9 (Avg. 23.3; temperature 0.6, top-p 0.95, as in Table 1), and the available teacher 68.1, 51.7 and 60.6 (Avg. 60.1). These means are descriptive; seed-paired differences are in Table 16.

	AIME24		AIME25		AIME26		Avg.	
arm	1–10	20–50	1–10	20–50	1–10	20–50	1–10	20–50
forward, teacher								
PS-OPD	28.1	34.5	23.7	27.1	16.7	22.0	22.8	27.9
OPD	25.5	31.9	21.8	25.5	14.2	19.4	20.5	25.6
reverse, teacher								
PS-OPD	32.7	40.3	25.8	29.9	19.3	26.1	26.0	32.1
OPD	31.2	32.8	25.1	26.6	18.6	21.1	25.0	26.8
forward, student								
PS-OPD	28.3	34.9	23.9	27.4	17.0	21.4	23.1	27.9
OPD	25.8	32.2	21.8	25.3	14.1	19.8	20.6	25.8
reverse, student								
PS-OPD	32.7	40.2	25.6	29.3	19.3	25.3	25.9	31.6
OPD	31.9	39.3	25.7	28.8	18.9	24.8	25.5	31.0

as its length explosion begins, then climbs to 0.904 at step 30 while most of its training rollouts still reach the cap. These readings use moving student rollouts, so changes confound policy, visited contexts and sampled problems. We report them as descriptive diagnostics, not as a mechanism.

C.5 Response length and truncation

With the matched teacher, OPD runs into the length cap mid-training, while PS-OPD does not. With the matched teacher, evaluated at temperature 0.6, top-p 0.95 as in Table 1, OPD becomes unstable under the main objective (forward KL over the teacher's top-128). Its length explosion starts at

Table 16: **PS-OPD minus OPD per benchmark, available teacher.** Mean difference in points, paired within seeds 1–3, and its 95% t interval (two degrees of freedom; an interval excluding zero is $|t| > 4.3$, two-sided $p < 0.05$ unadjusted). Step-range means average each seed's checkpoints before pairing, as in Table 6. Reverse KL on the student's support was not run on the NVIDIA cluster. Per-benchmark columns are descriptive.

steps	AIME24	AIME25	AIME26	Avg.
Available teacher, NVIDIA cluster				
forward, teacher				
1–10	+3.55 [+2.69, +4.40]	+1.33 [−0.18, +2.83]	+2.91 [+1.74, +4.08]	+2.59 [+2.22, +2.97]
20–50	+1.81 [+0.96, +2.65]	+1.25 [−0.12, +2.62]	+2.16 [−0.71, +5.02]	+1.74 [+0.58, +2.89]
50	+2.50 [−2.68, +7.68]	+0.83 [−3.77, +5.43]	+3.02 [+0.24, +5.80]	+2.12 [+0.59, +3.64]
reverse, teacher				
1–10	+1.18 [+0.20, +2.16]	−0.38 [−1.34, +0.58]	+1.31 [+0.93, +1.70]	+0.70 [+0.21, +1.20]
20–50	+7.83 [+3.88, +11.78]	+3.28 [+1.53, +5.03]	+5.26 [+3.14, +7.39]	+5.46 [+3.83, +7.09]
50	+17.15 [+10.11, +24.20]	+5.76 [+1.58, +9.95]	+11.14 [+10.83, +11.44]	+11.35 [+7.99, +14.71]
forward, student				
1–10	+2.83 [+1.85, +3.80]	+1.47 [−1.31, +4.24]	+2.67 [+0.38, +4.96]	+2.32 [+0.33, +4.31]
20–50	+2.41 [+1.77, +3.05]	+1.79 [+1.18, +2.40]	+2.19 [+0.03, +4.35]	+2.13 [+1.41, +2.85]
50	+1.74 [−2.93, +6.40]	+1.60 [−2.55, +5.75]	−0.07 [−8.13, +7.98]	+1.09 [−1.91, +4.08]
Available teacher, AMD cluster				
forward, teacher				
1–10	+2.60 [−0.24, +5.45]	+1.88 [+0.61, +3.15]	+2.51 [+1.52, +3.50]	+2.33 [+0.79, +3.87]
20–50	+2.64 [+0.36, +4.92]	+1.60 [−0.28, +3.47]	+2.59 [+1.88, +3.29]	+2.27 [+1.51, +3.04]
50	+2.92 [−2.34, +8.17]	+1.11 [−7.27, +9.49]	+2.16 [−4.42, +8.73]	+2.06 [−2.68, +6.80]
reverse, teacher				
1–10	+1.57 [+1.02, +2.12]	+0.66 [+0.24, +1.08]	+0.73 [−0.10, +1.55]	+0.98 [+0.41, +1.56]
20–50	+7.48 [+2.99, +11.97]	+3.30 [+1.07, +5.53]	+4.97 [+3.05, +6.90]	+5.25 [+2.45, +8.05]
50	+16.74 [+13.79, +19.68]	+6.60 [+3.61, +9.59]	+8.76 [+5.82, +11.71]	+10.70 [+9.37, +12.03]
forward, student				
1–10	+2.50 [+1.32, +3.68]	+2.06 [+0.85, +3.26]	+2.93 [+1.22, +4.65]	+2.50 [+1.47, +3.52]
20–50	+2.67 [+0.36, +4.99]	+2.12 [−0.02, +4.26]	+1.62 [−0.33, +3.56]	+2.14 [+1.78, +2.49]
50	+1.18 [−5.55, +7.92]	+2.78 [−0.51, +6.06]	+1.08 [+0.15, +2.00]	+1.68 [−1.52, +4.87]
reverse, student				
1–10	+0.84 [+0.22, +1.46]	−0.12 [−0.55, +0.32]	+0.42 [−0.22, +1.07]	+0.38 [−0.08, +0.84]
20–50	+0.87 [−1.59, +3.33]	+0.50 [−0.50, +1.51]	+0.50 [−6.38, +7.39]	+0.62 [−2.42, +3.67]
50	+3.89 [−1.55, +9.33]	+2.01 [−3.17, +7.20]	+1.29 [−7.07, +9.66]	+2.40 [−3.47, +8.27]

Table 17: **Top-128 support overlap.** Fraction of the 128 scored token ids that are also in the student's own top-128, averaged over loss-carrying positions; three-seed means, AMD cluster; columns are training steps. Defined only for the teacher's support; in the two student-support objectives it is identically 1 and is not logged.

objective	arm	1	5	10	20	30	40	50
Matched teacher								
forward, teacher	OPD	0.606	0.655	0.557	0.810	0.904	0.895	0.889
reverse, teacher	OPD	0.607	0.676	0.751	0.882	0.889	0.869	0.835
Available teacher								
forward, teacher	PS-OPD	0.628	0.629	0.597	0.605	0.615	0.613	0.621
	OPD	0.651	0.713	0.694	0.698	0.702	0.700	0.702
reverse, teacher	PS-OPD	0.629	0.594	0.533	0.581	0.590	0.569	0.508
	OPD	0.651	0.668	0.668	0.675	0.650	0.448	0.286

training step 7, with 17.2, 20.4, 23.8 and 30.4% of its evaluation responses at the cap over steps 7–10. Its accuracy then falls to 8.9% at training step 20 and 8.1% at step 30, with 79.4% and 89.5% of its responses at the cap. PS-OPD's share at the cap never exceeds 12.8% at any scored checkpoint (the untrained student's is 15.1% at this decoding). Over training steps 1–6, before the explosion starts, PS-OPD already leads by +3.01 points ([+0.76, +5.26]), so its early lead is not only OPD coming

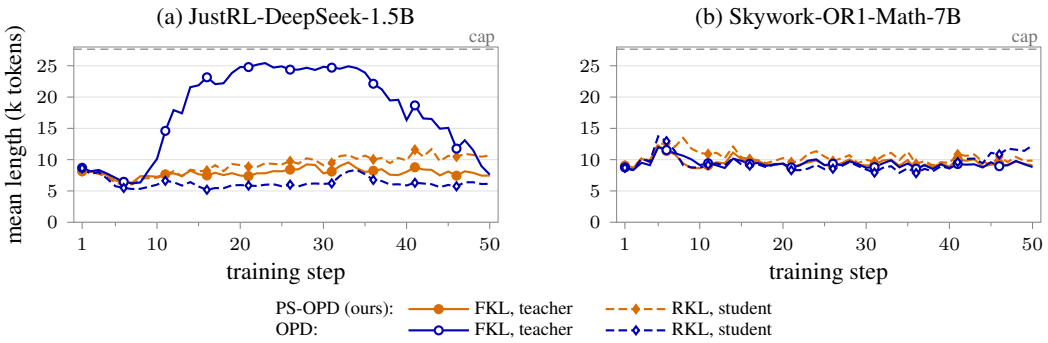


Figure 8: **Length of training rollouts.** Mean rollout length per training step, runs of Figure 4: (a) matched, (b) available teacher. PS-OPD's includes the teacher prefix. Dashed line: 27,648-token cap. AMD cluster, three-seed means.

Table 18: **Length and truncation by objective.** Available teacher unless marked; AMD cluster; three-seed means; columns are training steps. (a) Share of evaluation samples (temperature 1.0, top-p 1.0) at the 31,744-token response cap and (b) their mean length, on 90 AIME problems (accuracies use 89). (c) Logged share of a step's 256 training rollouts at the 27,648-token training cap (floor $1/256$, shown as 0.4; no rollout was aborted); its last row is the matched teacher's. Available-teacher forward-KL rows are omitted from (a) and (c) (at most 0.3 and 1.4 at every checkpoint shown).

objective	arm	1	5	10	20	30	40	50
(a) Evaluation responses at the cap (%)								
reverse, teacher	PS-OPD	0.0	9.5	9.5	1.0	0.5	0.6	7.4
	OPD	0.0	0.1	0.0	0.0	2.4	25.9	47.8
reverse, student	PS-OPD	0.0	10.3	10.0	2.5	1.6	2.8	1.3
	OPD	0.0	0.5	0.0	0.0	0.0	6.1	37.9
(b) Mean evaluation response length (tokens)								
forward, teacher	PS-OPD	11,654	13,851	13,329	13,627	13,659	13,805	13,376
	OPD	11,552	12,205	11,173	11,750	11,885	11,896	12,020
reverse, teacher	PS-OPD	11,858	18,327	16,987	16,204	15,712	15,607	15,985
	OPD	11,848	13,720	11,149	11,039	11,648	17,227	21,348
forward, student	PS-OPD	11,714	14,632	13,254	13,893	13,838	14,318	13,793
	OPD	11,613	12,526	11,376	12,015	11,988	12,085	12,170
reverse, student	PS-OPD	11,816	18,407	17,069	16,481	16,212	16,461	15,841
	OPD	11,794	14,284	11,503	12,281	12,043	13,667	19,075
(c) Training rollouts at the cap (%)								
reverse, teacher	PS-OPD	0.4	6.0	7.7	0.8	0.8	0.5	4.7
	OPD	0.4	3.0	0.4	0.4	0.8	12.2	31.1
reverse, student	PS-OPD	0.4	8.1	6.9	1.4	2.5	0.7	1.6
	OPD	0.4	3.3	0.4	0.4	0.4	1.3	21.5
matched teacher								
forward, student	OPD	0.4	0.4	10.9	83.7	81.4	27.1	12.4

apart. It leads by $+3.40$ ($[+2.99, +3.81]$) over steps 1–10 and by $+18.79$ ($[+15.94, +21.63]$) over steps 20–50, and in every seed first reaches OPD's step-50 accuracy (34.3%) at training step 20. At training step 50, 98.5% of PS-OPD's responses finish, against 82.5% of OPD's. With forward KL over the student's top-128, OPD falls to 11.3% at training step 20 with 76.2% of its responses at the cap, and its training rollouts reach the training cap on up to 88.4% of a batch (training step 25, between the columns of Table 18).

Figure 8 gives the mean training-rollout length for the runs of Figure 4, including PS-OPD's teacher prefix. With the matched teacher under the main objective, OPD's mean rollout rises from 8.7k tokens at training step 1 to 25.4k at step 23 and returns to 7.6k at step 50, while PS-OPD's stays between 6.3k and 9.6k. In the other three teacher-objective settings, both arms stay between 5.2k and 13.7k.

Table 19: **Reaching the cap and finishing, available-teacher arms of Table 1.** Three-seed means (%) over training steps 1–10, 20–50 and at step 50, on the 90 AIME problems of Table 18; AMD cluster, temperature 1.0, top-p 1.0. *At cap*: samples at the 31,744-token response cap. *Finished*: samples giving a final answer, possibly before the cut. *On finished*: accuracy among those.

		at cap			finished			on finished		
arm	**objective**	1–10	20–50	50	1–10	20–50	50	1–10	20–50	50
Teacher	(step 0)			0.9			99.1			61.1
OPD	forward, teacher	0.0	0.0	0.0	99.9	99.6	99.6	21.5	26.6	27.4
	reverse, student	0.1	11.0	37.9	99.9	80.4	50.3	26.4	42.5	60.6
Off-policy distillation		5.6	0.4	0.6	95.0	99.8	99.7	25.0	26.2	26.8
FH	forward, teacher	0.1	0.0	0.0	99.8	99.7	99.7	22.1	27.5	27.6
TrOPD		0.0	0.0	0.0	99.9	99.8	99.7	23.7	27.0	26.8
PS-OPD	forward, teacher	0.0	0.1	0.0	100.0	100.0	99.9	23.8	28.8	29.3
	reverse, student	8.8	2.1	1.3	92.3	98.4	99.0	29.2	32.9	32.2

In every reverse-KL setting with the available teacher, truncation in training rises with truncation at evaluation (Table 18), so the length explosion is a training phenomenon, not an evaluation artefact. Evaluation lags when truncation falls again. With the matched teacher under the main objective, 10.0% of OPD's training rollouts but 20.5% of its evaluation responses are at the cap at training step 50. The teachers themselves reach the evaluation cap on 2.7% (matched, temperature 0.6) and 0.9% (available, temperature 1.0) of their AIME samples on the AMD cluster. With the available teacher, early truncation comes from PS-OPD under reverse KL, peaking at 15.6% (teacher's support) and 19.2% (student's) at training step 7, a checkpoint between the columns of Table 18. Off-policy distillation truncates 8.1–10.3% over steps 4–6. From training step 20 on, these are at most 2.8% and 0.6%, except PS-OPD under reverse KL on the teacher's support, which returns to 7.4% at step 50. Under forward KL with the available teacher, neither arm exceeds 0.3% at any checkpoint and PS-OPD finishes at least 99.8% of its responses, so its leads there are accuracy on comparable outputs.

Table 19 gives, for the available-teacher arms of Table 1, the shares of responses at the cap and with a final answer, and accuracy on finished answers. The last drops exactly the responses that ran out of room or stopped without an answer, so it flatters the arm that finishes less. OPD under reverse KL on the available teacher's support scores 78.2% on its finished answers at training step 50, above that teacher's own 61.1%. We therefore never rank arms by it.

What truncated responses contain. We check the last 3,000 characters of every truncated response in seed 1. At OPD's peak checkpoint in its unstable setting with the available teacher (reverse KL on the teacher's support, step 50), every one ends in non-linguistic token output (bare punctuation and digits, or mixed-script token sequences) rather than in reasoning cut short. The other truncations with this teacher are ordinary reasoning that ran out of room: all 13 of the teacher's own, and most of PS-OPD's at step 7 under reverse KL on the student's support and of off-policy distillation's at step 6. The check is heuristic, covers one seed, and we did not evaluate at a larger budget.

D Controls and Ablations

The controls and ablations of Section 4.3, all exploratory, use the available teacher, NVIDIA cluster, forward KL over the teacher's top-128 and seeds 1–3 unless stated otherwise. Appendix K.1 describes the decoupled and sequential controls.

Neither of PS-OPD's ingredients, the prefix as rollout context or as cross-entropy target, alone matches it in Avg. over both training steps 1–10 and 20–50 (Table 20, Figure 9). *Context alone* (RH) reaches OPD's step-50 accuracy almost as early as PS-OPD (training step 9 against 7), but PS-OPD is higher over both periods and at step 50. *Loss alone* (Decoupled: unprefixed OPD plus a separate cross-entropy on the same prefixes) recovers most of the early gain (+2.18 points against OPD over steps 1–10, Table 25). It loses the later lead, first reaching 26.8% at training step 40, and differs from PS-OPD in both rollout context and continuation budget. *Loss first* (SFT→OPD: ten SFT steps, then 40 OPD steps) ends below PS-OPD at step 50 and below OPD over steps 20–50 (24.1% against 26.4%). PS-OPD's rise follows a transient step-4 dip, and RH dips deeper (Ap-

Table 20: **Ingredients and weight.** avg@16 accuracy in %, mean of seeds 1–3 (available teacher, NVIDIA cluster, temperature 1.0, top-p 1.0, learning rate 10^{-5}). Columns: training step or range. RH (λ=0) uses the prefix only as context, Decoupled only as cross-entropy target, $\lambda > 0$ as both, OPD as neither. SFT→OPD: ten SFT, then 40 OPD steps. *R*: first checkpoint whose Avg. mean reaches OPD's step-50 mean rounded down (26.8%), none later over 1 point below (–: never; per benchmark: Table 9). $^{\downarrow}$: below PS-OPD in Avg., seed-paired 95% interval excluding zero (Tables 6, 25).

	AIME24			AIME25			AIME26			Avg.				
arm	1–10	20–50	50	1–10	20–50	50	1–10	20–50	50	4	1–10	20–50	50	*R*
OPD	25.3	33.3	33.2	22.1	25.7	26.5	14.6	20.1	20.8	17.3	20.7$^{\downarrow}$	26.4$^{\downarrow}$	26.8$^{\downarrow}$	40
Decoupled	28.2	32.4	33.0	23.6	26.8	26.9	16.7	21.1	22.1	19.1	22.8	26.8$^{\downarrow}$	27.4	40
SFT→OPD	–	30.6	30.8	–	23.4	23.8	–	18.2	18.5	–	–	24.1$^{\downarrow}$	24.4$^{\downarrow}$	–
RH (λ=0)	26.7	33.5	33.8	21.8	26.9	26.0	15.7	21.7	21.7	12.7	21.4$^{\downarrow}$	27.4$^{\downarrow}$	27.1$^{\downarrow}$	9
λ=1	29.6	34.0	34.2	24.1	26.6	26.9	17.3	21.1	21.8	21.3	23.7	27.3	27.7$^{\downarrow}$	8
λ=3	28.8	33.7	32.9	23.6	26.7	26.9	16.1	19.9	19.6	23.0	22.8	26.8	26.5$^{\downarrow}$	30
PS-OPD (λ=0.3)	28.9	35.1	35.7	23.4	27.0	27.3	17.5	22.3	23.9	16.7	23.3	28.1	28.9	7

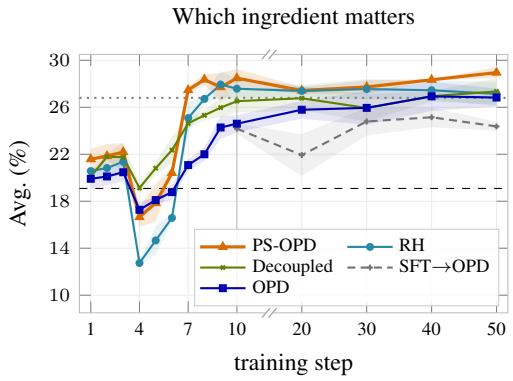


Figure 9: **Which ingredient matters.** Avg. (available teacher, NVIDIA cluster, temperature 1.0, top-p 1.0). Lines: mean of seeds 1–3; shaded: ±1 standard deviation across these seeds. Dashed: untrained student (19.1%); dotted: OPD's step-50 accuracy (26.8%). After training step 10 (break marks), the step axis is compressed. SFT→OPD is scored from step 10.

pendix F). In a single-seed sweep of the cut, its depth tracks the weight rather than the prefix length (Appendix G).

E Weight Ablation: Paired Statistics

Every positive prefix-loss weight leads RH and OPD early in training (Section 4.3), but later the ordering changes and several differences are unresolved. Table 20 gives per-benchmark means. The four arms (trained configurations) share the $r = 0.5$ prefix-conditioned rollout and differ only in the cross-entropy weight. Every statistic below is Avg., paired within seed (95% t intervals in brackets).

Over steps 1–10, every $\lambda > 0$ averages above RH, with 23.69% ($\lambda = 1$), 23.25% ($\lambda = 0.3$) and 22.84% ($\lambda = 3$) against 21.40%. The weight $\lambda = 1$ leads RH by 2.28 points ($t = 8.3$, $p \approx 0.014$ unadjusted), and $\lambda = 0.3$ by 1.85 ($t = 22.1$). Every $\lambda > 0$ also leads OPD over these steps ($t \geq 11.6$).

Over steps 20–50, $\lambda = 1$ leads OPD (0.88 points, $t = 7.6$), while the leads of RH (1.01, $t = 2.5$) and $\lambda = 3$ (0.40, $t = 0.9$) are unresolved. The difference $\lambda = 1$ minus RH is -0.1 $[-2.3, +2.1]$ points. Of the four checkpoint differences $\lambda = 1$ minus RH at steps 20, 30, 40 and 50, only the step-20 one, -0.9 $[-1.2, -0.6]$ points (negative in all three seeds, $t = -13.1$), is resolved. It alone remains significant after Holm correction (Appendix J). These exploratory tests establish neither an asymptotic ceiling nor a stable ordering of the weights.

At step 50, $\lambda = 0.3$ leads RH by 1.8 points ($t = 4.6$) and leads $\lambda = 1$ by 1.27 ($t = 9.0$). Because we chose $\lambda = 0.3$ by its step-50 mean, these tests restate the selection rule and are not independent evidence that 0.3 is the best weight.

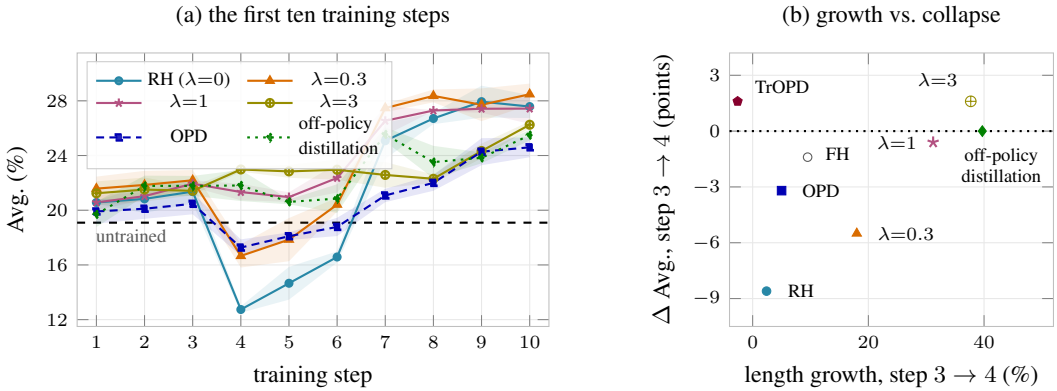


Figure 10: **The step-4 dip** (available teacher, NVIDIA cluster, temperature 1.0, top-p 1.0). **(a)** Avg. over the first ten training steps at 10^{-5}. Lines: mean of three seeds; shaded: ± 1 standard deviation across seeds; dashed: untrained student (19.1%). Off-policy distillation is the 256-sequence-batch run. **(b)** Change in Avg. from step 3 to step 4 against the growth of each arm's mean response length over the same training step (three seeds pooled; values in Table 21).

Table 21: Step 3 $\rightarrow$ 4 by arm (available teacher, NVIDIA cluster). Δ: change in accuracy (points, three-seed mean; for RH and OPD, bootstrap 95% CI of the Avg. change over 5,000 problem resamples) and in mean response length (three seeds pooled). Wait density (whole-word "Wait" per 1,000 generated tokens at step 4; untrained: 8.2) and Δ p_{90} length: seed 1. Calibration: Spearman correlation at step 4 between a problem's mean response length and the untrained model's accuracy on it (untrained: -0.55), three seeds pooled.

	RH (λ=0)	λ=0.3	λ=1	λ=3	OPD	Off-policy distillation	FH	TrOPD
Δ AIME24	-11.1	-8.4	$+1.3$	$+1.2$	-2.4	$+1.1$	-2.7	$+1.8$
Δ AIME25	-9.7	-3.7	$+1.3$	$+2.5$	-2.4	$+0.1$	-0.1	$+1.3$
Δ AIME26	-5.1	-4.5	-4.4	$+1.0$	-4.7	-1.1	-1.4	$+1.7$
Δ Avg.	-8.6	-5.5	-0.6	$+1.6$	-3.2	$+0.0$	-1.4	$+1.6$
95% CI	$[-12.1, -5.4]$				$[-4.8, -1.7]$			
Δ mean length	$+2.4\%$	$+18.0\%$	$+31.2\%$	$+37.7\%$	$+5.0\%$	$+39.7\%$	$+9.5\%$	-2.6%
Wait density	27.9	26.0	21.3	16.4	18.7	20.7	18.9	16.0
Δ p_{90} length	-3.4%	$+16\%$	$+29\%$	$+40\%$	$+1.7\%$	$+51\%$	$+6.3\%$	-4.3%
Calibration	$+0.14$	-0.01	-0.24	-0.44	-0.12	-0.35	-0.21	-0.35

F THE STEP-4 DIP: SUPPORTING ANALYSIS

The step-4 dip (Section 4.3) is transient, and among the weights only the two whose responses lengthen least fall below the untrained student (observational, partly single-seed measurements). In Avg., RH's step-4 score, 12.74%, is not merely a local minimum but *below* the untrained student (19.09%), by 6.35 points (Figure 10(a)). The bootstrap 95% CI of this gap is $[-9.8, -3.2]$ (5,000 resamples paired over problems). OPD dips more mildly, to an unresolved 1.82 points below the student. The means of RH and OPD stay below the untrained student for three consecutive steps and that of $\lambda = 0.3$ for two, while those of $\lambda = 1$, $\lambda = 3$ and off-policy distillation never do. From step 7 on, every seed of every three-seed arm scores above the untrained student in Avg. (Tables 5 and 24), and RH's mean is back above its step-3 score (25.1% against 21.4%). Table 21 and Figure 10(b) relate the dip to each arm's length growth at that step.

We do not rank the four weights' dips by depth, because not every gap between adjacent weights exceeds seed variation, but they split into two groups. The two whose mean response length grows least, RH (+2.4%) and $\lambda = 0.3$ (+18.0%), fall below the untrained student at step 4 in every seed in Avg. Both $\lambda = 1$ (+31.2%) and $\lambda = 3$ (+37.7%) stay above it in every seed (Table 5). Off-policy distillation lengthens most (a lower bound, since it truncates 8.7% of its step-4 responses) and does not dip. Arms whose rollouts carry no prefix do not follow this pattern. OPD barely lengthens

yet loses 3.2 points (CI $[-4.8, -1.7]$), so a prefix is not necessary for the step-4 drop. Our design does not separate two further differences between OPD and RH. In RH, the prefix places every graded token in the teacher's late, backtrack-dense reasoning, and prefixed problems train under a generation budget reduced by their prefix length (Appendix A). FH loses 1.4 points against OPD's 3.2, a difference three seeds do not resolve ($t = 1.3$). Yet its mean length grows more ($+9.5\%$ against $+5.0\%$; paired over seeds, $+4.5$ percentage points of growth, $t = 6.8$). TrOPD gains an unresolved 1.6 points while its mean length falls 2.6% (in every seed). Thus, across arms, smaller length growth does not imply a deeper dip.

What the dip is not. Six simple explanations do not fit.

- *Formatting:* boxed answers and `</think>` closure are at $\geq 99.9\%$ at RH's steps 3 and 4, against 99.7–99.8% in the untrained model.
- *Evaluation cap:* no OPD response reaches the 31,744-token cap at steps 1–10 (other objectives: Appendix C), and RH reaches it on 1 of 21,360 responses over steps 3–7.
- *Truncation* runs against the dip. At steps 4–6 it is 2.8–2.9% for λ=3 and 8.7–10.6% for off-policy distillation (at most 0.2% at step 3), yet neither arm dips in Avg.
- *Redundancy:* the compression ratio of the trailing 20k characters (lower means more redundant) falls monotonically *through* the recovery, from 0.342 at step 3 to 0.295 at step 7 (no response below 0.12), though this does not exclude every form of reasoning degeneration.
- *Problem order:* the three seeds train on different problem orders but bottom out at the same step.
- *Continuation:* openings identifiable as mid-stream continuations are 0.1% of responses at RH's step 4, as at step 3, step 7 and in the untrained model.

What is lost. Between steps 3 and 4, RH's pass@16 on the 89 Avg. problems falls from 48.3% to 28.0% (maj@16: $32.2\% \rightarrow 19.4\%$; per seed). This is a loss of sampled solving success, not proof of erased knowledge. Problems the step-3 checkpoint could already solve account for 8.7 points of the 8.6-point net loss, offset by a 0.1-point gain on the others. Pooling the three seeds (48 samples per problem), 25 of the 53 problems with a non-zero step-3 score get zero on all 48 samples at step 4, and 36 are back at or above their step-3 level by step 7. By tercile of the untrained model's accuracy (three seeds pooled), RH moves (step $3 \rightarrow 4 \rightarrow 7$) $58.3\% \rightarrow 37.5\% \rightarrow 65.4\%$ on the easiest, $4.6\% \rightarrow 0.4\% \rightarrow 7.7\%$ on the middle and $0.8\% \rightarrow 0.1\% \rightarrow 1.5\%$ on the hardest, overshooting step 3 on every tercile by step 7.

Observations motivating a mechanism. Three observations suggest a mismatch between discourse changes and response-length allocation. First, whole-word "Wait" markers roughly double at this step in every arm we counted except TrOPD, including arms that never dip (seed 1; RH $149 \rightarrow 323$ per response). Within the λ arms what differs is their *density* (Table 21), as arms that lengthen absorb the same increase over more tokens. Second, RH's 90th-percentile length falls while its 10th percentile rises, compressing their ratio from 2.74 to 2.03 (back to 3.57, past the base model's 3.08, by step 7, when accuracy recovers). Third, calibration, negative when harder problems get longer responses, recovers for RH to -0.51 by step 7 and -0.62 by step 10. It orders all eight arms as their mean step-3 $\rightarrow$ 4 changes in Avg. are ordered, up to the tie between TrOPD and λ=3. This rank agreement between means is descriptive, not a test, and holds on no single benchmark. These measurements do not establish that a longer response would make the new style productive, or that missing prefix imitation is the cause.

What this does not establish. The length–accuracy relationship reverses with the level of analysis. Within a checkpoint, longer responses to the same problem score worse. Across the λ arms and off-policy distillation, the arms that lengthen most stay above the untrained student at step 4, although TrOPD shortens without dipping. No observational design here settles this reversal, nor do we claim that the compression is the sole cause of the accuracy loss rather than a correlate. Training step 4 is the least remarkable of steps 1–7 by every training-log statistic we checked (smallest pre-clip gradient norm, smallest loss). These aggregate logs do not rule out an optimization effect or interactions with AdamW's moment history (Appendix I).

G Prefix-Length Ablation

In one seed, the step-4 dip tracks the cross-entropy weight rather than the prefix length (Table 22). Every prefixed arm above cuts the teacher's reasoning at $r = 0.5$, i.e., $K_i = \lfloor 0.5\, M_{\text{think}}(i) \rfloor$ capped

at 8,192 tokens. We also train $r \in \{0.125, 0.75\}$ at $\lambda = 0$ (context only) and $\lambda = 1$, one seed each on the dense early checkpoint schedule. These are single-seed observations, against a typical across-seed standard deviation of 0.7 points per checkpoint (Section 4.1). The values $r \in \{0.125, 0.5, 0.75\}$ give mean prefixes of 988, 3,751 and 5,013 tokens, i.e., 12.5%, 49.1% and 70.1% of the reasoning span. The cap binds on the longest traces, so we report the realized mean $\bar{K}$, not r. All prefixes are cut from the *same* teacher trace per problem, on the same 15,492 of 17,917 training problems, and the $r = 0.5$ prefixes are byte-identical to those of the main results. The prefix occupies part of the 27,648-token response, and the remaining budget is not equalized (Appendix A). Prefixed problems keep a mean of 26,660, 23,897 and 22,635 tokens at $r = 0.125$, 0.5 and 0.75, so the ablation varies demonstration length and remaining budget together. Numbers below are Avg. unless a benchmark is named.

Table 22: Prefix-length ablation (available teacher, NVIDIA cluster, seed 1): avg@16 accuracy in %. $\bar{K}$: realized mean prefix length in tokens. Avg. at single checkpoints and averaged over training steps 1–10; $\Delta_{3\to4}$: change from step 3 to step 4 in points. Untrained student: 19.1 (Avg.); 26.5, 19.4 and 11.4 per benchmark (Table 4). The $r = 0.5$ rows are seed 1 of RH and $\lambda = 1$.

			Avg.							**step 4** ($\Delta_{3\to4}$)		
λ	r	$\bar{K}$	4	$\Delta_{3\to4}$	1–10	20	30	40	50	AIME24	AIME25	AIME26
0	0.125	988	14.5	−6.4	21.6	27.1	26.7	28.0	25.5	17.9 (−8.7)	15.8 (−6.2)	9.7 (−4.1)
0	0.5	3,751	13.0	−7.6	21.6	27.8	27.0	27.7	26.6	16.9 (−8.5)	13.5 (−9.4)	8.6 (−5.0)
0	0.75	5,013	13.8	−7.3	21.2	25.4	26.2	26.1	26.9	16.5 (−9.2)	15.4 (−6.9)	9.5 (−5.8)
1	0.125	988	23.3	+1.2	22.6	26.8	27.5	26.9	25.8	30.0 (+1.7)	23.1 (+0.4)	16.8 (+1.5)
1	0.5	3,751	22.0	+0.4	24.2	27.0	27.3	28.5	27.5	28.3 (+3.5)	24.2 (+1.9)	13.6 (−4.3)
1	0.75	5,013	22.0	+0.7	24.1	26.7	26.7	27.3	27.7	28.5 (+2.7)	23.5 (+1.0)	13.8 (−1.5)

Step-4 dip: weight, not length (one seed). At $\lambda = 0$, every prefix length dips at step 4 to between 13.0% and 14.5%, below the untrained student in Avg. and on each benchmark, with first-ten-step means within 0.4 points of one another. At $\lambda = 1$, every prefix length stays above the untrained student at step 4, in Avg. and on each benchmark. On AIME26, however, the step-3 $\to$ 4 change is −4.3 points at $r = 0.5$ and −1.5 at $r = 0.75$. Across a five-fold range of realized prefix length, λ separates the arms that fall below the untrained student from those that do not, while the amount of reasoning handed over does not. This contrasts $\lambda = 1$ with $\lambda = 0$ only. At $r = 0.5$, PS-OPD ($\lambda = 0.3$) still dips below the untrained student in all three seeds (Avg., Table 5).

The shortest supervised prefix falls later (one seed). At $\lambda = 1$, $r = 0.125$ does not dip at step 4. After a high of 25.6% at step 5, it falls to 22.7%, 21.6% and 20.8% at steps 6–8 (still above the untrained student's 19.1%), before reaching 25.7% at step 10. The two longer prefixes score between 27.2% and 28.3% at steps 7–9. From step 20 on, every prefix length at either λ lies between 25.4% and 28.5%. These readings rest on one seed, with a generation budget that falls as r rises.

H Robustness at a Ten Times Smaller Learning Rate

At a ten times smaller learning rate, PS-OPD still reaches OPD's step-50 accuracy before OPD, in Avg. and on each benchmark, after about ten times as many training steps (Table 23). Every arm in the main study trains at a constant 10^{-5}, ten times the 10^{-6} of several recent OPD recipes for math-reasoning students of 0.6B–8B parameters (Li et al., 2026; Zhao et al., 2026; Liu et al., 2026a; Xu et al., 2026), and twice TrOPD's 5×10^{-6} on this teacher–student pair (Xing et al., 2026). Ko et al. (2026) run OPD at 10^{-5} on this pair, and Zhang et al. (2026) use 5×10^{-5} with base students (rates only loosely comparable across optimizer steps per rollout batch, batch size, loss aggregation and response cap). We repeat the six-arm comparison at 10^{-6} for 150 training steps with one seed, checkpointed every 10 steps (Table 23). Off-policy distillation at this rate is the earlier 64-sequence-batch run. With one seed and two rates, this describes how the curves move with the rate, without a variance estimate, and does not show that the rate alone produces PS-OPD's early crossing.

Cost to reach OPD's step-50 accuracy. With the Avg. threshold $T_a = 26.8\%$ of Appendix B.4, at 10^{-6} PS-OPD first reaches T_a at training step 70 (26.9%, after 26.7% at 60) and never falls back below T_a minus 1 point. OPD never reaches it, peaking at 25.7% at training step 110. At 10^{-5}, the same seed reaches T_a at training step 7 for PS-OPD and 40 for OPD. Measured as in Table 10, the

Table 23: Learning rate 10^{-6} (available teacher, NVIDIA cluster, temperature 1.0, top-p 1.0, seed 1): Avg. accuracy in % at every checkpoint, and the mean over the six checkpoints from 100 to 150 in Avg. and on AIME24, AIME25 and AIME26 (A24–A26; bold: best arm). RH is $\lambda = 0$, PS-OPD $\lambda = 0.3$. Off-policy distillation (Off-pol.) is the earlier 64-sequence-batch run; FH and TrOPD were not run at this rate. Steps 20, 40, 60 and 80 are not in the provenance manifest (Appendix K.2).

	Avg. at training step															steps 100–150			
arm	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	Avg.	A24	A25	A26
OPD	20.7	20.0	22.1	20.4	21.9	22.8	22.5	24.2	24.4	24.1	25.7	23.9	25.1	24.7	24.4	24.6	31.3	25.3	17.3
RH	22.2	21.6	20.6	21.6	24.5	25.9	25.7	25.8	26.4	26.0	26.3	27.1	27.5	26.1	27.1	26.7	33.3	25.8	**21.0**
PS-OPD	20.4	21.9	21.9	21.7	25.9	26.7	26.9	26.8	26.2	26.6	26.2	27.5	26.3	28.0	27.7	27.0	33.2	**27.1**	20.8
λ=1	19.8	21.6	23.5	23.7	25.3	26.6	25.7	26.4	27.0	26.4	25.9	25.9	27.0	25.0	28.2	26.4	32.9	26.6	19.7
λ=3	21.2	19.5	22.4	25.1	24.2	25.5	25.6	26.8	27.0	26.8	27.7	26.4	26.3	28.1	27.6	**27.2**	34.5	26.2	20.8
Off-pol.	20.1	20.6	20.8	24.3	25.4	25.0	27.1	26.9	26.2	27.3	26.6	26.4	27.3	28.3	26.0	27.0	**34.7**	25.8	20.5

150-step PS-OPD run took 133.9 GPU-hours, 0.893 per step. Its crossing takes about 62 GPU-hours of training (75 with the 12.74 GPU-hours of teacher traces), against 6.2 (18.9) at 10^{-5}, i.e., about ten times the training steps and training GPU-hours. With per-benchmark thresholds set the same way (33.1%, 26.4% and 20.8% on AIME24, AIME25 and AIME26), PS-OPD first crosses at steps 60, 70 and 90 at 10^{-6}, against 7, 8 and 7 at 10^{-5}. OPD crosses only on AIME25, at step 130, and falls back, whereas at 10^{-5} it crosses at steps 40, 50 and 40. PS-OPD thus crosses before OPD at both rates, in Avg. and on each benchmark.

Level within 150 training steps. In Avg., the steps-100–150 means place RH, the three weights and off-policy distillation between 26.4% and 27.2%, with OPD lowest at 24.6%, also on each benchmark. PS-OPD's 27.0% is not the highest, and $\lambda = 0.3$ was selected at 10^{-5}. PS-OPD peaks at 27.95% (step 140) after 133.9 GPU-hours, against 29.0% at step 50 at 10^{-5} after 42.5 GPU-hours. It averages 27.0% over steps 100–150, against 28.0% over steps 20–50 at 10^{-5}. At 10^{-5}, OPD is within 0.3 points of RH by step 50 in Avg. (three-seed means). At 10^{-6}, it remains lower at step 150, in Avg. and on each benchmark. Whether it would catch up is unresolved, so the experiment does not establish a shared final ceiling. Off-policy distillation is the only arm with sustained truncation at this rate, peaking at 5.4% of responses at step 40 and falling below 1% by step 120. The λ=3 arm reaches 1.7% at step 40, and every other arm stays under 0.3%.

The baseline's rate. On seed 1, the shared rate does not handicap OPD relative to the two other rates we ran. OPD's step-50 score is highest at the shared rate: 27.6% in Avg. at 10^{-5}, against 24.7% at 2×10^{-5} (Appendix B.3) and at most 25.7% within 150 training steps at 10^{-6}. The same holds on each benchmark.

I Discussion and Limitations

This appendix bounds our claims about distilling DeepSeek-R1-Distill-Qwen-1.5B from the available teacher Skywork-OR1-Math-7B and the matched teacher JustRL-DeepSeek-1.5B.

Early progress, late accuracy, and compute are separate. The early advantage is an average over training steps, not a wall-clock comparison. Off-policy distillation continued to 200 training steps comes within an unresolved 0.8 points of PS-OPD's step-50 accuracy, without rollouts or teacher scoring (Appendix B.3). GPU-hour comparisons include prefix construction but are quantised by the checkpoint grid, cover one pair and are seed-sensitive (Appendix B.4). Under seed resampling, PS-OPD saves a median of 16 GPU-hours to reach OPD's step-50 accuracy, but the 2.5% quantile is 7. We chose the prefix-loss weight $\lambda = 0.3$ on the reported seeds and problems. A preregistered test on three fresh seeds confirms its later lead over OPD, with uncertain effect size (interval lower bound +0.3 points).

Scope of the available-teacher setting. This setting reuses a released teacher, here RL-finetuned from the larger DeepSeek-R1-Distill-Qwen-7B, not our student. It aims to avoid RL-finetuning the student checkpoint as a *prerequisite*, not to remove the teacher's own training cost. We claim neither efficiency nor superiority over RL finetuning of the student itself.

Two teacher–student pairs. Each teacher runs at three seeds under four objectives (Appendix C). With both, PS-OPD leads OPD over training steps 1–10 under forward KL over the teacher's top-128. Two pairs neither establish a claim about teacher regimes nor form a controlled contrast of teacher identity. Each teacher has its own prefix pool, coverage and mean prefix length (Table 12). The available teacher's headroom over the student is nearly twice the matched one's. We ran no context-only, decoupled or sequential control and no λ sweep on the matched pair. Thus, Section 4.3, Appendix D and the two-way test of Section 3 remain available-teacher results. We do not claim that the prefix's context or target role carries the matched pair's result. Nor does the prefix's help to the frozen student establish a learning mechanism. The decoupled control reproduces most of PS-OPD's average gain over training steps 1–10, mainly by dipping less, but not the later lead in an exploratory contrast (Appendices K.1 and B.4).

Evaluation and domain scope. Our contamination audit found no training matches for AIME24 or AIME25, which does not prove them uncontaminated. The Avg. suite (89 problems) is small. Every reported accuracy, the fresh-seed confirmation included, comes from the three benchmarks used to select λ. Our intervals cover training seeds only. In the main study, we report problem-level uncertainty only for the step-4 dip. We did not compare short and long training allowances. Thus, long-horizon reasoning describes our single-turn mathematics setting, not a demonstrated source of PS-OPD's advantage. Nor do we claim that the construction generalizes to chat, long-input retrieval, multi-turn agents or other domains merely because it needs no verifier.

Length, style, and the dip. A teacher prefix shortens the student's token allowance by its length, and we ran no horizon-matched control. Over training steps 1–10 of the available teacher's forward-KL variants, only 11 of 7,680 context-only rollouts (0.14%) reach the 27,648-token cap, prefix included, and 1 of OPD's. With $\lambda = 0.3$, 1 and 3, which dip less or not at all, the shares are 0.22%, 0.47% and 1.24%. Thus, the variant that dips most truncates least, arguing against truncation as the direct cause. Still, a smaller allowance could shape generation before the cap binds. OPD's destabilisation with the matched teacher, where up to 79.3% of its rollouts in a training step reach the cap under the main objective, is a different event (Figure 4). Ordinary gradient norms and losses at step 4 do not rule out an optimizer effect through AdamW's history-dependent moments.

Learning-rate scope. All main-study methods share the learning rate 10^{-5}, so PS-OPD's 7-step crossing of OPD's final accuracy compares methods at this rate. Evidence at 10^{-6} and 2×10^{-5} rests on one seed each (Appendices H and B.3). The coarsely checkpointed 10^{-6} study neither replicates nor rejects the three-step dip. Nor is 10^{-5} safe for every OPD configuration, as OPD destabilises at this rate with both teachers. In both forward-KL settings of the matched pair, it falls to 8.1% over the teacher's top-128 (step 30) and to 11.3% over the student's (step 20). Conversely, PS-OPD under the main objective never has over 12.8% of its responses at the cap (Appendix C). The finding applies to the measured recipe, not every OPD loss or learning rate.

Demonstration selection and prefix length. Off-policy distillation uses complete teacher traces without correctness filtering (Table 12). Filtering would change quality, coverage and difficulty, with unknown net effect. We do not claim that the cut fraction $r = 0.5$, giving a mean prefix of 3,751 tokens with the available teacher, is optimal. The one-seed length ablation varies prefix length and remaining suffix allowance together (Appendix G).

J Statistics and Decontamination

Seed variability. Over the 84 checkpoints of the six main methods, the across-seed standard deviation of Avg. has median 0.73 and mean 0.70 points. This gives a descriptive scale of 0.40 points for a three-seed mean (mean standard deviation/$\sqrt{3}$), not a significance cutoff. Comparisons over a window of steps first average checkpoints within each seed, since checkpoints from one run are not independent replicates. A non-significant difference does not establish equal accuracy ceilings. The step-4 dip intervals use a bootstrap over evaluation problems, preserving the Avg. weighting and conditioning on the trained seeds.

Confirmatory and exploratory tests. The only confirmatory comparison is the fresh-seed test of the available teacher's later lead under forward KL over the teacher's top-128, fixed before submitting its runs (Appendix K.2). Tables 6 and 25 give seed-paired differences on seeds 1–3 with 95% t intervals, reflecting training-seed variation only. They have two degrees of freedom ($t_{0.975} = 4.303$),

or five for pooled seeds 1–6 ($t_{0.975} = 2.571$). These contrasts are exploratory, since we examined the windows during analysis and selected $\lambda = 0.3$ on these seeds. Holm's step-down correction at $\alpha = 0.05$ on the unadjusted two-sided p-values leaves significant:

- Table 6: 2 of 28, PS-OPD minus OPD over steps 1–10 ($t = 29.8$) and minus off-policy distillation over steps 20–50 ($t = 29.3$). Restricting the family to the 12 contrasts against the four baselines (steps 1–10, 20–50 and step 50) leaves the same 2.
- Table 25: 0 of 18, and 0 of 46 for both tables together.
- Checkpoint tests of $\lambda = 1$ minus RH (prefix only as context): 1 of 4 (step 20).
- All 79 distinct seed-paired tests as one family: 1, off-policy distillation at step 200 over OPD at step 50 ($t = 45.3$). Three seeds cannot reliably estimate its small standard deviation (Appendix B.3).

The objective grid and matched-teacher pair (Appendix C) form a separate exploratory family outside every Holm family, without multiplicity adjustment or confirmatory conclusion.

Contamination audit. We scanned every AIME24, AIME25 and AIME26 problem against all 17,917 DAPO-Math-17k training problems in two passes, and all reported numbers exclude the single match. The first pass checks word 8-gram containment after stripping instruction boilerplate and reducing LaTeX markup to plain tokens. The second checks bag-of-words containment (content words and numbers) to catch paraphrases. Two independent LLM readers judged every candidate pair, each seeing both problems but not the other's verdict. Each labeled it SAME (allowing rewording, formatting, or one of DAPO's answer rewrites), VARIANT (same template, different numbers or asked quantity), or DIFFERENT. A majority-of-three tiebreak was never needed. AIME24 and AIME25 received no SAME verdict in either pass. AIME26 had exactly one match, problem 0 (present verbatim, ground truth 277), leaving 89 Avg. problems.

K Controls in Detail, Confirmatory Test and Baselines

This appendix details the decoupled and sequential controls (Appendix K.1), the confirmatory fresh-seed test (Appendix K.2), and the FH (Appendix K.3) and TrOPD (Appendix K.4) baselines. All use the available teacher on the NVIDIA cluster with three seeds, decoded at temperature 1.0 and top-p 1.0 (Appendix A). In the text, accuracies and tests use Avg. Each trained method is an *arm*, and a seed-paired difference is *unresolved* when its confidence interval includes zero.

K.1 Decoupled and sequential controls

Two controls, not new method variants, test whether simpler combinations of imitation and OPD match PS-OPD. The prefix loss beside unprefixed OPD recovers most of the early lead of PS-OPD, but not the later lead. SFT followed by OPD, with the split of training steps we ran, reproduces the accuracy of neither PS-OPD nor OPD at equal total training steps.

Design. Comparing $\lambda > 0$ with RH ($\lambda = 0$, prefix only as rollout context) isolates the effect of supervising the prefix, not whether it is needed as context, which the *decoupled* control tests. It trains unprefixed OPD (top-128 forward KL, learning rate 10^{-5}, the OPD code path). For each prompt with a PS-OPD prefix, it adds one teacher-forced cross-entropy example on that same $r = 0.5$ prefix, weighted by its number of rollouts to match the exposure of PS-OPD. Its total weight is $\lambda(1 - e) = 0.3 \times 0.995$ that of the OPD term (logged ratio 0.2985 at every training step of every seed). Micro-batch gradients are summed so that clipping binds as for PS-OPD (Appendix A). No training step of this control or of OPD was skipped for a non-finite gradient. Relative to PS-OPD, this control thus drops the prefix from the rollout context and restores OPD's full generation allowance together. The *sequential* control trains off-policy distillation for 10 training steps (256-sequence batch, 10^{-5}), then OPD (top-128 forward KL, 10^{-5}) for 40 steps from this step-10 SFT model. We compare it at equal total training steps, counting the warm start (per-seed scores in Table 24).

Decoupled control results. Over training steps 1–10 it averages 22.84%, leading OPD by 2.18 points ($t = 15.8$) and RH by 1.43 ($t = 7.5$). PS-OPD leads OPD by 2.59 points over the same steps, so most of that lead is reproduced. The remaining 0.42-point gap to PS-OPD appears in all three seeds but is unresolved ($t = -3.7$). From step 3 to 4 its accuracy changes by -2.6 points, close to OPD's -3.2 and shallower than the -5.5 of PS-OPD by an unresolved 2.9 ($t = 2.9$). Over steps

Table 24: **Per-seed accuracy of the controls** (available teacher, NVIDIA cluster, learning rate 10^{-5}): Avg. avg@16 accuracy in % per seed and their mean, by training step. The sequential control trains off-policy distillation for 10 training steps, then OPD for 40. Its columns count total training steps, so its step-10 entry ends the warm start.

control	seed	1	2	3	4	5	6	7	8	9	10	20	30	40	50
Decoupled	1	20.2	22.0	20.4	19.1	20.9	23.5	24.6	25.1	25.5	27.2	27.7	26.8	26.9	27.2
	2	20.3	21.8	22.8	19.5	21.3	22.3	24.2	25.8	25.5	25.4	25.6	26.4	26.5	26.3
	3	19.8	21.5	22.1	18.8	20.2	21.3	25.0	25.1	26.9	27.0	27.0	24.6	27.4	28.5
	mean	20.1	21.8	21.7	19.1	20.8	22.4	24.6	25.3	26.0	26.5	26.8	25.9	26.9	27.4
Sequential	1										24.4	23.4	25.9	25.0	24.1
	2										23.7	22.5	24.9	26.1	24.8
	3										24.4	20.0	23.5	24.4	24.2
	mean										24.2	21.9	24.8	25.2	24.4

Table 25: **Controls minus each arm** (available teacher, NVIDIA cluster, paired within seeds 1–3). Entries: mean difference in points (t). Last column: 95% t interval (CI) of the Avg. difference (two degrees of freedom). Bold: interval excludes zero ($|t| > 4.3$, unadjusted). Per-benchmark columns are descriptive. PS-OPD is $\lambda = 0.3$, RH $\lambda = 0$. The sequential control is compared at equal total training steps.

minus	steps	AIME24	AIME25	AIME26	Avg.	Avg. CI
Decoupled control minus						
PS-OPD	1–10	**−0.64 (−8.2)**	+0.15 (0.7)	**−0.77 (−15.3)**	−0.42 (−3.7)	[−0.91, +0.07]
	20–50	**−2.71 (−5.6)**	−0.19 (−0.3)	−1.19 (−3.4)	**−1.36 (−5.1)**	[−2.52, −0.20]
	50	−2.71 (−2.0)	−0.35 (−0.3)	−1.72 (−1.6)	−1.59 (−4.1)	[−3.28, +0.09]
RH	1–10	**+1.57 (4.4)**	**+1.74 (9.5)**	+0.99 (3.2)	**+1.43 (7.5)**	[+0.61, +2.25]
	20–50	−1.15 (−2.9)	−0.16 (−0.4)	−0.59 (−0.9)	−0.63 (−2.3)	[−1.80, +0.53]
	50	−0.76 (−0.7)	+0.97 (1.0)	+0.43 (0.5)	+0.21 (1.1)	[−0.63, +1.05]
OPD	1–10	**+2.91 (10.5)**	**+1.48 (9.3)**	**+2.14 (7.0)**	**+2.18 (15.8)**	[+1.58, +2.77]
	20–50	−0.90 (−3.2)	+1.06 (3.2)	+0.97 (1.8)	+0.38 (1.4)	[−0.80, +1.55]
	50	−0.21 (−0.9)	+0.49 (0.8)	+1.29 (0.7)	+0.52 (0.8)	[−2.19, +3.23]
Off-policy distillation	1–10	+0.13 (0.4)	+0.34 (1.5)	**+0.56 (5.5)**	+0.34 (1.7)	[−0.55, +1.24]
	20–50	+0.61 (2.7)	+1.44 (2.3)	**+1.74 (7.5)**	+1.26 (3.5)	[−0.28, +2.81]
	50	−1.18 (−1.2)	+0.69 (0.7)	+2.16 (1.4)	+0.56 (2.2)	[−0.51, +1.62]
Sequential control minus						
PS-OPD	20–50	**−4.48 (−10.1)**	−3.58 (−3.0)	**−4.09 (−10.7)**	**−4.05 (−6.1)**	[−6.92, −1.18]
	50	**−4.86 (−7.3)**	−3.47 (−4.1)	−5.39 (−4.2)	**−4.57 (−9.7)**	[−6.59, −2.55]
OPD	20–50	**−2.67 (−4.5)**	−2.33 (−2.6)	**−1.94 (−5.6)**	**−2.31 (−5.6)**	[−4.08, −0.55]
	50	**−2.36 (−4.4)**	−2.64 (−4.0)	−2.37 (−2.1)	−2.46 (−3.7)	[−5.29, +0.37]
Off-policy distillation	20–50	−1.16 (−1.3)	−1.94 (−2.3)	−1.17 (−2.9)	−1.43 (−2.1)	[−4.36, +1.51]
	50	−3.33 (−2.1)	−2.43 (−3.4)	−1.51 (−0.9)	−2.42 (−3.1)	[−5.76, +0.91]

20–50 it averages 26.75%, unresolved against OPD (+0.38, $t = 1.4$) and RH (−0.63, $t = -2.3$), and 1.36 points below PS-OPD ($t = -5.1$, 95% CI $[-2.5, -0.2]$). At step 50 it scores 27.35%, below PS-OPD in every seed by an unresolved 1.59 points ($t = -4.1$). None of the 18 Avg. tests of both controls in Table 25 is significant after Holm correction.

Interpretation. These results show that, on these seeds, the average gain of PS-OPD over training steps 1–10 comes largely from auxiliary imitation, and most of it does not need the imitated prefix as rollout context. The two curves still differ in shape over these steps (Appendix B.4). The missing later lead is consistent with the later benefit depending on coupling the imitation to the rollout, but four caveats leave this unestablished. First, we examined the later steps after selecting $\lambda = 0.3$ on these seeds, so the contrast is exploratory and favours PS-OPD. Second, the step-50 difference alone is unresolved. Third, per benchmark, the later lead of PS-OPD is resolved on AIME24 only. Finally, rollout context and generation allowance change together. We did not repeat this contrast on fresh seeds.

Sequential control results. At the end of the warm start (total 10) the control scores 24.18%, unresolved against OPD at step 10 (−0.4, $t = -0.9$). Ten OPD training steps later it is lower,

at 21.93% (an unresolved change, $t = -2.0$), and below OPD (-3.9, $t = -6.2$). Intermediate checkpoints were not scored, so the timing of this change is unknown. Its mean is below that of PS-OPD at every total ($t = -6.3$, -4.9, an unresolved -3.0, -6.3 and -9.7 at totals 10–50). Over totals 20–50 it averages 24.06%, below PS-OPD by 4.05 points ($t = -6.1$), OPD by 2.31 ($t = -5.6$), and off-policy distillation by an unresolved 1.43 ($t = -2.1$). We ran only this split, so this control does not test the SFT-before-OPD benefit that Li et al. (2026) report for other splits or pairs. The warm start is its own 10-step SFT run, not the step-10 checkpoint of the off-policy distillation arm. The two differ by an unresolved -1.3 points ($t = -2.8$), a gap of the size seen between same-seed off-policy distillation runs (Appendix B.3). Table 10 gives the compute of both controls, warm start included.

K.2 Confirmatory fresh-seed test and provenance manifest

A test planned before any fresh run confirms the later lead of PS-OPD over OPD on three new seeds.

Confirmatory result. We selected $\lambda = 0.3$ by its step-50 mean on the three seeds and 89 problems of the five-arm comparison. Before submitting any fresh run, we fixed the arms (PS-OPD at $\lambda = 0.3$, OPD, off-policy distillation), the contrasts (steps 20–50 and step 50) and the number of seeds, and we report the result whatever its sign. The primary test is PS-OPD ($\lambda = 0.3$) minus OPD in Avg., averaged over checkpoints 20–50 within each of seeds 4–6. It confirms the lead if the paired mean is positive with a 95% t interval excluding zero, with 1 point fixed as the smallest effect of interest. The per-seed differences are $+1.47$, $+2.85$ and $+1.91$ points (mean $+2.08$, $t = 5.12$, 95% CI $[+0.33, +3.83]$). The lead is thus confirmed, although the lower bound does not exclude effects below 1 point. The fresh seeds are new training runs, scored on the 89 problems used to select the weight. In secondary contrasts on these seeds, PS-OPD leads OPD by $+1.40$ points at step 50 alone ($[+0.20, +2.61]$), and off-policy distillation by $+2.20$ over steps 20–50 ($[+0.95, +3.45]$) and $+2.43$ at step 50 ($[+0.85, +4.02]$). Pooling seeds 1–6 (5 degrees of freedom, $t_{0.975} = 2.571$), which include the selection seeds, gives $+1.91$ points over 20–50 ($[+1.31, +2.50]$) and $+1.76$ at step 50 ($[+1.10, +2.42]$) against OPD, and $+2.41$ ($[+1.99, +2.84]$) and $+2.29$ ($[+1.57, +3.01]$) against off-policy distillation. All nine fresh runs completed without reruns. The only deviation is that the off-policy distillation runs saved only model shards, which does not affect scores.

Provenance manifest. A manifest released with the code links each Avg. and per-benchmark value it lists (table entries, plotted values, and values quoted in text and captions) to its seed-level scores. Each entry holds one score with its cluster (AMD or NVIDIA), evaluation file and checksum, benchmark, Avg., problem-exclusion version, samples per problem, per-benchmark problem counts, seed, checkpoint and, where they exist, training and evaluation job identifiers. A seed assembled from two runs lists both segments. A script recomputes every listed value from these files. A few values read directly from evaluation files, such as bootstrap intervals, are not listed.

Statistics and cost. Tables 6 and 25 give the seed-paired intervals, and Appendix J the multiple-testing correction. Appendix B.4 reports training steps and GPU-hours to fixed accuracy thresholds, cost per training step (Table 10) and accuracy at common GPU-hour budgets (Table 11).

K.3 Forward-horizon baseline

FH tests whether grading only the first H tokens of the student's own rollout, without teacher text, improves on OPD. Averaged over the later training steps, it is above OPD and an unresolved margin below PS-OPD.

Design. A teacher prefix changes *where the rollout starts*, whereas FH changes *which part of the rollout is graded*. It adapts the prefix-restricted OPD of Zhang et al. (2026), who distill with the sampled-token reverse KL (Equation 11), but uses the objective of the arms it is compared with. Unlike their early-stopped rollouts, ours are never truncated. The loss is masked after H tokens, so the rest of the pipeline, including how rollouts end, is unchanged from OPD. FH thus tests their loss restriction, not their compute saving, and we make no GPU-hour comparison (Table 10, note b). It shares OPD's unprefixed data, objective, learning rate, budget, seeds and checkpoint schedule (Table 4), so its control is full-horizon OPD. The learning rate is the shared 10^{-5}, not the 5×10^{-5} that Zhang et al. (2026) use with base students. The FH arms of Table 1 are separate AMD-cluster runs, not compared across clusters.

Positive control. A horizon that narrowed the token-mean denominator but not the graded span would leave FH numerically identical to OPD. We therefore logged the graded fraction per micro-batch. With $H = 4{,}096$ it is below 100% from the start (40.0%, 43.3% and 42.0% at training step 1 for seeds 1–3) and between 29% and 50% throughout.

Results. Table 4 gives three-seed accuracy and Table 6 the seed-paired comparisons with PS-OPD. Over steps 20–50 FH averages 27.03%, 0.66 points above OPD ($t = 7.2$) and an unresolved 1.08 points below PS-OPD ($t = 3.1$). At step 50 it scores 26.55%, unresolved against OPD (-0.28, $t = -0.5$).

K.4 TrOPD BASELINE

TrOPD (Xing et al., 2026), the published method closest to PS-OPD, was studied on this same teacher–student pair. Like PS-OPD, it leads OPD early in training, but this lead does not persist, and PS-OPD is above it later.

Design. We run the released implementation (https://github.com/Xingrun-Xing2/TrOPD), changing only the settings common to all our arms: student, teacher, DAPO data, learning rate 10^{-5} (the release uses 5×10^{-6}), optimizer settings including weight decay 0.01 (the release uses 0.1), 32 prompts with 8 rollouts each, a $1{,}536/27{,}648$ prompt/rollout token budget, 50 training steps, and the checkpoint schedule. The algorithm follows the release throughout. The TrOPD paper (its Section 4.3) describes a teacher-generated prefix annealed to zero, but the released code, which we run, grades teacher-forced sequences separately from the student rollouts.

- *Student rollouts.* Let $s = \log p_S$ and $t = \log p_T$ be the student's and teacher's log-probabilities of a sampled token before the training step. The token is accepted with probability $\min(1, p_T/p_S)$ and receives advantage $t - s$. Otherwise it receives $-F$, where F is the forward KL over the teacher's top-64 tokens. All advantages are frozen before the first optimizer step.

- *Teacher-forced sequences.* Each prompt adds one sequence holding the teacher's complete response to that problem, not selected on correctness. Only the first $\lfloor \alpha L \rfloor$ of its L tokens are graded, with $\alpha = \frac{1}{2}\big(1 + \cos(\pi g/200)\big)$ at training step g and advantage $0.001\,(s - t)$.

- *Optimization.* Dual-clip PPO with clip range $0.2/0.28$ and $c = 10$, token-mean aggregation, and no loss clamp. We keep the release's per-micro-batch loss normalization and its mini-batch split, so the 256 rollouts and 32 teacher-forced sequences give two optimizer steps per training step.

Consequences of shared settings. The release fixes the anneal length at 200 training steps, so within our 50-step budget the graded share of each teacher-forced sequence falls only from 100% to 85%. The release's teacher responses are not public, so we use Skywork traces on our data. Teacher-forced sequences add 32 sequences per training step to both teacher scoring and student training, so we also compare TrOPD at equal cost (Appendix B.4).

Verification. Before training, a unit test on synthetic inputs matched our implementation exactly to the released code in the advantage, top-64 KL construction and loss on a short leftover mini-batch. At training step 1 of seeds 1–3, the acceptance rate matched its expectation $\mathbb{E}[\min(1, p_T/p_S)]$ (85.57%, 86.42% and 86.89% against 85.57%, 86.42% and 86.90%). The minimum F was negative, ruling out a hidden clamp, and teacher-forced tokens were 10.2–10.8% of the total. Finally, start-of-step and current-policy F agreed exactly on the first optimizer step and differed on the second.

Results. Table 4 gives the three-seed means (per seed in Table 5) and Table 6 the paired comparisons. Over steps 1–10 TrOPD averages 22.81%, 2.15 points above OPD ($t = 12.1$) and an unresolved 0.45 points below PS-OPD ($t = 2.7$). It does not dip at step 4 on any benchmark. Over steps 20–50 TrOPD averages 26.23%, unresolved against OPD (-0.14, $t = -0.3$) and 1.88 points below PS-OPD ($t = 5.0$).

L Extended Background and Related Work

This appendix defines the sampled reverse-KL estimator and reviews related work. The estimator (Lu & Thinking Machines Lab, 2025) uses only the sampled token y_t, i.e.,

$$D_{\mathrm{KL}}\big(\pi_{\bar\theta}(\cdot \mid c_t)\|\pi_T(\cdot \mid c_t)\big) = \mathbb{E}_{y_t\sim\pi_{\bar\theta}(\cdot|c_t)}\left[\log \frac{\pi_{\bar\theta}(y_t \mid c_t)}{\pi_T(y_t \mid c_t)}\right],$$
$$A_t = \log \frac{\pi_T(y_t \mid c_t)}{\pi_{\bar\theta}(y_t \mid c_t)}, \qquad \nabla_\theta \widehat{\mathcal{L}}_{\mathrm{RKL}} = -\frac{1}{L}\sum_{t=1}^{L} A_t \nabla_\theta \log \pi_\theta(y_t \mid c_t), \qquad y_t \sim \pi_{\bar\theta}(\cdot \mid c_t). \tag{11}$$

L.1 Prior evidence on prefixes in both directions

Prior work reports each effect of Section 3 on other model pairs, apart from TrOPD's training ablation on ours, and one study reports both. Our test measures both directions on one pair, with the available teacher.

Prefixes come from evaluation responses generated on the NVIDIA cluster, without the 8,192-token cap of PS-OPD. Continuations and unprefixed responses are generated on the AMD cluster under the evaluation protocol, with the prefix counted against the token budget. As controls, each model continues its own response, cut at half its reasoning span or at half another sample's on the same problem. The first cut is length-dependent, so it need not match the unprefixed accuracy. The second should, as a check of the prefix insertion. On every benchmark, both effects grow with the cut, while every own-response interval covers zero (Table 26). Thus, the effect follows who wrote the prefix. Because it scores only final answers, the test measures a prefix's usefulness as context and is only a proxy for the quality of the teacher's token-level feedback.

Teacher-written prefixes help a weaker continuer. Untrained Qwen3-1.7B-Base and Qwen3-8B-Base models continuing truncated Qwen3-8B solutions gain AIME-24 accuracy sharply over the first few thousand teacher tokens, then flatten (Zhang et al., 2026, Section 2.3). In Speculative Thinking (Yang et al., 2025), a 32B reasoning model that takes over reflective steps raises a 1.5B model's MATH500 accuracy from 83.2% to 89.4%. UFT (Liu et al., 2025, Section 3) and Prefix-RFT (Huang et al., 2026) train by RL on the policy's continuation of a solution or demonstration prefix. Prefix-RFT's prefixed rollouts earn higher rewards, by a margin that shrinks during training (their Figure 4). TrOPD (Xing et al., 2026, Section 4.3) builds its OPD trajectories the same way on our pair. This guidance raises the AIME-24/AIME-25/AMC-23 average of each of its three variants by 0.8–1.1 points (their Table 5).

Student-written prefixes hurt a stronger continuer. Li et al. (2026, Section 6.1, Figure 11b) let the teacher continue student rollouts longer than 16K tokens on DAPO-Math-17K prompts, truncated at several depths. Its accuracy advantage over the student falls monotonically from 37 points at a 1K-token prefix to 2 points at 16K. Qwen3-14B solves 62.1% of GPQA-Diamond alone, but 46.0% when continuing randomly truncated Qwen3-1.7B rollouts (Zhu et al., 2026, Section 5.1, Appendix A.22). Xie et al. (2026) find that the teacher's chance of a correct answer falls toward the student's as the student prefix grows. For a Qwen3-0.6B student with a Qwen3-4B-Instruct-2507 teacher, they also find OPD on only the first 30% of rollout tokens matching full OPD, while OPD on the last 30% barely learns. Zhou et al. (2026) call this off-policy teacher decay. On MATH-500, Qwen3-1.7B continuing Qwen2.5-Math-1.5B rollouts falls to the student's accuracy within about 300 student tokens. At the token level, the teacher's top-1/top-2 margin and the spread of advantages decay along student rollouts (Liu et al., 2026a). Teacher–student log-probability gaps widen at later positions (Fu et al., 2026a). The teacher's next-token distribution also becomes less confident and discriminative, which Liu et al. (2026b) call supervision fidelity decay. All these are cross-model settings with a stronger teacher.

Both directions. Relay-OPD (Xu et al., 2026, Sections 1 and 3.2, Figure 2) lets Qwen3-4B-Instruct-2507 take over from a Qwen3-1.7B student at reflection triggers on DAPO problems. Replacing only the reflection token at each trigger raises accuracy from 27.7% to 35.0%. Moving a fixed-length takeover from the earliest to later triggers lowers it from 42.0% to 34.0% and then 29.5%. With 28.5% of tokens written by the teacher, relayed trajectories reach 43.6%, against 60.6% for the teacher alone. The authors attribute the loss from late takeovers to the teacher being pulled along by the student's context.

Table 26: **Effect of the prefix, per benchmark** (available teacher; Figure 2 plots the Avg. block). Accuracy (%) of a frozen model continuing a response cut at a fraction (*cut*) of its reasoning span. *none*: unprefixed responses sampled for this test on the AMD cluster, not the scores of Appendix C. *own*: the model's own response, cut at 50% of its span (*same cut*) or of another sample's (*partner cut*). *Change*: points from *none* (*other − own*: from *own, same cut*), with 95% bootstrap intervals over problems, descriptive per benchmark. AIME24 and AIME25 have 30 problems, AIME26 29; temperature 1.0, top-p 1.0.

		student continues		teacher continues	
prefix	**cut**	**acc.**	**change [95% CI]**	**acc.**	**change [95% CI]**
AIME24					
none	–	24.2	–	67.9	–
other model	25%	32.7	+8.5 [+3.3, +14.6]	58.8	−9.2 [−13.7, −4.8]
other model	50%	41.3	+17.1 [+10.8, +23.8]	46.9	−21.0 [−29.8, −12.7]
other model	75%	55.0	+30.8 [+22.3, +39.8]	39.6	−28.3 [−37.9, −19.6]
own, same cut	50%	24.4	+0.2 [−4.2, +4.8]	67.3	−0.6 [−4.4, +3.1]
own, partner cut	50%	26.0	+1.9 [−1.5, +5.0]	69.8	+1.9 [−1.9, +5.8]
other − own	50%	–	+16.9 [+11.3, +22.9]	–	−20.4 [−30.0, −11.5]
AIME25					
none	–	21.3	–	52.5	–
other model	25%	26.9	+5.6 [+2.5, +9.6]	42.5	−10.0 [−16.0, −4.6]
other model	50%	33.8	+12.5 [+6.3, +19.6]	35.4	−17.1 [−24.8, −10.0]
other model	75%	38.5	+17.3 [+10.2, +25.4]	29.2	−23.3 [−32.7, −14.8]
own, same cut	50%	18.8	−2.5 [−6.7, +1.3]	50.2	−2.3 [−5.6, +0.6]
own, partner cut	50%	19.4	−1.9 [−4.4, +0.6]	52.9	+0.4 [−4.0, +4.8]
other − own	50%	–	+15.0 [+7.9, +22.7]	–	−14.8 [−23.1, −7.5]
AIME26					
none	–	13.4	–	59.3	–
other model	25%	22.6	+9.3 [+4.3, +15.1]	54.5	−4.7 [−8.2, −1.3]
other model	50%	31.9	+18.5 [+10.1, +27.6]	39.7	−19.6 [−25.9, −13.6]
other model	75%	42.2	+28.9 [+19.0, +39.2]	26.5	−32.8 [−41.2, −24.4]
own, same cut	50%	13.6	+0.2 [−3.4, +4.1]	59.1	−0.2 [−2.6, +2.2]
own, partner cut	50%	14.0	+0.6 [−3.9, +5.6]	59.1	−0.2 [−2.4, +1.9]
other − own	50%	–	+18.3 [+10.6, +26.9]	–	−19.4 [−25.4, −13.8]
Avg.					
none	–	19.6	–	59.9	–
other model	25%	27.4	+7.8 [+5.1, +10.8]	51.9	−8.0 [−10.6, −5.4]
other model	50%	35.6	+16.0 [+11.8, +20.4]	40.6	−19.2 [−23.6, −15.1]
other model	75%	45.3	+25.7 [+20.6, +31.0]	31.8	−28.1 [−33.4, −23.3]
own, same cut	50%	18.9	−0.7 [−3.0, +1.6]	58.9	−1.0 [−2.9, +0.7]
own, partner cut	50%	19.8	+0.2 [−1.9, +2.4]	60.6	+0.7 [−1.4, +2.8]
other − own	50%	–	+16.7 [+12.7, +21.0]	–	−18.2 [−22.8, −13.8]

L.2 Matched and available teachers

Our matched teacher, JustRL (He et al., 2025a), is RL-finetuned from the student checkpoint itself, DeepSeek-R1-Distill-Qwen-1.5B. Building such a teacher for a new student requires a separate RL stage, here 4,380 RL steps. Our available teacher, Skywork-OR1-Math-7B, is RL-finetuned from the larger DeepSeek-R1-Distill-Qwen-7B (He et al., 2025b). It shares the student's distillation lineage, not its checkpoint or scale. In Li et al. (2026, Section 3.2), it is the stronger of two teachers for our student, transferring more than the checkpoint it was RL-finetuned from. There, teachers given further RL, Skywork among them, also yield larger gains than teachers from the same pipeline without it. Thus, RL finetuning by itself does not make a teacher matched.

Controlled comparisons favor the matched teacher. With our student, OPD from JustRL-1.5B recovers more than 80% of the gap to that teacher, whereas the larger and slightly stronger DeepSeek-R1-Distill-Qwen-7B yields no improvement (Li et al., 2026, Section 4.1). A Qwen3-1.7B student learns more from a copy of its own checkpoint trained with DAPO for 200 steps than from the comparably accurate Qwen3-8B (Zhu et al., 2026, Section 6.2). DeepSeek-V4 (DeepSeek-AI, 2026,

Section 5.1) and Kimi K3 (Kimi Team, 2026, Section 4.1) distill RL-trained specialists into one student by multi-teacher OPD at scale, without a controlled teacher comparison. In our runs, PS-OPD reaches 89–92% of the matched teacher's Avg. in 50 training steps (Table 2). However, under both forward-KL objectives OPD falls below the untrained student mid-training (Appendix C.5), so a matched teacher is not uniformly the easier case.

L.3 TOP-k OBJECTIVES IN PRIOR WORK

We present none of the four objectives of our grid (Table 2) as new. The grid is one controlled comparison under a single implementation. Fu et al. (2026a) use teacher top-k local support matching with a truncated reverse-KL objective. This is the published counterpart of our reverse KL over the teacher's top-k, not that objective itself. Reverse KL over the student's top-k likewise exists in the training framework our code builds on. PG-OPD (Zhao et al., 2026) scores short student probes by teacher–student top-k overlap to decide which rollouts to continue. SPOT (Qu et al., 2026) adds outcome-calibrated soft targets at a few probed positions.