UNDERSTANDING THE SYNERGY BETWEEN SFT, RLVR, AND OPD IN LLM POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern LLM post-training composes supervised fine-tuning (SFT), reinforcement learning with verifiable rewards (RLVR), and on-policy distillation (OPD) into multi-stage pipelines, yet these stages are typically designed and evaluated in isolation. We show that this composition is consequential: a stage that improves the current model can make the next stage less effective. Through controlled experiments with Qwen3 models on math and science reasoning, we first characterize OPD across nine student–teacher pairs spanning $2\times$ to $53\times$ parameter ratios and show that OPD effectiveness depends on student–teacher compatibility rather than teacher scale alone. The surrounding stages of OPD reshape this compatibility in three ways: (1) A brief SFT warm-up improves subsequent OPD, while an RLVR-strengthened student regresses under distillation from the same teacher. (2) Adapting the teacher with RLVR raises downstream OPD accuracy in proportion to the capability it adds. Following these two interventions, we find that combining teacher adaptation and student warm-up alone raise average OPD accuracy from 29.2% to 43.8% (**50%** relative improvement) after the same number of distillation steps, with additional preparatory training. (3) At comparable accuracy, OPD leaves a stronger initialization for downstream RLVR than SFT, with a gap that widens as RL compute scales. Our results suggest that each post-training stage should be chosen not only for the capability it adds, but for the *learning interface* it creates for the next stage.

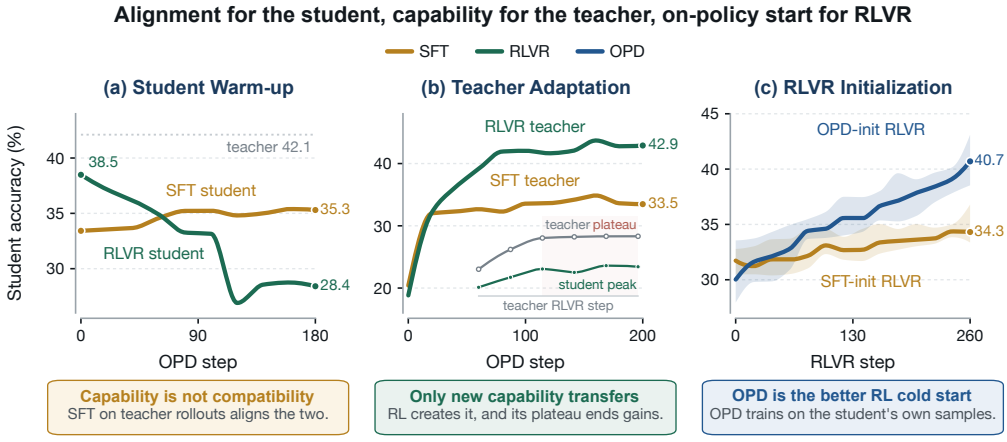


Figure 1: **The outcome of each post-training stage depends on the learning state it inherits.** **(a)** An RLVR-warmed student (green) starts stronger but regresses under OPD, while a briefly SFT-warmed student (gold) improves steadily from the same teacher. **(b)** An RLVR-adapted teacher transfers substantially more than a SFT-adapted teacher of the same origin. *Inset:* teacher accuracy along its RLVR training (gray) and final OPD accuracy of a student distilled from the teacher at each step (green); both plateau together, so student gains stop when the teacher's do. **(c)** At comparable starting accuracy, OPD-initialized models (blue) reach higher final accuracy under subsequent RLVR than SFT-initialized models, and the gap widens with RL compute.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Recent advances in post-training have substantially improved the reasoning capabilities of large language models (LLMs) (Guo et al., 2025; Lambert et al., 2025; Yang et al., 2025; Xiao et al., 2026; Zeng et al., 2026). Three post-training objectives drive this progress: supervised fine-tuning (SFT) on expert demonstrations (Ouyang et al., 2022; Wei et al., 2022), reinforcement learning with verifiable rewards (RLVR) (Lambert et al., 2025; Shao et al., 2024) with sparse outcome-level feedback on model-generated trajectories, and on-policy distillation (OPD) (Lu & Lab, 2025; Yang et al., 2025) with dense token-level teacher supervision on the student's own trajectories. In practice these stages are rarely applied in isolation: SFT, RLVR, and OPD are composed sequentially, each reshaping the policy that the next stage receives. Yet, despite the rapid adoption of multi-stage post-training pipelines, how these stages interact and in what order they should be composed remains poorly understood. We show that this problem is consequential: *the effect of a post-training stage depends on the learning state created by the stage before it*, and a stage that improves the current model can simultaneously make the subsequent stage less effective.

We begin by characterizing what determines OPD effectiveness in isolation (§3). Consistent with concurrent findings (Li et al., 2026; Zhu et al., 2026; Wang et al., 2026), we observe across nine Qwen3 student–teacher pairs spanning $2\times$ to $53\times$ parameter ratios that stronger teachers do not necessarily transfer better under OPD. Moreover, reverse KL continues to decrease after accuracy has already peaked, so closer teacher matching does not imply better task performance either. We show that the *reinforceable fraction*, the share of student-sampled tokens that receive a positive teacher signal, can serve as a complementary diagnostic that captures student–teacher compatibility where scalar divergence does not. Together, these results suggest that OPD effectiveness depends on the student–teacher pairing at initialization. Since SFT, RLVR, and OPD each reshape that pairing, their ordering cannot be treated as arbitrary.

We call this the *composition problem*: choosing the order, duration, and handoff points of SFT, RLVR, and OPD in a post-training pipeline (§4). Current frontier pipelines compose these stages in different ways: some warm-up the student with SFT and then apply RLVR (Guo et al., 2025; Shao et al., 2024), while recent work inserts OPD before RLVR (Xiao et al., 2026; Zeng et al., 2026). On the teacher side, some pipelines strengthen the teacher with RLVR before distillation (Yang et al., 2025), and others interleave domain-specific RL with OPD to consolidate multiple specializations into a single model (Yang et al., 2026). Yet these choices are rarely compared under a shared experimental protocol. Does improving the student with SFT or RLVR help subsequent OPD, and much SFT warm-up is useful before it begins to hinder distillation (Li et al., 2026; Zhu et al., 2026)? How do teacher accuracy gains from RLVR translate into student gains, compared with SFT on verified solutions (Zelikman et al., 2022; Zhu et al., 2026)? And when OPD is followed by RLVR, does distillation leave a better initialization than SFT (Aphale & Liu, 2026; Shen, 2026)?

To answer these questions, we design controlled experiments that vary the student state, teacher state, and order of training stages. Our main study uses a Qwen3-4B-Base student and a Qwen3-14B teacher, evaluated on six math and one science benchmark. Within each comparison, we keep the surrounding data, optimization setup, and evaluation protocol fixed so that the impact of each stage can be isolated. From these experiments, we derive the following findings:

- Teacher scaling does not consistently improve distillation across student sizes, model families, objectives, and training distributions, highlighting the importance of student–teacher compatibility.
- Brief SFT warm-up improves distillability while bringing the student closer to the teacher, whereas a stronger RLVR-prepared student can regress under distillation from the same teacher (Figure 1a).
- Teacher adaptation with RLVR improves downstream OPD as teacher accuracy rises, with diminishing returns once teacher performance plateaus (Figure 1b).
- At comparable starting accuracy, OPD provides a more effective initialization than SFT for subsequent RLVR and aggregate policy entropy alone does not account for this advantage (Figure 1c).

Across all experiments, the same pattern emerges: the outcome of a stage depends on the state it inherits, so a stage should be judged not only by the capability it adds but also by the state it hands to the next stage, which we call its *learning interface*. In addition, these findings motivate a simple composition recipe: **adapt the teacher with RLVR, briefly warm the student with SFT, and then distill**. Combining teacher adaptation and student warm-up raises average final OPD accuracy from 29.2% to 43.8% (50% relative improvement) at the same OPD step budget, with additional preparatory training.

We summarize the overall contributions of our paper as follows:

- We characterize OPD across model scales, families, objectives, and training distributions under a shared protocol, confirming that effective distillation depends on student–teacher compatibility rather than teacher scale alone (§3).
- We present a systematic controlled study of how SFT, RLVR, and OPD interact, showing that student capability and distillability can move in opposite directions, that teacher adaptation transfers in proportion to the capability it adds, and that OPD provides a stronger RLVR initialization than SFT at comparable accuracy (§4).
- We derive an empirically grounded composition recipe, where we first adapt teacher with RLVR, then briefly warm-up the student with SFT, and then transition to OPD. This recipe demonstrate significant improvement within the settings studied (§5).

By studying these interactions in open-weight models at scale, we aim to make LLM post-training dynamics more accessible to the research community and to inform the design of future pipelines.

2 Preliminaries

We define the post-training objectives studied throughout the paper and the quantities we use to characterize student–teacher transfer. Appendix A.1 gives full objective details, Appendix A.2 provides metric definitions, aggregation, and derivations, and Appendix A.3 extends the related-work discussion.

Post-Training Objectives. Given a labeled dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}$ of trajectories, **SFT** minimizes token-level cross entropy: $\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\sum_t \log \pi_{S_\theta}(y_t \mid x, y_{<t})\big]$. The supervision here is off-policy: prefixes and targets are fixed regardless of what the evolving student generates. **RLVR** optimizes the student on its own rollouts using a scalar outcome reward $r(x, y) \in \{0, 1\}$. We use GRPO (Shao et al., 2024), which samples G responses per prompt and normalizes within the group: $\hat{A}_i^{\text{RLVR}} = (r(x, y^{(i)}) - \mu_{\text{group}})/\sigma_{\text{group}}$. The supervision is on-policy but sparse: one outcome reward per self-generated trajectory, shared by every token. **OPD** also trains on student-generated trajectories with dense teacher supervision at every prefix (Gu et al., 2023; Agarwal et al., 2024; Lu & Lab, 2025). Our default formulation uses a sampled-token policy-gradient surrogate for token-level reverse-KL matching (Lu & Lab, 2025). The objective minimizes reverse KL between the student p_t and teacher q_t at each prefix: $\mathcal{L}_{\text{OPD}} = \mathbb{E}_{x,y\sim\pi_S}\big[\frac{1}{|y|}\sum_t D_{\text{KL}}(p_t \| q_t)\big]$. Our default implementation uses a sampled-token policy-gradient estimator optimized with the same clipped importance-ratio objective used for RLVR. We also evaluate a dense top-k forward-KL objective for robustness (Appendix A.1).

The three methods differ along two axes: SFT provides dense supervision on fixed trajectories, RLVR provides sparse outcome supervision on student trajectories, and OPD provides dense teacher supervision on student trajectories.

Measuring Transfer. We characterize student–teacher transfer with three quantities (formal definitions in Appendix A.2). Let A_S, $A_{S'}$, and A_T be the accuracies of the original student S, the distilled student S', and the teacher T. **Distillation effectiveness** $\text{DE} = (A_{S'} - A_S)/(A_T - A_S)$ is the fraction of the accuracy gap recovered. The **reinforceable fraction** ρ_0^+ is the share of student-sampled tokens receiving a positive teacher signal at initialization. It captures how often the teacher reinforces the student's own behavior, providing a compatibility signal that teacher accuracy alone cannot reveal. **Normalized reverse KL** $\kappa_s = \widehat{\text{rKL}}_s/\widehat{\text{rKL}}_0$, where $\widehat{\text{rKL}}_s$ estimates $\text{KL}(\pi_{S,s} \| \pi_T)$ on student samples at step s and tracks on-policy divergence during training. Together, these capture outcome, initialization-time compatibility, and training dynamics.

From Individual Stages to Composition. Several concurrent studies have characterized OPD dynamics in isolation. Li et al. (2026) identify thinking-pattern consistency and teacher-knowledge novelty as conditions for successful transfer. Wang et al. (2026) characterize OPD as an exploration catalyst and attribute failures to signal corruption. Zhu et al. (2026) document pitfalls of the reverse-KL approximation. On adjacent stages, prior work has shown that SFT cold starts improve OPD (Li et al., 2026; Zhu et al., 2026), that RLVR teacher adaptation aids downstream distillation (Zhu et al., 2026), and that pre-RL entropy predicts post-RLVR outcomes (Aphale & Liu, 2026). What remains less understood is how these stages *compose*: how changing one stage alters the learning conditions faced by the next. We address this gap by first characterizing OPD under a controlled protocol (§3), then studying how SFT, RLVR, and OPD reshape student–teacher compatibility when composed (§4). A detailed discussion of prior work is in Appendix A.3.

3 OPD Dynamics at Scale

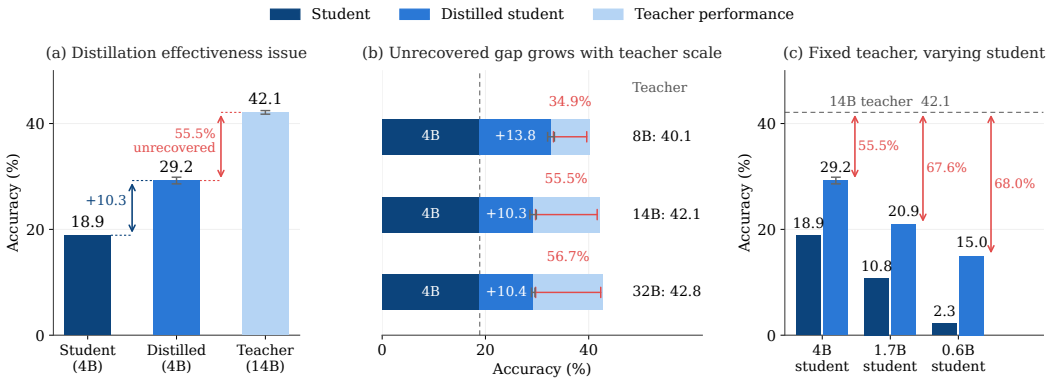


Figure 2: **A larger teacher does not guarantee better transfer.** Red lines show the unrecovered gap, 100 − DE (%). **(a)** Qwen3-4B-Base gains 10.3 points from Qwen3-14B but leaves 55.5% of the teacher's advantage unrecovered. **(b)** For the 4B student, DE decreases as teacher size increases from 8B to 32B. **(c)** With the 14B teacher fixed, smaller students recover a smaller share of the gap.

Setup. Before studying how stages compose, we characterize OPD in isolation to establish the baselines and diagnostics that §4 builds on. We distill Qwen3-{0.6B, 1.7B, 4B}-Base students from Qwen3-{8B, 14B, 32B} post-trained teachers (nine pairs, 2×–53× parameter ratios) using DAPO-Math-17K (Yu et al., 2025) with sampled-token reverse-KL distillation. All runs train for one epoch with 64 prompts and four responses per prompt per batch. Unless otherwise stated, we compare at step 200; 4B results average three seeds, and 0.6B and 1.7B use one run per teacher. We report `mean@4` accuracy averaged over AIME 2024, AIME 2025, AMC 2023, MATH-500, Minerva Math, OlympiadBench, and GPQA-Diamond, and quantify transfer using DE (§2). Full configurations are in Appendix B.

Observation 3.1: Increasing teacher scale does not reliably improve transfer. Using the largest teacher (Qwen3-32B) yields neither the highest accuracy nor the highest DE for any student in our scaling grid. For Qwen3-4B-Base, switching the teacher from 8B to 32B lowers student accuracy from 32.70% to 29.25% and DE from 65.1% to 43.3% (Figure 2). The preferred teacher also varies across students: Qwen3-8B performs best for both the 4B and 1.7B students, while Qwen3-14B leads for the 0.6B student (Appendix B). With teacher size held fixed, DE increases with student size for every teacher (Figure 2c). The complete grid and extended learning curves are in Appendix B.

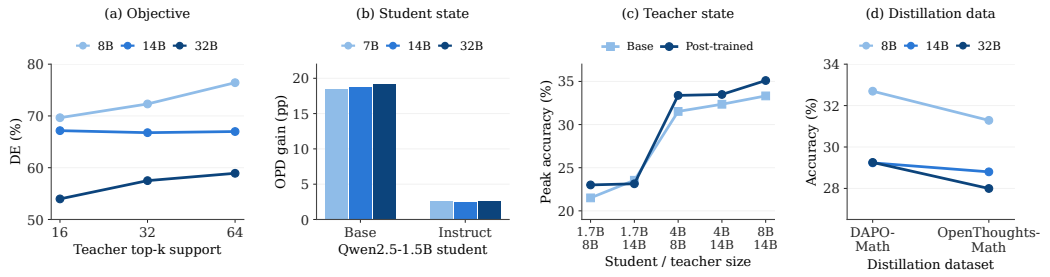


Figure 3: **OPD effectiveness varies across objectives, model states, and training data.** **(a)** Under forward KL, DE decreases with teacher size at every top-k support. **(b)** Qwen2.5-1.5B base students gain more from OPD than instruct students. **(c)** Post-trained teachers improve peak accuracy in four of five matched-size comparisons. **(d)** Qwen3-8B gives the highest accuracy on both training datasets for Qwen3-4B-Base.

Observation 3.2: OPD effectiveness depends on the distillation setting. The pattern from *Observation 3.1* holds across distillation objectives, student and teacher states, and training data (Figure 3). Under forward KL with top-k support, DE decreases with teacher size at every k (Figure 3a). OPD gains vary substantially with the student's starting state: Qwen2.5-1.5B base students gain 18.54–19.14 points, compared with 2.48–2.66 for instruct students (Figure 3b). Post-

trained teachers improve peak accuracy in four of five matched-size comparisons (Figure 3c). Qwen3-8B gives the highest accuracy on both DAPO-Math and OpenThoughts-Math for the 4B student (Figure 3d). Full results are in Appendix B.

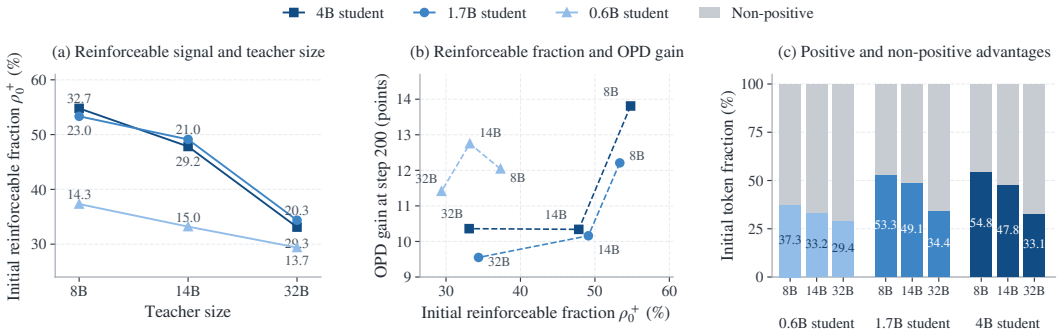


Figure 4: **Reinforceable signal varies across student–teacher pairings. (a)** Initial ρ_0^+ decreases with teacher size for every student, with point labels showing step-200 accuracy. **(b)** Selecting maximum ρ_0^+ identifies the largest step-200 gain for two of three students. **(c)** Larger teachers leave a greater share of tokens without positive distillation signal.

Observation 3.3: The reinforceable fraction complements scalar divergence. Measured before any OPD training, ρ_0^+ decreases with teacher size for every student (Figure 4a, Figure 11): larger teachers leave more tokens without positive distillation signal (Figure 4c), consistent with their lower OPD gains. Reverse KL follows a different pattern: for both 0.6B and 1.7B students, the 32B teacher has the smallest initial rKL but gives the lowest step-200 accuracy. Maximum ρ_0^+ identifies the best teacher for two of three students at step 200 (Figure 4b), distinguishing productive pairings that scalar divergence does not.

Together, these dynamics suggest that effective OPD depends not on teacher capability or scalar policy distance alone, but on the learning signal induced by the current student–teacher pairing. We next ask whether the stages surrounding OPD can deliberately reshape this pairing (§4).

4 Composing Post-Training Stages

§3 showed that greater teacher scale does not reliably improve OPD and that transfer depends on the current student–teacher pairing. We now ask how the stages surrounding OPD affect subsequent learning. SFT, RLVR, and OPD each reshape the policy that the next stage receives, motivating a closer examination of their roles in multistage post-training (Yang et al., 2025; Xiao et al., 2026; Zeng et al., 2026; Yang et al., 2026). We study this by changing one component at a time while holding everything else fixed:

- **Student warm-up** (§4.1): How does preparing the student affect OPD, and do the method and duration matter?
- **Teacher adaptation** (§4.2): Does improving the teacher help, and is it about domain adaptation or more?
- **Stage ordering** (§4.3): Can OPD be a better initialization for subsequent RLVR than SFT?

Setup. Unless stated otherwise, all experiments in this section use Qwen3-4B-Base as the student and Qwen3-14B as the teacher with the same optimization recipe and evaluation protocol as in §3. Preparatory stages (student warm-up, teacher adaptation, and pre-RLVR SFT or OPD) trained on OpenThoughts-Math (Guha et al., 2026), while the final OPD and RLVR stages trained on DAPO-Math (Yu et al., 2025). Further details are reported in Appendices C–E.

4.1 Student Warm-Up Before Distillation

If OPD effectiveness depends on the student–teacher pairing at initialization as observed in §3, a natural way to improve distillation is to prepare the student before it begins. We compare two warm-up methods: (i) SFT on correct teacher-generated trajectories, which moves the student toward the teacher, and (ii) RLVR on the target domain, which improves the student independently of the

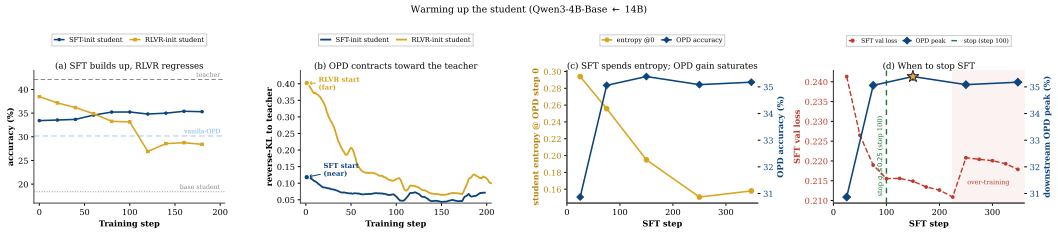


Figure 5: **Student warm-up: the method and the amount both matter. (a)** Brief SFT raises the observed OPD performance and reduces the subsequent drop, whereas the RLVR-prepared student regresses under OPD from the same teacher. **(b)** Both students contract toward the teacher in reverse KL—the RLVR-warmed student from much farther away—so the regression occurs despite successful teacher matching. **(c)** As SFT continues, student entropy keeps falling while downstream OPD accuracy saturates after a short warm-up. **(d)** The SFT validation-loss plateau signals when to stop: SFT beyond it no longer improves downstream OPD.

teacher. To understand how the amount of SFT shapes the student and the distillation that follows, we sweep the SFT duration across $K \in \{25, 75, 150, 250, 348\}$ steps and track student entropy at OPD initialization.

Observation 4.1: Higher student capability does not imply better distillability. SFT and RLVR both improve the student before OPD, but have sharply different effects on subsequent distillation (Figure 5a). A short SFT warm-up raises the OPD performance from 33.6% to 35.4% and removes the regression that follows the peak in the curve. The RLVR-warmed student instead enters OPD at 38.5%, already within 3.6 points of the teacher, yet falls to 28.4% by the end of distillation. A stronger student is therefore not necessarily a better starting point for distillation.

Observation 4.2: Closer teacher matching can accompany student regression. Reverse KL decreases during OPD from both SFT and RLVR initializations, even though their accuracy trajectories differ (Figure 5b). The SFT checkpoints begin downstream OPD with reported rKL values around 0.10–0.14 (Appendix C.5), while the illustrated RLVR initialization begins near 0.40. For the latter, continued teacher matching does not prevent accuracy from falling. This extends the distinction in §3: progress in the distillation objective need not imply progress on the task.

Observation 4.3: SFT warm-up benefits saturate early. Increasing the SFT warm-up initially improves downstream OPD, from 30.9% after 25 steps to 35.1–35.4% after 75–150 steps, and then saturates (Figure 5c). Beyond this region, additional SFT provides little improvement in subsequent distillation. Meanwhile, the reported initial entropy falls from 0.2559 at $K = 100$ to 0.1507 at $K = 250$. The first SFT validation-loss plateau, near step 100, lies close to the region where downstream gains saturate (Figure 5d). This retrospective agreement suggests a useful signal of diminishing returns, making this plateau a practical indicator of diminishing warm-up returns.

Discussion. These experiments separate student *capability* from student *distillability*. RLVR can produce a substantially stronger student while simultaneously making supervision from the same teacher less useful, whereas a brief SFT warm-up can improve the student–teacher interface even with a much smaller gain in standalone capability. The contrast is also diagnostic: reverse KL alone does not distinguish the vanilla and RLVR-adapted student states, while the reinforceable fraction changes sharply with their downstream OPD behavior. Taken together, the results suggest that student preparation should be evaluated by how it reshapes subsequent learning, not only by how much it improves the student before distillation.

Guidance. For the studied pairing, a brief SFT warm-up is a useful preparation for OPD. The validation-loss plateau provides a candidate point to check whether additional SFT still improves downstream learning.

4.2 Teacher Adaptation Before Distillation

Student warm-up changes the policy that *receives* the distillation signal and teacher adaptation changes the policy that *produces* it. We ask: does improving the teacher make it more effective for OPD, and does *how* the teacher is improved matter? We compare two adaptation methods on the

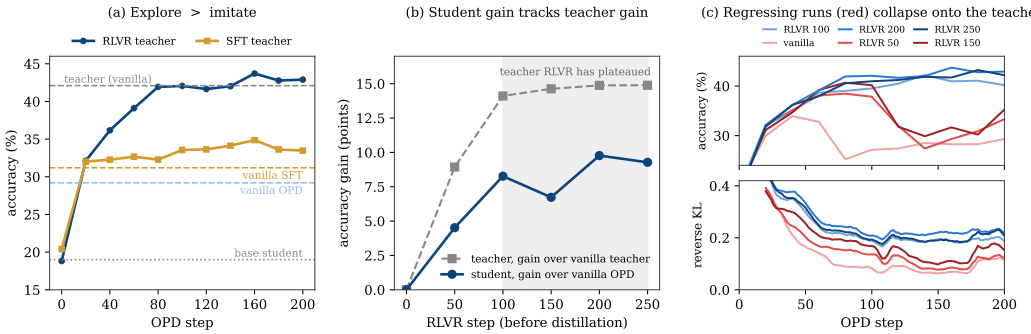


Figure 6: **Teacher adaptation transfers capability, not proximity.** **(a)** OPD from an RLVR-adapted teacher surpasses the off-the-shelf teacher's own accuracy, whereas a SFT-trained teacher improves slightly over vanilla OPD. **(b)** The student's gain tracks the teacher's gain across RLVR checkpoints and flattens once the teacher plateaus (shaded). **(c)** Some runs regress while reverse KL continues to decrease, so teacher matching alone does not explain accuracy retention.

same target-domain prompts: *RLVR*, which improves the teacher on-policy with verifiable outcome rewards, and *STaR-style SFT* (Zelikman et al., 2022), which trains the teacher to imitate its own verified solutions. [1] Both methods use the same target-domain prompts, and we take checkpoints at $K \in \{50, 100, 150, 200\}$ for both, plus $K = 250$ for RLVR. Each resulting teacher distills the same Qwen3-4B-Base student with identical OPD data, initialization, and budget (Appendix D).

Observation 4.4: Teacher capability gains transfer through OPD; a control that does not improve the teacher yields no gain. Distilling from an RLVR-adapted teacher raises student accuracy from 18.9% to 38.5–43.7% depending on the checkpoint, against 33.6% from the off-the-shelf teacher (Figure 6a). The STaR-adapted teacher does not meaningfully improve (33.4–34.9% across checkpoints), and downstream student accuracy is correspondingly unchanged. This confirms that only genuine teacher capability gains transfer.

Observation 4.5: The student's gain is a stable fraction of the teacher's gain, so distillation saturates when teacher RLVR does. Teacher accuracy rises from 42.1% before adaptation to 51.0% at $K = 50$ and 56.2% at $K = 100$, then remains within 56.7–57.0%. The corresponding student OPD peaks rise from approximately 33.6% to 38.5% and 42.2%, with later checkpoints yielding 40.7–43.7% (Figure 6b). Evaluating DE at these observed peaks, students recover approximately 60% of the gap between the teacher and the initial student. This complements §3: while scaling teacher size from 14B to 32B did not help, raising the 14B teacher's accuracy from 42% to 57% through RLVR substantially improves transfer.

Discussion. RLVR and STaR-style self-training see the same prompts but change the teacher differently. RLVR lets the teacher explore and reinforces what succeeds, raising its accuracy by 9–15 points; self-training imitates solutions the teacher can already produce and leaves its accuracy unchanged. Only the first gives the student more to learn, in line with evidence that RL generalizes where SFT mainly memorizes (Chu et al., 2025) and that OPD transfers what is new to the student (Li et al., 2026). The benefit lasts as long as the teacher keeps improving: student gains track teacher gains and flatten once teacher accuracy plateaus. Our control cannot tell whether RLVR-induced capability would transfer better at matched teacher accuracy. A stronger teacher also does not prevent regression, since some students lose accuracy while reverse KL keeps falling (Figure 6c). As a complementary check at fixed teacher size, the checkpoint with the largest ρ_0^+ yields the best Qwen3-1.7B-Base student (29.77% versus 28.00% for minimum rKL), in a single-run comparison where it is also the latest checkpoint (Appendix D.5).

Guidance. Teacher accuracy is useful to monitor during adaptation, with diminishing downstream returns once it plateaus in this sweep. Teacher checkpoint selection should also consider the subsequent OPD trajectory, since a high peak need not persist to the end of training.

[1] We also call this SFT teacher, which fine-tuned on one verified-correct self-generated rollout per prompt whenever at least one of four sampled responses is correct.

4.3 Stage Ordering for RLVR Initialization

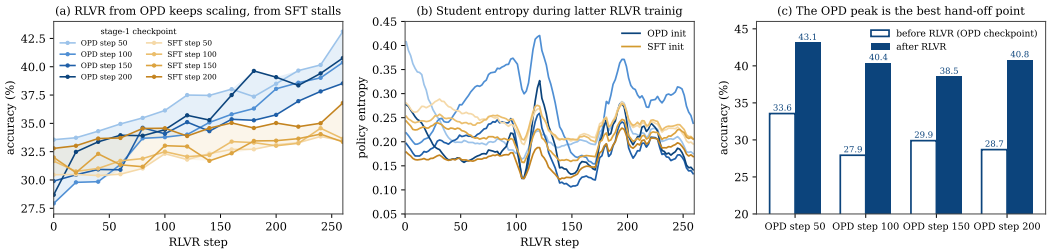


Figure 7: **An OPD checkpoint is a better RLVR initialization than an SFT checkpoint trained on the same trajectories. (a)** RLVR from OPD checkpoints (blue) is still rising at step 260, while RLVR from SFT checkpoints (gold) stalls by step 100. Shading marks the range across checkpoints. **(b)** OPD-initialized runs spend entropy as accuracy rises; SFT-initialized runs keep theirs largely unchanged. **(c)** Among the tested OPD handoffs, the step-50 checkpoint has both the highest pre-RLVR accuracy and the highest post-RLVR accuracy, reaching 43.1%.

We now ask the reverse question: which student state is the better initialization for a subsequent RLVR stage? To understand that, we compare two separate preparation pipelines, vanilla OPD followed by RLVR and SFT followed by RLVR. SFT uses verified teacher-generated solutions, whereas OPD uses teacher feedback on student-generated responses. Both start from Qwen3-4B-Base and use the same fixed Qwen3-14B teacher, followed by the same GRPO procedure on DAPO-Math-17K. Preparation prompts come from the same retained OpenThoughts-Math pool, with the SFT validation split and other preparation differences detailed in Appendix E.1. The comparison matters because RLVR is on-policy, so the initialization determines which behaviors are available for exploration and reinforcement, and because current recipes commonly place SFT before RLVR (Yang et al., 2025; Xiao et al., 2026). We repeat it for Qwen3-1.7B-Base and Qwen2.5-1.5B-Base with their corresponding 14B teachers (Appendix E.5).

Observation 4.6: OPD initialization enables substantially stronger scaling with RLVR. The two initialization families start RLVR at similar accuracy, with SFT slightly ahead on average. Their trajectories then diverge sharply. After 260 RLVR steps, every OPD-initialized run (38.5–43.1%) outperforms every SFT-initialized run (33.4–36.8%), and the separation grows over the observed horizon: OPD-initialized models are still improving at the end of a full pass over DAPO-Math-17K, while SFT-initialized models gain much less (Figure 7a). OPD therefore leaves not merely a stronger checkpoint but *substantially more headroom for scaling the downstream RLVR stage.*

Observation 4.7: OPD and SFT checkpoints with comparable entropy yield different RLVR gains. A natural explanation is that OPD preserves more policy entropy before RLVR (Aphale & Liu, 2026), but SFT and OPD checkpoints have substantially overlapping initial entropy yet produce clearly separated RLVR trajectories (Figure 7b). Some SFT checkpoints start with higher entropy than OPD checkpoints but still finish lower. This agrees with Shen (2026), where OPD's higher entropy at RL initialization did not predict a better endpoint, and locates the advantage in something more structured than aggregate uncertainty.

Observation 4.8: The highest-accuracy tested OPD checkpoint is also the best RLVR handoff. Section 3 showed that OPD reaches its best task performance well before teacher matching converges. The same checkpoint also provides the strongest initialization for downstream RLVR. RLVR from the peak OPD checkpoint (step 50) reaches 43.11%, while later OPD checkpoints that have already regressed finish at 38.5–40.8% (Figure 7c). RLVR recovers part of the lost performance but does not close the gap. Continuing OPD beyond its accuracy peak therefore degrades not only the current model but also its potential for subsequent RLVR.

Discussion. When RLVR follows teacher supervision, our results favor OPD over SFT at comparable student accuracy. One plausible explanation is that OPD provides teacher feedback on the student's own trajectories, aligning the training distribution with the one that RLVR must subsequently explore—but the advantage cannot be reduced to entropy alone. Testing whether this advantage grows with model scale, data scale, and substantially larger RLVR compute is a natural next step for frontier post-training systems.

Guidance. When RLVR follows teacher supervision, our results favor OPD over SFT at comparable student accuracy. OPD should hand off near its task-performance peak rather than continue toward teacher convergence, and scalar entropy alone is not sufficient to choose this transition. Within the training horizon we study, OPD-initialized models also continue to benefit more from additional RLVR training. They also reinforce the central composition principle: evaluate a stage by both the capability it adds and the learning conditions it leaves for the next stage.

5 A Simple Composition Recipe

§4 studied student preparation and teacher adaptation separately. We now combine them: adapt Qwen3-14B with RLVR until convergence of teacher ($K = 200$), briefly warm Qwen3-4B-Base with SFT on the adapted teacher's trajectories ($K = 75$), and then distill from that teacher. At the common 200-step OPD endpoint, vanilla OPD reaches 29.2% average accuracy, student warm-up reaches 35.3%, teacher adaptation reaches 42.2%, and both combined reach 43.8% (Figure 8). The combined pipeline improves over vanilla OPD by 14.6 percentage points, or 50% relative, at the same OPD step budget with additional preparatory training. Teacher adaptation provides most of the improvement, while adding student warm-up yields a further 1.6 points. The two interventions therefore combine effectively, although their individual gains are not additive.

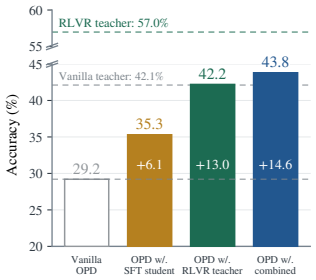


Figure 8: Final OPD accuracy combined with teacher RLVR adaptation and SFT warm-up.

Together with §4.3, these results suggest a simple composition principle: **adapt the teacher with RLVR, briefly warm the student with SFT, distill with OPD, and finally distill to the cold-start student.** More broadly, post-training stages are most effective when each prepares a productive learning interface for the one that follows. Our aim in this work is not a new algorithm but a first controlled map of how these stages interact. Whether the rule holds across model scales, domains, and longer training horizons is beyond our scope, and we hope it invites that study. For interested readers, Appendix F examines how teacher ordering shapes continued improvement when post-training stages are composed recursively.

6 Conclusion

In this work, we presented a controlled investigation into how SFT, RLVR, and OPD interact in LLM post-training. By changing one stage at a time under a shared protocol, we clarified why the same stage can help or hurt depending on the state it inherits. Brief SFT improves distillability whereas an RLVR-strengthened student can regress, and teacher RLVR improves transfer until teacher accuracy plateaus. Combining teacher adaptation with student warm-up raises final OPD accuracy from 29.2% to 43.8% at the same OPD step budget. Notably, we also show that OPD supports stronger downstream RLVR scaling than SFT within the studied budgets, despite similar starting accuracy and overlapping aggregate entropy in the main Qwen3-4B-Base comparison. Together, these results show that effective post-training composition depends on both the capability each stage adds and the *learning interface* it leaves for the next stage.

Limitations and Future Work. Our experiments show that each stage shapes both current capability and subsequent learning, with composition studied mainly using Qwen3-4B and Qwen3-14B. Teacher RLVR improves downstream OPD under our protocol, while transferability at matched teacher accuracy and the underlying mechanisms remain open. These leave our central finding intact: across controlled interventions, a post-training stage's effect depends on the learning state it inherits. Future work can extend these findings to larger model scales with mixture-of-expert (MoE) architectures and also agentic learning, where tool use and long-horizon interaction introduce new demands on exploration and transfer. Our recursive OPD case study (Appendix F) motivates studying adaptive stage transitions and progressively refreshed teachers as ways to sustain learning as students and tasks evolve. The broader goal is to develop post-training pipelines that improve current performance while preserving the capacity for continued learning. We hope this study provides an empirical foundation for post-training pipelines that improve current performance while preparing models for continued learning.

AI USE STATEMENT

In this work, we used generative AI tools to edit and tighten the manuscript text, to check the consistency of numbers between the text, tables, and figures, and to help write and debug plotting scripts. We did not use generative AI tools to develop theoretical frameworks, formulate or prove mathematical claims, propose hypotheses, design experiments, implement the training methods, generate or clean data, or produce or interpret the reported results. All AI-assisted text and code were reviewed and verified by the authors. We take responsibility for the final content of this work, including all text, mathematical statements, experimental claims, and figures.

ETHICS STATEMENT

This work studies post-training using open-weight language models, existing reasoning datasets, and model-generated trajectories. Our goal is to make interactions between training stages more transparent and accessible to the research community. Distillation may also transfer errors or biases from teachers, so improved benchmark accuracy should not be interpreted as evidence of safe deployment. Applications beyond the studied settings require appropriate safety and reliability evaluation.

REPRODUCIBILITY STATEMENT

We document the experimental procedures and supporting analyses in Appendices A–G Appendix A defines the objectives and diagnostics, while Appendix B provides model and dataset details, training and evaluation settings, hardware configurations, and metric aggregation. Appendices C–E describe student warm-up, teacher adaptation, and downstream RLVR initialization, respectively, including data preparation, hyperparameters, and checkpoint selection. Appendix F details the recursive case study, and Appendix G discusses limitations and unsuccessful attempts. The main Qwen3-4B-Base scaling experiments use three rollout seeds (42, 43, and 44) per teacher, with results reported as means and sample standard deviations. The smaller-student scaling experiments use one run per teacher, and checkpoint sweeps are distinguished from seed replication. To support independent reproduction and further research, we will release training and evaluation code, data-preparation scripts, experiment configurations, available training logs, evaluation outputs, and scripts for reproducing the tables and figures.

BIBLIOGRAPHY

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*, volume 2024, pp. 21246–21263, 2024.

Siddharth Aphale and Kelly Liu. Sft overtraining predicts rank inversion via entropy collapse under rlvr. *arXiv preprint arXiv:2606.18487*, 2026.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6621–6642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24j.html.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801. IEEE, 2019.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10818–10838. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chu25c.html.

Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-strong generalization via direct on-policy distillation. *arXiv preprint arXiv:2607.05394*, 2026.

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*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

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

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tommy He, Jerome Sieber, and Matteo Saponati. A predictive law for on-policy self-distillation from world feedback. *arXiv preprint arXiv:2605.30070*, 2026.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in neural information processing systems*, 35:33716–33727, 2022.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH.

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

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25366–25394, 2025.

Runze Liu, Jiashun Liu, Xu Wan, Yuqian Fu, and Ling Pan. When rl fails after sft: Rejuvenating model plasticity for robust sft-to-rl handoff. *arXiv preprint arXiv:2606.09932*, 2026.

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.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. Kl for a kl: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan

Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

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

Jianxiong Shen. Stage-1 controls the entropy regime, not the outcome. *arXiv preprint arXiv:2606.09059*, 2026.

Zhennan Shen, Yanshu Li, Qingyu Yin, Chak Tou Leong, Zhilin Wang, Yanxu Chen, Rongduo Han, Sunbowen Lee, and Yi R Fung. On the geometry of on-policy distillation. *arXiv preprint arXiv:2606.07082*, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations. *arXiv preprint arXiv:2607.13399*, 2026.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

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

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

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

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, juncai liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru 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, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Weixiao Zhan, Yongcheng Jing, Leszek Rutkowski, and Dacheng Tao. Distillation traps and guards: A calibration knob for llm distillability. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19814–19833, 2026.

Hanyang Zhao, Haoxian Chen, Han Lin, Genta Indra Winata, David Yao, and Wenpin Tang. Opd+: Rethinking the advantage design for on-policy distillation. *arXiv preprint arXiv:2606.01039*, 2026.

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.

APPENDIX

A Background and Related Work **16**

A.1 Post-Training Objectives 16

A.2 Measuring Transfer 17

A.3 What Is Known, and What Is Not 17

B Experimental Details and Additional Analyses on OPD Dynamics **18**

B.1 Experimental Setup and Training Configurations 18

B.2 Metric Definitions and Aggregation 20

B.3 Full Student–Teacher Scaling Results 20

B.4 Robustness Across Distillation Objectives 21

B.5 Model States, Training Data, and Evaluation Domains 22

B.6 Reinforceable Fraction and Teacher Selection 23

B.7 Teacher-Dependent Training Dynamics 24

B.8 OPD can lose task performance despite continued teacher matching. 25

B.9 Early Stopping and Training-Budget Accounting 25

C Student Warm-Up Before Distillation **26**

C.1 SFT Warm-Up 26

C.2 RLVR Warm-Up 28

C.3 Downstream OPD Setup 28

C.4 Sensitivity to SFT Warm-Up Duration 30

C.5 Initialization Diagnostics 30

D Teacher Adaptation Before Distillation **30**

D.1 Experimental Setup 31

D.2 Teacher Adaptation with SFT 31

D.3 RLVR Teacher Adaptation 31

D.4 Downstream OPD Configuration and Evaluation 32

D.5 Teacher Selection at Fixed Model Size 32

D.6 Additional Results 33

E Downstream RLVR Initialization **33**

E.1 Experimental Setting 34

E.2 SFT Initialization 34

E.3 OPD Initialization 34

E.4 Common Downstream RLVR Stage 34

E.5 Generalization Across Student Scale, Model Family, and Handoff Checkpoint . . . 36

F Case Study: Can Post-Training Composition Be Recursive? 37

G Further Discussions 38

G.1 Why Composition Matters 38

G.2 Limitations 38

G.3 Unsuccessful Attempts 39

A Background and Related Work

We first define the three post-training objectives studied throughout the paper (§A.1) and the quantities we use to characterize student–teacher transfer (§A.2), placing each definition next to the prior work it draws on. We also summarize what is established about OPD and identify the gap this paper studies (§A.3).

A.1 Post-Training Objectives

Supervised fine-tuning (SFT). Given a fixed dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}$ of trajectories to imitate, SFT minimizes the token-level cross entropy

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{t=1}^{|y|} \log \pi_{S_\theta}(y_t \mid h_t)\right], \tag{1}$$

where $h_t = (x, y_{<t})$. The supervision is off-policy: targets are fixed and do not depend on what the student produces, so the student is trained to reproduce trajectories it may never generate on its own.

Reinforcement learning with verifiable rewards (RLVR). RLVR optimizes the student on its own rollouts using a scalar outcome reward $r(x, y) \in \{0, 1\}$ obtained by verifying the final answer. We use GRPO (Shao et al., 2024), which samples a group of G responses per prompt and normalizes within the group,

$$\hat{A}_i^{\text{RLVR}} = \frac{r(x,y^{(i)}) - \mu_{\text{group}}}{\sigma_{\text{group}}}, \tag{2}$$

assigning the same advantage to every token of response i. The supervision is on-policy but sparse: one scalar per trajectory, indicating whether the answer was correct but not which tokens were responsible.

On-policy distillation (OPD). OPD also trains on student-generated trajectories, but replaces the sparse outcome signal with teacher supervision at every generated prefix (Gu et al., 2023; Agarwal et al., 2024; Lu & Lab, 2025). Writing $p_t(v) = \pi_S(v \mid h_t)$ and $q_t(v) = \pi_T(v \mid h_t)$, the reverse-KL objective is

$$\mathcal{L}_{\text{OPD}} = \mathbb{E}_{x,\, y\sim\pi_S}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\text{KL}}(p_t \,\|\, q_t)\right]. \tag{3}$$

Following standard practice (Lu & Lab, 2025), our default implementation uses a sampled-token policy-gradient estimator, which requires only one teacher log-probability per sampled token rather than the full teacher distribution. For the token y_t sampled by the student we define the teacher advantage

$$A_t = \text{sg}[\log \pi_T(y_t \mid h_t) - \log \pi_{S_{\text{old}}}(y_t \mid h_t)], \tag{4}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation, and optimize it with the same clipped importance-ratio objective used for RLVR. The sign of A_t determines the effect of each update. When $A_t > 0$ the update reinforces a student-sampled token that the teacher prefers more strongly. When $A_t < 0$ it suppresses that token without specifying which alternative should replace it. The mean of A_t under the student's own samples equals the negative reverse KL at that prefix (Appendix B, Proposition 1), so the mean carries no information beyond the loss itself. The mean therefore carries no information beyond reverse KL, motivating diagnostics that capture other properties of the distribution of A_t, including its sign structure.

Recent work has questioned the bias, variance, and robustness of this sampled-token estimator (Zhao et al., 2026; Oh et al., 2026; Xing et al., 2026; Fu et al., 2026). To test whether our observations depend on this formulation, we additionally evaluate a dense top-k forward-KL objective,

$$\mathcal{L}_{\text{FKL}}^{(k)} = \mathbb{E}_{x,\, y\sim\pi_S}\left[\frac{1}{|y|}\sum_{t=1}^{|y|}\sum_{v\in\mathcal{V}_k^T(h_t)} q_t(v)\log\frac{q_t(v)}{p_t(v)}\right], \quad \mathcal{V}_k^T(h_t) = \text{TopK}(q_t, k), \tag{5}$$

which matches the teacher's distribution over its k most likely tokens and therefore does specify replacements.

The three methods differ along two axes: SFT provides dense supervision on fixed trajectories, RLVR provides sparse outcome supervision on student trajectories, and OPD provides dense teacher supervision on student trajectories.

A.2 Measuring Transfer

We characterize student–teacher transfer from three views: the outcome of a completed distillation run, the structure of the teacher's signal on the student's behavior at initialization, and the on-policy distance between the two policies during training.

Distillation effectiveness. Let A_S and $A_{S'}$ denote the student's accuracy before and after distillation, and A_T the teacher's accuracy. We define

$$\mathrm{DE}(S \leftarrow T) = \frac{A_{S'} - A_S}{A_T - A_S}, \tag{6}$$

the fraction of the teacher–student gap recovered through distillation, with $1 - \mathrm{DE}$ the *unrecovered share*. The quantity is well defined when the teacher outperforms the student ($A_T > A_S$), and negative values indicate regression. The same quantity is used concurrently by Li et al. (2026) under the name *gap recovery rate*. We adopt it because raw accuracy gains are not comparable across students that begin at different accuracies.

Reinforceable fraction. Partitioning sampled tokens by the sign of A_t in Eq. 4, we define

$$\rho_0^+ = \mathbb{P}_{y \sim \pi_{S_0}} \left[A_t > 0 \right], \tag{7}$$

the fraction of tokens on the student's rollout that the teacher would reinforce rather than suppress, evaluated at initialization before any parameter update. The same sign decomposition is introduced concurrently by Wang et al. (2026), who use it to diagnose signal corruption and to motivate advantage clipping. Li et al. (2026) track a closely related quantity, the top-k overlap between student and teacher at student-visited states, and show that it separates successful from failing runs. Our use differs in purpose rather than in construction. Because ρ_0^+ depends only on the student's initial distribution, it requires a single scoring pass and no training, and therefore gives an initialization-time view of how much of the student's current on-policy behavior the teacher can directly reinforce. In §3 we test whether this quantity, measured before any training, identifies which teacher will distill most effectively. Pre-training predictors of distillation gain have also been proposed in the self-distillation setting (He et al., 2026).

On-policy teacher–student divergence. We also measure the reverse KL on student-generated trajectories after s optimizer steps,

$$\widehat{\mathrm{rKL}}_s = \mathbb{E}_{x,\, y \sim \pi_{S_s}} \left[\frac{1}{|y|} \sum_t \log \frac{\pi_{S_s}(y_t|h_t)}{\pi_T(y_t|h_t)} \right], \qquad \kappa_s = \widehat{\mathrm{rKL}}_s / \widehat{\mathrm{rKL}}_0. \tag{8}$$

The first measures teacher–student distance on the states the student actually visits, and provides the natural reference for asking whether distillation effectiveness can be explained by scalar policy divergence alone. The normalized form κ_s places pairs with different initial divergence on a common axis and is what we plot when studying training dynamics. Both are computed from the student's rollouts on training prompts, so they can be monitored during training without additional evaluation passes.

A.3 What Is Known, and What Is Not

OPD Training Dynamics Landscape Increasing teacher capability alone does not guarantee more effective distillation. Classical knowledge-distillation work observed that higher teacher accuracy need not produce a better student (Cho & Hariharan, 2019), motivating intermediate teacher assistants (Mirzadeh et al., 2020) and revised objectives for strong teachers (Huang et al., 2022). The effect reappears for reasoning models, where traces from strong teachers can underperform simpler alternatives for small students (Li et al., 2025). In the on-policy regime, three concurrent studies reach converging conclusions. Li et al. (2026) identify thinking-pattern consistency and genuine novelty of teacher knowledge as two important conditions for successful transfer, and show that successful runs are marked by rising top-k overlap at student-visited states. Wang et al. (2026) characterize OPD as an exploration catalyst rather than a capability expander and attribute failures to signal corruption under student–teacher mismatch and to length exploitation. Zhu et al. (2026) document pitfalls of the top-K reverse-KL approximation and of locally incompatible supervision on drifted prefixes. Section 3 re-examines these findings under a single controlled protocol across a 3×3 grid of student and teacher sizes, and tests robustness to model family, distillation objective, and training data.

Interventions on Individual Training Stages Prior work has also examined each stage adjacent to OPD in isolation. On the student side, an off-policy cold start (SFT on teacher-generated rollouts before distillation) raises initial overlap and improves both early optimization and the final ceiling (Li et al., 2026; Zhu et al., 2026). On the teacher side, RLVR adaptation of the teacher improves downstream distillation and moves the teacher closer to the student's distribution (Zhu et al., 2026), rejection fine-tuning has been used to calibrate teachers for distillability (Zhan et al., 2026), and the teacher's RL-induced policy shift can itself be transferred in place of the teacher policy (Feng et al., 2026). On the stage that follows, pre-RL policy entropy has been shown to predict post-RLVR outcomes better than pre-RL accuracy (Aphale & Liu, 2026), over-trained SFT checkpoints can lose plasticity needed for downstream RL (Liu et al., 2026), and SFT, OPD, and RLVR leave structurally different updates in parameter space (Shen et al., 2026). A concurrent study of two-stage post-training for vision-language models compares SFT and OPD warm-starts directly and finds that OPD's higher entropy at RL initialization does not translate into a better endpoint in their setting (Shen, 2026). We return to this contrast in §4.

The Gap We Think Prior work has established several effects of individual stages and adjacent stage transitions. What remains much less understood is how these stages *compose*, that is, how changing one stage alters the learning conditions faced by the next. We first characterize the OPD dynamics that determine whether a student–teacher pairing is productive (§3), then study how SFT, RLVR, and OPD reshape that pairing when composed (§4), and finally ask whether it can be maintained recursively as the student evolves (§F).

B Experimental Details and Additional Analyses on OPD Dynamics

This appendix details §3 with complete configurations, learning curves, and additional comparisons. The analyses examine how teacher scale interacts with the student state, distillation objective, and training duration. For interested readers, we also provide more context on the reinforceable fraction and the conditions under which an early handoff can preserve task performance.

B.1 Experimental Setup and Training Configurations

Our main training grid pairs Qwen3-0.6B, 1.7B, 4B-Base students with Qwen3-8B, 14B, 32B post-trained teachers, systematically varying both model sizes to characterize how OPD dynamics depend on student and teacher scale. All nine training runs use DAPO-Math-17k-Processed[2] with the sampled-token reverse-KL objective during distillation (Lu & Lab, 2025), built on verl as our base repository (Sheng et al., 2024). Table 1 lists the shared training configuration used across all scaling experiments. For our main experiments with the largest student, Qwen3-4B-Base, we report three runs with seeds 42, 43, and 44 (as shown in Figure 2) to ensure experimental reliability. Qwen3-0.6B-Base and Qwen3-1.7B-Base comparisons use a single run per teacher due to computational cost. Since the 4B experiments already establish that our main claim is robust to seed variation, these smaller scales serve primarily to test whether the observation generalizes across student sizes.

The main manuscript uses step 200 as the primary comparison point to reflect our computational budget. However, during our internal experiments, some runs continued beyond this point, and we prefer to include their longer trajectories here to provide a fuller picture of the training dynamics for transparency, so that interested readers can build on them in future work. That said, our core runs follow a one-epoch schedule with an optimizer horizon of 279 steps. We report results at step 260, their last shared evaluation, alongside step 200 for comparison with the main text. The forward-KL experiments are compared at step 180, which is available for every teacher and every value of k; where as model-state and training-data comparisons use step 200. Each table specifies its evaluation horizon, and we report variability across runs only where multiple seeds are available.

[2] https://huggingface.co/datasets/open-r1/DAPO-Math-17k-Processed

Table 1: Shared OPD configuration for the main experiment grid in §3.

Setting	Value
Training data	DAPO-Math-17k
Parameter update	Full-parameter training, LoRA rank 0
Optimizer	AdamW
Learning rate / (β_1, β_2)	10^{-6} / $(0.9, 0.999)$
Schedule / warm-up	Constant learning rate / 10 steps
Weight decay / gradient norm clipping	0.01 / 1.0
Prompt batch / PPO minibatch	64 / 64
PPO epochs / clipping ratio	1 / 0.2
Rollout responses / temperature / top-p	4 / 1.0 / 1.0
Maximum prompt / response length	2048 / 4096 tokens
Loss mode / policy-gradient update	Sampled-token reverse KL (`k1`) / enabled
Distillation coefficient / task reward	1.0 / disabled
Auxiliary KL reward / KL loss	Disabled / disabled
Loss aggregation / entropy coefficient	Token mean / 0
Configured loss ceiling / log-probability floor	10 / -10
Data shuffling / overlong-prompt filtering	Disabled / enabled
Qwen3 thinking mode	Disabled
Actor backend / rollout backend	FSDP / vLLM
Actor / teacher GPU allocation	4 / 4 GPUs
Rollout / teacher tensor parallelism	1 / 1
Sequence parallelism / compute dtype	1 / bfloat16
Dynamic batching / actor token budget	Enabled / 24,576 tokens per GPU
Teacher inference token budget	4096 tokens per batch
Training schedule	One epoch with no explicit trainer step cap
Evaluation interval	Every 20 steps

Table 2: Forward-KL configuration with optimization, sampling, and length settings match Table 1. The objective uses teacher top-k support.

Setting	Value
Student / teachers	Qwen3-4B-Base / Qwen3-{8B, 14B, 32B}
Training data	DAPO-Math-17k
Loss mode / top-k	`forward_kl_topk` / $k \in \{16, 32, 64\}$
Update	Direct distillation loss with policy-gradient mode disabled
Optimizer / learning rate	AdamW / 10^{-6}
Schedule / warm-up	Constant learning rate / 10 steps
Prompt batch / responses per prompt	64 / 4
Maximum prompt / response length	2,048 / 4096 tokens

Table 3: Evaluation protocol where the primary aggregate assigns equal weight to each of the seven benchmarks.

Setting	Value
Primary benchmarks	AIME 2024, AIME 2025, AMC 2023, MATH-500, Minerva Math, OlympiadBench, GPQA-Diamond
Responses per question	4
Temperature / top-p / top-k	0.6 / 0.95 / disabled
Prompt / response limit	2,048 / 32,768 tokens
Per-benchmark score	Mean correctness over four responses (`mean@4`)
Configured evaluator	`ttrl_math.reward_func` in `verl`
Primary aggregate	Unweighted mean of the seven benchmark scores
Initial accuracy	Evaluation at training step 0
Initial training diagnostics	First logged training step, recorded as step 1
Across-run summary	Mean and sample standard deviation where $n = 3$

Table 4: Training and evaluation datasets, prompt lengths use the Qwen3 tokenizer with the training chat template.

Dataset	Unique prompts	Average prompt tokens
Training		
DAPO-Math-17k	17,917	122
OpenThoughts3-Math	30,000	113
Evaluation		
AIME 2024	30	150
AIME 2025	30	145
AMC 2023	83	119
MATH-500	500	98
Minerva Math	272	158
OlympiadBench	675	127
GPQA-Diamond	198	240

B.2 Metric Definitions and Aggregation

For benchmark b with N_b questions, let c_{bqj} indicate whether sampled response j to question q is correct. The benchmark score and primary aggregate are:

$$A_b = \frac{100}{4N_b} \sum_{q=1}^{N_b} \sum_{j=1}^{4} c_{bqj}, \qquad A = \frac{1}{7} \sum_{b=1}^{7} A_b \tag{9}$$

Thus, `mean@4` measures average correctness across sampled responses. It does not apply majority voting or count a question as solved whenever one of its four responses is correct.

Let $A_{S \leftarrow T,h}$ denote the student's accuracy after h OPD steps. We compute gain and distillation efficiency using:

$$G_h = A_{S \leftarrow T,h} - A_{S,0}, \qquad \mathrm{DE}_h = \frac{A_{S \leftarrow T,h} - A_{S,0}}{A_T - A_{S,0}} \tag{10}$$

The initial student score is pooled across the runs in the corresponding main-grid row. These baselines are 2.25%, 10.79%, and 18.89% for the 0.6B, 1.7B, and 4B students. The fixed teacher reference accuracies used for DE are 40.09%, 42.11%, and 42.80% for the 8B, 14B, and 32B teachers. We report absolute accuracy alongside DE because a larger teacher reference increases the denominator of Equation (10). The two quantities therefore answer different questions about transfer.

For a student-generated token y_t at prefix h_t, define the log-probability advantage as

$$a_t = \log p_T(y_t \mid h_t) - \log p_S(y_t \mid h_t) \tag{11}$$

Within a valid response-token block $\mathcal{B}$, the reinforceable fraction is

$$\widehat{\rho}^{+}_{\mathcal{B}} = \frac{\sum_{t \in \mathcal{B}} m_t \mathbf{1}[a_t > 0]}{\sum_{t \in \mathcal{B}} m_t} \tag{12}$$

where m_t is the response mask. The reported statistic follows the trainer's mean aggregation of these masked token statistics, while zero advantages are excluded from both the positive and negative fractions. The term *reinforceable* refers to the sign of the distillation signal and it does not classify the correctness of a solution. On the other hand, rKL curves use the sampled-token statistic logged by the same training code.

All accuracy summaries first aggregate benchmarks within a run and then aggregate runs. Error bars represent sample standard deviation across rollout seeds. For peak summaries, we take the maximum within each run before averaging. These observed maxima provide retrospective reference values and they are distinct from selecting a checkpoint using an independent validation set.

B.3 Full Student–Teacher Scaling Results

Table 5 reports the complete grid at two training horizons. At step 260, the 8B teacher produces the highest accuracy for all three students. For Qwen3-4B-Base, its 31.49% accuracy exceeds the 14B and 32B outcomes by 2.30 and 2.92 points. The corresponding DE values show that increased teacher scale does not consistently translate into greater recovery of the teacher's capability.

Table 5: Complete Qwen3 base-student grid where accuracy is the seven-benchmark mean in percent. Gain and DE use the pooled initial score for each student.

Student	Teacher	A_{200}	A_{260}	Gain_{260}	DE_{260} (%)
0.6B	8B	14.29	14.91	+12.65	33.4
0.6B	14B	15.00	13.24	+10.99	27.6
0.6B	32B	13.66	13.24	+10.99	27.1
1.7B	8B	23.00	22.26	+11.47	39.1
1.7B	14B	20.95	20.13	+9.34	29.8
1.7B	32B	20.34	21.36	+10.58	33.0
4B	8B	32.70	31.49	+12.60	59.4
4B	14B	29.23	29.19	+10.30	44.4
4B	32B	29.25	28.57	+9.68	40.5

Figure 9 shows that this endpoint comparison coexists with different learning dynamics. The 0.6B student undergoes a pronounced intermediate decline with the 8B teacher before recovering. At step 200, the 14B teacher instead leads this row by 0.71 points. The 1.7B and 4B students favor the 8B teacher at both reported horizons. For 4B, a strong early improvement is followed by a decline and partial recovery. These results suggest that teacher selection should be interpreted relative to the intended training budget. They also show why a single endpoint does not fully describe transfer.

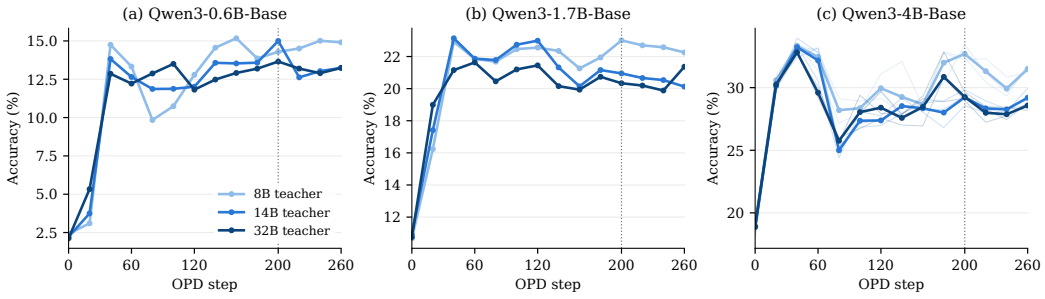


Figure 9: **Complete OPD learning curves across student and teacher sizes.** Each panel compares three teachers for one base student. The 0.6B and 1.7B curves show individual runs. For 4B, bold lines show the three-run mean and faint lines show individual runs. The dotted line marks step 200. All curves extend to the shared final evaluation at step 260.

B.4 Robustness Across Distillation Objectives

To assess whether our teacher-scaling observations extend across distillation objectives, we further examine Qwen3-4B-Base under forward KL with teacher top-k support for $k \in 16, 32, 64$. Table 6 compares all nine configurations at a common budget of 180 steps, showing that distillation from Qwen3-32B yields the lowest accuracy and DE at every value of k. Among smaller teachers, Qwen3-14B achieves higher accuracy at $k = 16$ and $k = 32$, while Qwen3-8B performs better at $k = 64$. These results extend our finding that larger teachers do not consistently improve transfer, while showing that the preferred teacher can depend on how much of its output distribution is matched.

Table 6: Forward-KL results. All cells use Qwen3-4B-Base and one run and peak accuracy is restricted to the same budget. Baseline and teacher references match Table 5.

k	Teacher	A (%)	Gain	DE (%)	Peak (%)
16	8B	33.66	+14.77	69.7	34.43
16	14B	34.48	+15.59	67.2	34.48
16	32B	31.79	+12.91	54.0	33.48
32	8B	34.22	+15.33	72.3	35.11
32	14B	34.39	+15.50	66.8	35.28
32	32B	32.64	+13.75	57.5	33.12
64	8B	35.09	+16.20	76.4	35.09
64	14B	34.44	+15.55	67.0	35.56
64	32B	32.98	+14.09	58.9	33.68

This comparison also highlights why accuracy and DE should be considered together. At $k = 16$, distillation from Qwen3-14B yields a student accuracy of 34.48% at step 180, compared with 33.66% when using Qwen3-8B. Despite this accuracy

advantage, Qwen3-14B has lower DE because its larger capability gap over the initial student increases the normalization denominator. A larger teacher can therefore produce a more accurate student while transferring a smaller fraction of its capability advantage, indicating that teacher selection depends on both the distillation objective and how transfer is measured.

B.5 Model States, Training Data, and Evaluation Domains

Student state and model family. To examine how OPD effectiveness varies across model families and student starting states, Table 7 extends our analysis to post-trained Qwen3 students and both base and instruct Qwen2.5 students at a common budget of 200 steps. For Qwen2.5-1.5B, OPD improves base-student accuracy by 18.54 to 19.14 points across three teachers, compared with gains of 2.48 to 2.66 points for its instruct counterpart. In contrast, post-trained Qwen3-4B starts at 39.58% accuracy and finishes below this baseline with every teacher. These results show that OPD gains depend substantially on the student's starting state, extending this observation beyond our main Qwen3 base-model grid.

Table 7: Student-state and model-family comparisons after 200 OPD steps. Entries report accuracy (%) with gains over the pooled initial student score in parentheses. The smallest teacher is 8B for Qwen3 and 7B for Qwen2.5. Results average three runs per teacher for Qwen3-4B-Base and use one run per configuration otherwise.

Student	State	Initial	Qwen2.5-7B/Qwen3-8B teacher	14B teacher	32B teacher
Qwen3-1.7B	Base	10.79	23.00 (+12.22)	20.95 (+10.16)	20.34 (+9.55)
Qwen3-1.7B	Post-trained	29.51	29.82 (+0.31)	31.17 (+1.65)	29.18 (-0.33)
Qwen3-4B	Base	18.89	32.70 (+13.81)	29.23 (+10.34)	29.25 (+10.36)
Qwen3-4B	Post-trained	39.58	38.03 (-1.55)	39.28 (-0.31)	37.69 (-1.89)
Qwen2.5-1.5B	Base	1.30	19.85 (+18.54)	20.01 (+18.71)	20.44 (+19.14)
Qwen2.5-1.5B	Instruct	18.08	20.69 (+2.61)	20.56 (+2.48)	20.73 (+2.66)
Qwen2.5-3B	Base	18.61	25.87 (+7.27)	26.10 (+7.50)	25.28 (+6.68)
Qwen2.5-3B	Instruct	25.11	24.40 (-0.70)	25.24 (+0.14)	25.56 (+0.45)

A more capable post-trained student has less measured headroom for improvement, which can change how much it benefits from teacher supervision and motivates our study of student preparation in §4. Teacher rankings also vary across model families, as illustrated by Qwen2.5-1.5B-Base, which achieves its highest step-200 accuracy with the 32B teacher. These results suggest that teacher selection should account for the student's model family and starting state rather than relying on a fixed ranking of teachers.

Teacher training state. Table 8 compares base and post-trained Qwen3 teachers at matched model sizes, with post-trained teachers yielding higher peak student accuracy in four of five pairs. For Qwen3-4B-Base, peak accuracy increases from 31.52% to 33.38% with an 8B teacher and from 32.35% to 33.49% with a 14B teacher. The exception is Qwen3-1.7B-Base paired with a 14B teacher, where the base teacher yields a higher peak. Endpoint rankings also depend on how much of this improvement persists through training, suggesting that teacher training state can affect both peak performance and its retention. This distinction motivates our controlled teacher-adaptation experiments in Section 4.

Table 8: Comparison of base and post-trained Qwen3 teachers at matched model sizes, using base students throughout. Endpoint accuracy is measured at step 200, and peaks are computed over evaluations through that step. Results with post-trained teachers average three runs for the 4B student, while all other configurations use one run.

Student	Teacher size	Base A	Post A	Base peak	Post peak
1.7B	8B	19.66	23.00	21.50	23.00
1.7B	14B	22.30	20.95	23.54	23.14
4B	8B	25.05	32.70	31.52	33.38
4B	14B	30.81	29.23	32.35	33.49
8B	14B	31.19	29.34	33.32	35.10

Training distribution. To examine whether our teacher-scaling observations depend on the distillation data, Table 9 compares Qwen3-4B-Base trained on DAPO-Math and OpenThoughts-Math at a common budget of 200 OPD steps. Qwen3-8B yields the highest student accuracy on both datasets, reaching 32.70% on DAPO-Math and 31.29% on OpenThoughts-Math. This agreement indicates that its advantage in this student setting extends beyond a single training distribution. At step 240 on OpenThoughts-Math, Qwen3-8B remains the best teacher, although Qwen3-32B overtakes Qwen3-14B. The comparison therefore supports a consistent preference for the 8B teacher across these two datasets, while showing that the relative performance of larger teachers can change with training duration.

Table 9: Qwen3-4B-Base accuracy (%) by distillation dataset. DAPO-Math reports mean ± s.d. over three runs. OpenThoughts-Math is run per teacher.

	DAPO-Math	OpenThoughts-Math	
Teacher	Step 200	Step 200	Step 240
8B	32.70 ± 0.66	31.29	31.10
14B	29.23 ± 0.63	28.80	29.18
32B	29.25 ± 0.34	28.00	30.35

Evaluation domain. To examine whether teacher rankings depend on the evaluation domain, Table 10 reports benchmark-level results for Qwen3-4B-Base at step 260. Qwen3-8B achieves the highest mean accuracy on all six math benchmarks, leading both the math-six aggregate at 33.22% and our primary core-seven aggregate at 31.49%. Qwen3-14B instead performs best on GPQA-Diamond and achieves the highest ten-benchmark average at 28.14% after including SciKnowEval. These results show that the preferred teacher depends on the target task mixture, even for a fixed student and training recipe. We retain core-seven as the primary metric and use the additional benchmarks to identify where teacher rankings differ across domains.

Table 10: Qwen3-4B-Base accuracy (%) at step 260, reported as mean ± sample standard deviation over three rollout seeds. Aggregates average the first six, first seven, or all ten benchmarks with equal weight.

Benchmark / aggregate	8B teacher	14B teacher	32B teacher
AIME 2024	13.89 ± 1.73	13.61 ± 2.68	13.33 ± 0.83
AIME 2025	12.78 ± 0.48	12.50 ± 0.83	11.11 ± 1.27
AMC 2023	42.97 ± 1.06	39.86 ± 0.92	36.45 ± 1.59
MATH-500	72.23 ± 0.70	61.57 ± 0.70	64.13 ± 1.25
Minerva Math	21.57 ± 0.70	18.29 ± 0.09	17.68 ± 0.41
OlympiadBench	35.89 ± 0.94	30.83 ± 0.97	31.00 ± 0.13
GPQA-Diamond	21.11 ± 0.51	27.71 ± 0.86	26.31 ± 1.68
SciKnowEval-Biology	14.67 ± 5.20	15.17 ± 1.04	15.33 ± 1.26
SciKnowEval-Chemistry	18.25 ± 2.00	23.75 ± 0.25	20.00 ± 1.09
SciKnowEval-Physics	16.04 ± 0.18	38.12 ± 1.13	32.40 ± 6.59
Math-six	33.22 ± 0.39	29.44 ± 0.87	28.95 ± 0.63
Core-seven	31.49 ± 0.29	29.19 ± 0.68	28.57 ± 0.34
All ten	26.94 ± 0.51	28.14 ± 0.44	26.77 ± 0.38

B.6 Reinforceable Fraction and Teacher Selection

Table 11: Initial reverse KL and ρ_0^+ for each student–teacher pair.

Student	0.6B			1.7B			4B		
Teacher	8B	14B	32B	8B	14B	32B	8B	14B	32B
Initial rKL	0.740	0.680	0.648	0.447	0.478	0.426	0.370	0.376	0.382
ρ_0^+ (%)	37.32	33.20	29.39	53.34	49.12	34.37	54.79	47.83	33.11

To relate teacher selection to the initial student–teacher pairing, Table 11 reports reinforceable fraction and reverse KL from the same runs used in our accuracy comparisons. The reinforceable fraction

decreases with teacher size for every student, falling from 54.79% to 33.11% for Qwen3-4B-Base as teacher size increases from 8B to 32B. Reverse KL follows a different pattern: for both 0.6B and 1.7B students, Qwen3-32B has the lowest initial rKL but does not yield the highest accuracy at either step 200 or step 260. A teacher's proximity to the student therefore does not necessarily identify the most effective source of supervision, motivating a direct comparison of selection rules.

We quantify each teacher choice through selection regret, defined as the difference between student accuracy obtained with the best observed teacher and that obtained with the selected teacher:

$$\mathcal{R}_h(S, \widehat{T}) = \max_{T \in \mathcal{T}} A_{S \leftarrow T,h} - A_{S \leftarrow \widehat{T},h}, \tag{13}$$

where $\mathcal{T}$ contains the three candidate teachers, h denotes the number of OPD steps, and $A_{S \leftarrow T,h}$ is the resulting student accuracy in percent. For the replicated 4B experiments, outcomes and initial diagnostics are averaged across runs before applying each selection rule.

Table 12 compares four rules at steps 200 and 260. Selecting the largest initial reinforceable fraction identifies the best-performing teacher for all three students at step 260, whereas selecting the smallest initial rKL incurs 1.67 and 0.89 points of regret for the 0.6B and 1.7B students. At step 200, the reinforceable-fraction rule incurs 0.71 points of regret for the 0.6B student and zero for the other two students, showing that its agreement with the best observed teacher also depends on the training horizon.

Table 12: Teacher-selection regret in accuracy points relative to the best observed teacher at each horizon. Lower is better. 4B comparisons use mean outcomes and diagnostics across three runs.

	Step 200			Step 260		
Selector	0.6B	1.7B	4B	0.6B	1.7B	4B
Largest teacher	1.35	2.67	3.44	1.67	0.89	2.92
Smallest teacher	0.71	0.00	0.00	0.00	0.00	0.00
Largest ρ_0^+	0.71	0.00	0.00	0.00	0.00	0.00
Smallest initial rKL	1.35	2.67	0.00	1.67	0.89	0.00

In this grid, selecting the largest reinforceable fraction gives the same teacher choices as selecting the smallest teacher, so these results do not establish a selection advantage over teacher size alone. They do show that ρ_0^+ distinguishes useful pairings that minimum rKL fails to identify. We therefore interpret ρ_0^+ as a diagnostic of the supervision encountered by a particular student, capturing the fraction of sampled response tokens with positive distillation advantages. This sign information complements the distributional discrepancy summarized by rKL, although neither statistic alone fully describes downstream task gains. These findings motivate our composition experiments, which examine how student preparation changes the supervision it receives from a fixed teacher.

B.7 Teacher-Dependent Training Dynamics

We examine how teacher size shapes the evolution of distillation diagnostics for Qwen3-4B-Base. Figure 11 summarizes three runs per teacher. We normalize each run's reverse KL by its first logged value and average diagnostics over consecutive 20-step intervals before computing means and sample standard deviations across runs.

All three teachers show a pronounced early decrease in normalized reverse KL, accompanying the initial accuracy gains in Figure 9c. After this initial transition, normalized reverse KL remains lowest for the 8B teacher, intermediate for 14B, and highest for 32B. The smaller teacher therefore permits a greater proportional reduction in the initial student–teacher discrepancy. However, continued teacher matching does not consistently improve task performance: reverse KL remains low even when accuracy falls below its earlier maximum. The reinforceable fraction follows a different pattern. The 8B teacher initially provides the largest fraction of positive supervision and the 32B teacher the smallest, but these differences narrow and their ordering changes during training. Thus, the persistent ordering in normalized reverse KL does not imply a persistent ordering in positive signal. Together, these trajectories show that teacher matching, reinforceable supervision, and task performance capture distinct aspects of distillation. Initialization diagnostics help characterize the starting pairing, while subsequent training should be assessed alongside task accuracy.

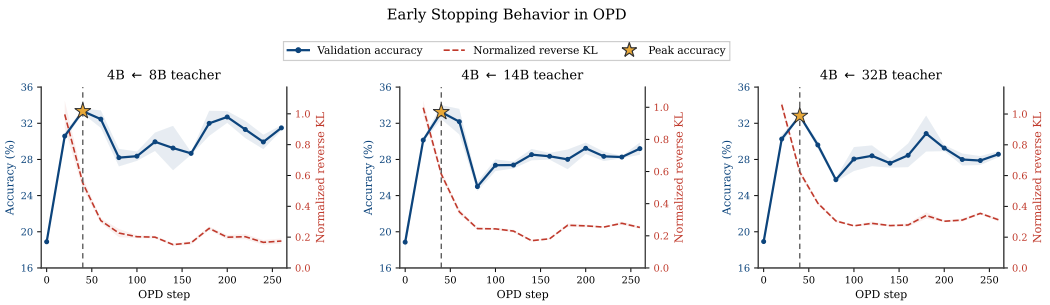


Figure 10: **Task accuracy and teacher matching follow different trajectories.** Colored curves show mean benchmark accuracy for Qwen3-4B-Base, with sample-deviation bands across three runs per teacher. Dashed gray curves show mean rKL normalized by each run's first logged value, using the right axis. Dotted vertical lines mark the peak of each mean accuracy curve through step 200.

B.8 OPD CAN LOSE TASK PERFORMANCE DESPITE CONTINUED TEACHER MATCHING.

We observed that 4B student runs improve rapidly, but further teacher matching does not consistently preserve their early gains. Across the nine runs, stopping at step 40 gives 33.13% mean accuracy, only 0.10 points below the mean of their individual-run peaks through step 200. Continuing to step 200 instead yields 30.39%, a mean gap of 2.84 points. The step-40 states therefore retain more accuracy with 80% fewer optimizer updates than step 200. An exploratory $\kappa \leq 0.6$ stopping rule selects step 48.9 on average and gives 32.14% accuracy, with a 1.09-point gap and 75.6% fewer updates than step 200. The fixed step-40 comparison performs better in these runs, while the threshold rule illustrates how a training-time diagnostic can be evaluated against a simple budget-based alternative. These comparisons are retrospective. Appendix B.9 provides the threshold sensitivity analysis and step accounting, including the extended step-260 reference. These results motivate evaluating whether OPD still improves task performance when allocating training to a subsequent stage.

B.9 EARLY STOPPING AND TRAINING-BUDGET ACCOUNTING

We compare fixed-step and KL-based stopping rules across the nine Qwen3-4B-Base runs to quantify how much task performance can be retained with fewer OPD updates. The analysis uses recorded evaluations at 20-step intervals through step 260. For each run, we measure the accuracy gap in percentage points between its selected evaluation and its highest observed accuracy within this horizon, then average these gaps across runs. A smaller gap indicates better retention of the observed peak. These *retrospective comparisons* use evaluated policy states and do not assume that every state was retained as a saved checkpoint.

For the KL-based rule, let $d_r(s)$ denote run r's logged sampled-token reverse KL at step s. We normalize it by the first logged value:

$$\kappa_r(s) = \frac{d_r(s)}{d_r(1)}, \tag{14}$$

and stop at the first evaluated step $s \geq 20$ satisfying $\kappa_r(s) \leq \tau$, using step 260 if no crossing occurs. The rule uses the unsmoothed statistic at each evaluation step. We compare $\tau \in \{0.4, 0.5, 0.6, 0.7, 0.8\}$, where a smaller threshold requires a larger reduction in rKL before stopping.

Table 13 shows that a fixed stop at step 40 achieves 33.13% mean accuracy, only 0.10 points below the mean of the individual-run peaks (33.23%). Continuing to steps 200 and 260 instead yields 30.39% and 29.75%. The step-40 state therefore retains more accuracy with 80% fewer updates than step 200 and 84.6% fewer than step 260. The $\tau = 0.6$ rule stops at step 48.9 on average and achieves 32.14% accuracy, with a mean gap of 1.09 points. Its effect varies across teachers: the gaps are 0.04, 0.00, and 3.23 points for the 8B, 14B, and 32B teachers, respectively. Thresholds 0.7 and 0.8 select step 40 in every run and match the fixed-step result, while lower thresholds wait longer and produce larger mean gaps in this sweep. These comparisons identify an effective early stopping opportunity in the observed runs, but do not independently validate a stopping step or threshold for new student–teacher pairs.

Table 13: Stopping comparisons averaged over nine Qwen3-4B-Base runs. Gap is measured from each run's observed peak through step 260. Update savings use step 260 as the reference.

Stopping rule	Mean step	Accuracy (%)	Gap (pp)	Updates saved (%)
Fixed step 40	40.0	33.13	0.10	84.6
Fixed step 60	60.0	31.41	1.82	76.9
Fixed step 100	100.0	27.92	5.31	61.5
Fixed step 200	200.0	30.39	2.84	23.1
Fixed step 260	260.0	29.75	3.48	0.0
$\tau = 0.4$	66.7	30.13	3.11	74.4
$\tau = 0.5$	60.0	31.41	1.82	76.9
$\tau = 0.6$	48.9	32.14	1.09	81.2
$\tau \in \{0.7, 0.8\}$	40.0	33.13	0.10	84.6

With 64 prompts per update and four responses per prompt, step 40 requires 2,560 prompt presentations and 10,240 sampled responses. The $\tau = 0.6$ rule averages approximately 3,129 prompt presentations and 12,516 responses, compared with 16,640 and 66,560 at step 260. These counts include repeated presentations of a question. Reported savings concern optimizer updates and sampled responses, while token and compute costs also depend on response length, teacher scoring, optimization, and evaluation. The practical implication is to assess whether further OPD still improves task accuracy when deciding how much training to allocate before a subsequent stage.

C Student Warm-Up Before Distillation

This appendix provides additional implementation details and analyses for the student warm-up experiments in §4.1. We describe the SFT and RLVR warm-up procedures, the shared downstream OPD setup, and the effect of SFT duration on subsequent distillation.

C.1 SFT Warm-Up

We warm up Qwen3-4B-Base with full-parameter SFT on correct trajectories generated by Qwen3-14B for OpenThoughts-Math prompts. We sweep the resulting student checkpoints at $K \in \{25, 75, 150, 250, 348\}$ training steps. Each checkpoint independently initializes a downstream OPD run against the same fixed Qwen3-14B teacher, allowing us to isolate how the amount of SFT before distillation changes the subsequent optimization. Table 14 reports the complete training configuration.

Table 14: Student SFT warm-up hyperparameters.

Setting	**Student SFT Warm-Up**
Starting student	Qwen3-4B-Base
Solution generator	Qwen3-14B
Prompt source	OpenThoughts-Math
Supervision	Fixed teacher-generated solutions
Update scope	Full-parameter SFT
Training backend / GPUs	FSDP / 8
Microbatch per GPU	4
Learning rate	10^{-5}
Training duration	3 epochs
Maximum sequence length	14,336 tokens
Truncation	Right
Sequence parallel size	2
Padding removal	Enabled
Save / validation interval	25 / 25 steps
Validation criterion	SFT loss

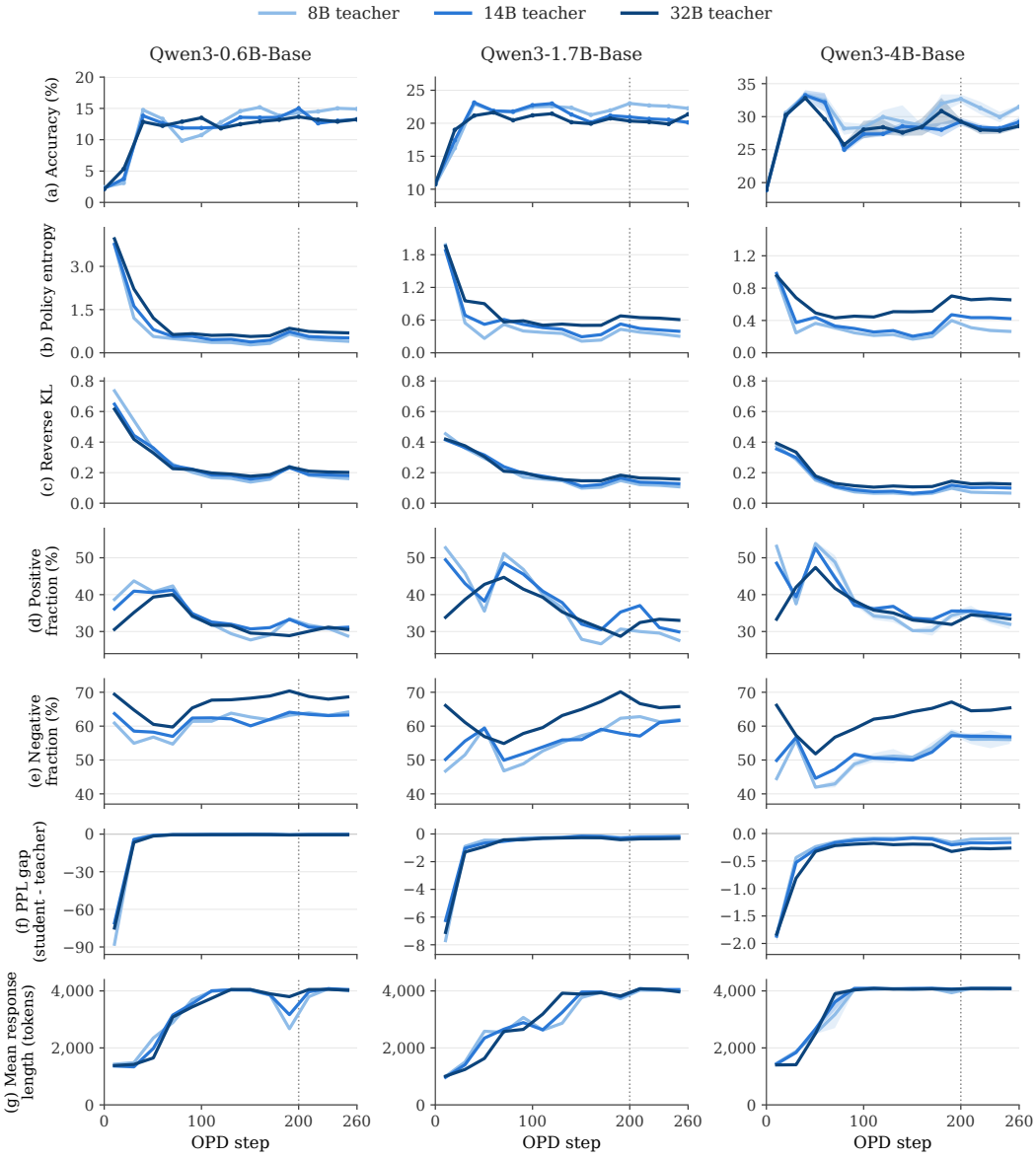


Figure 11: OPD training dynamics across student and teacher scales.

C.2 RLVR WARM-UP

For the RLVR warm-up control, we instead train Qwen3-4B-Base on OpenThoughts-Math with GRPO and a binary verifiable outcome reward. For each prompt, the student samples eight responses and receives reward according to solution correctness. Unlike SFT warm-up, this stage improves the student directly from task reward and does not use the downstream distillation teacher as supervision.

Selected RLVR checkpoints are subsequently distilled from the same fixed Qwen3-14B teacher on DAPO-Math-17K. The OPD step counter is restarted at this handoff, so warm-up optimization and downstream distillation are treated as separate stages. Table 15 reports the RLVR configuration.

Table 15: Student RLVR warm-up hyperparameters.

Setting	**Student RLVR Warm-Up**
Starting student	Qwen3-4B-Base (Thinking Disabled)
Algorithm	GRPO
Warm-up data	OpenThoughts-Math
Reward	Verifiable outcome (binary)
Update scope	Full-parameter training
Prompt / PPO minibatch size	64 / 64
Responses per prompt	8
Rollout temperature	1.0
Prompt / response limit	2,048 / 8,192 tokens
Learning rate / warm-up	10^{-6} / 10 steps
Weight decay / gradient clip	0.01 / 1.0
Lower / upper clipping ratio	0.20 / 0.28
Rollout importance correction	Token level; threshold 2.0
KL reward / KL loss	Disabled / disabled
Entropy coefficient / loss aggregation	0 / token mean
Data shuffling / overlong-prompt filtering	Disabled / enabled
Training schedule	1 epoch; no explicit step cap
Save / evaluation interval	20 / 20 steps
Evaluation temperature / top-p	0.6 / 0.95
Evaluation responses / top-k	4 / disabled
Training backend / GPUs	FSDP / 8
Rollout backend / tensor parallelism	vLLM / 1
Sequence parallel size	1
Dynamic token budget per GPU	32,768
Padding removal / gradient checkpointing	Enabled / enabled

C.3 DOWNSTREAM OPD SETUP

All warm-up checkpoints are evaluated under the same downstream OPD recipe. We distill from a fixed Qwen3-14B teacher on DAPO-Math-17K using the sampled-token reverse-KL policy-gradient objective described in §2. Student initialization is therefore the only quantity changed across the SFT checkpoint sweep. We evaluate every 20 OPD steps using four responses per problem at temperature 0.6 and top-$p = 0.95$. Reported accuracy is mean@4 across AIME24, AIME25, AMC23, GPQA, MATH-500, Minerva, and Olympiad-Bench. Table 16 gives the complete downstream configuration.

Table 16: Downstream OPD hyperparameters for the SFT warm-up sweep.

Setting	**Downstream OPD**
Student initialization	Selected SFT checkpoint
Teacher	Qwen3-14B
Training data	DAPO-Math-17K
Objective	Sampled-token reverse KL policy gradient
Task rewards / auxiliary KL loss	Disabled / disabled
Optimizer	AdamW
Learning rate / warm-up	10^{-6} / 10 steps
LR schedule / Adam betas	Constant / $(0.9, 0.999)$
Weight decay / gradient clip	0.01 / 1.0
Prompt / PPO minibatch size	64 / 64
PPO epochs / clipping ratio	1 / 0.2
Responses per prompt	4
Rollout temperature / top-p	1.0 / 1.0
Prompt / response limit	2,048 / 8,192 tokens
Loss aggregation / entropy coefficient	Token mean / 0
Distillation coefficient / configured top-k	1 / 64
Loss / log-probability clamps	10 / -10
Training duration	1 epoch; 200-step cap
Save / evaluation interval	50 / 20 steps
Evaluation temperature / top-p	0.6 / 0.95
Evaluation responses / top-k	4 / disabled
Data shuffling / overlong-prompt filtering	Disabled / enabled
Student / teacher inference dtype	bfloat16 / bfloat16
Student / teacher GPUs	4 / 4
Student rollout / teacher TP	1 / 2
Student / teacher GPU memory fraction	0.70 / 0.85
Dynamic token budget per GPU	24,576
GPU type	NVIDIA H200

C.4 Sensitivity to SFT Warm-Up Duration

Figure 12 shows that the benefit of SFT warm-up appears quickly and then saturates. The downstream OPD peak increases from 30.87% at $K = 25$ to 35.06% at $K = 75$ and 35.39% at $K = 150$. Extending SFT further gives peaks of 35.08% at $K = 250$ and 35.18% at $K = 348$. Thus, most of the downstream gain is already realized by $K = 75$–150, and additional SFT does not raise the OPD ceiling. The longer warm-ups also largely preserve their peak performance throughout OPD. In particular, the $K = 150$ run finishes at 35.30%, only 0.09 points below its best observed accuracy, while the $K = 250$ run finishes at its peak. By contrast, the minimally warmed $K = 25$ student both reaches a lower maximum and exhibits a larger subsequent drop. These trajectories support using a short SFT phase to prepare the student rather than continuing SFT until convergence.

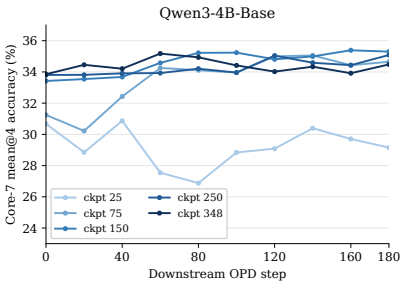


Figure 12: **Downstream OPD trajectories from different SFT warm-up checkpoints.**

Table 17: Downstream OPD results across SFT warm-up checkpoints.

SFT Checkpoint K	Initial (%)	Peak (%)	Peak t	Final (%)	Drop
25	30.68	30.87	40	29.15	1.71
75	31.25	35.06	140	34.65	0.41
150	33.42	35.39	160	35.30	0.09
250	33.81	35.08	180	35.08	0.00
348	33.85	35.18	60	34.48	0.70

C.5 Initialization Diagnostics

The checkpoint sweep also shows that continued SFT keeps changing the student even after its downstream benefit has saturated. Student entropy at the beginning of OPD decreases from 0.2940 at $K = 25$ to 0.1950 at $K = 150$ and 0.1507 at $K = 250$, yet the strongest downstream OPD result is already obtained at $K = 150$. Lower initialization entropy therefore does not monotonically translate into better distillation.

A similar pattern appears in the initial teacher–student reverse KL. It falls from 0.1417 at $K = 25$ to its minimum of 0.1028 at $K = 150$, but further SFT does not produce additional downstream gains. Together with the validation-loss curve in Figure 5d, this places the useful warm-up region around $K = 75$–150: the first validation-loss plateau coincides with saturation in downstream OPD. We use this as a practical signal of diminishing returns rather than as a universal stopping rule.

Table 18: Student diagnostics at the beginning of downstream OPD.

SFT Checkpoint K	OPD Step t	Student Entropy	Reverse KL
25	1	0.2940	0.1417
75	1	0.2559	0.1227
150	1	0.1950	0.1028
250	1	0.1507	0.1153
348	1	0.1580	0.1117

D Teacher Adaptation Before Distillation

This appendix provides the details of teacher preparation and the corresponding downstream OPD settings for §4.2, along with additional results and discussion for interested readers.

D.1 Experimental Setup

We compare Qwen3-14B teachers adapted using either SFT on positive self-generated rollouts ((Zelikman et al., 2022)) or RLVR on OpenThoughts3-Math. Each teacher then distills a fresh Qwen3-4B-Base student on DAPO-Math-17K under identical downstream settings, including student initialization, OPD objective, learning rate, batch size, sampling, and evaluation. Table 19 reports the SFT teacher-adaptation hyperparameters, while Table 21 reports the RLVR hyperparameters, separating teacher preparation from downstream student training.

D.2 Teacher Adaptation with SFT

For the SFT-adapted teacher, we start from Qwen3-14B with thinking disabled and construct supervision directly from its own rollouts on OpenThoughts3-Math. For each prompt, we sample up to four independent responses and retain the prompt only if at least one rollout is verified as correct by the task reward. Incorrect generations are discarded, and sampling is repeated, up to the four-generation budget, until a verified-correct solution is obtained. This produces approximately 16K unique prompts paired with verified self-generated solutions, matching the number of unique training prompts used for RLVR teacher adaptation and thereby controlling for adaptation-data scale. We then perform full-parameter SFT on this fixed dataset for three epochs using AdamW with a learning rate of 10^{-6} and a maximum sequence length of 14,336 tokens. Selected checkpoints from this training trajectory are subsequently used as teachers for downstream OPD.

Table 19: SFT teacher training hyperparameters.

Setting	**SFT Teacher**
Starting teacher	Qwen3-14B (Thinking Disabled)
Adaptation data	OpenThoughts3-Math
SFT format	Full-parameter
Supervision	Fixed, verified **self**-generated solutions
Unique prompts in batch	64
Learning rate	10^{-6}
Maximum sequence length	14,336 tokens
Truncation	Right
Training schedule	3 epochs
Optimizer	AdamW
Training backend	FSDP
GPUs	8

D.3 RLVR Teacher Adaptation

For the RLVR-adapted teacher, we start from the same Qwen3-14B model and use the same approximately 16K OpenThoughts3-Math training prompts considered for SFT adaptation in §D.2. This controls the underlying adaptation data and isolates the effect of the teacher-training objective. We train the teacher with GRPO using the common RLVR optimization and rollout recipe adopted throughout this work: for each prompt, we sample eight responses at temperature 1.0 and optimize against a binary verifiable-outcome reward indicating whether the generated solution is correct. Training uses batches of 64 unique prompts, a learning rate of 10^{-6} with a 10-step warm-up, and a single pass over the adaptation set. We disable both KL regularization and entropy bonuses and otherwise retain the RLVR optimization, clipping, importance-correction, and rollout settings used in our main experiments. To study how progressive RLVR adaptation of the teacher affects subsequent distillation, we save checkpoints every 20 training steps and select five checkpoints along the training trajectory for downstream OPD evaluation. Each checkpoint is used to distill a freshly initialized Qwen3-4B-Base student under the same downstream OPD configuration, allowing differences in student performance to be attributed to the teacher's degree of RLVR adaptation. Table 20 reports the complete RLVR teacher-training configuration.

Table 20: RLVR teacher training hyperparameters.

Setting	**RLVR-Adapted Teacher**
Starting teacher	Qwen3-14B
Algorithm	GRPO
Adaptation data	OpenThoughts-Math
Reward	Verifiable outcome (binary)
Unique prompts in batch	64
Responses per prompt / temperature	8 / 1.0
Prompt / response limit	2,048 / 8,192 tokens
Learning rate / warm-up	10^{-6} / 10 steps
Weight decay / gradient clip	0.01 / 1.0
Lower / upper clipping ratio	0.20 / 0.28
Rollout importance correction	Token level w/. threshold 2.0
KL loss	Disabled
Entropy coefficient / loss aggregation	0 / token mean
Data shuffling / overlong-prompt filtering	Disabled / enabled
Training schedule	1 epoch; no explicit step cap
Save / evaluation interval	20 / 20 steps
Evaluation temperature / top-p	0.6 / 0.95
Evaluation samples / top-k	4 / disabled
Training GPUs / rollout TP	8 / 1
Rollout backend / GPU memory fraction	vLLM / 0.60
Actor parameter / optimizer offload	Enabled / enabled

D.4 Downstream OPD Configuration and Evaluation

Table 21 shows our downstream OPD hyperpaarameters after both SFT and RLVR teacher. Its uses sampled-token reverse-KL policy gradients on student-generated responses, with task rewards and auxiliary KL regularization disabled. The student uses the non-thinking chat-template setting throughout OPD. The SFT and RLVR teacher runs both runs for 200 steps and validated in every 20 steps. During evaluaton, for each checkpoint, we sample four responses per evaluation question at temperature 0.6 and top-$p = 0.95$. We compute accuracy as the unweighted mean of the seven benchmark-specific mean@4 scores: AIME24, AIME25, AMC23, MATH-500, Minerva, Olympiad-Bench, and GPQA. Thus, each benchmark receives equal weight, and mean@4 averages correctness over the four responses.

D.5 Teacher Selection at Fixed Model Size

To examine whether initialization diagnostics distinguish useful teachers beyond parameter count, we compare four RLVR-adapted Qwen3-14B checkpoints using Qwen3-1.7B-Base. Each checkpoint distills a freshly initialized student under identical downstream OPD settings. Table 22 reports reverse KL and reinforceable fraction at the first logged training step, together with accuracy at OPD step 200. All runs use rollout seed 42, and checkpoint labels follow Figure 13.

Table 22: Initial diagnostics and step-200 accuracy for Qwen3-1.7B-Base with RLVR-adapted Qwen3-14B teachers. Each configuration uses one run.

Teacher K	rKL_0	ρ_0^+ (%)	A_{200} (%)
50	0.505	47.23	27.41
100	**0.476**	52.40	28.00
150	0.555	49.08	29.16
200	0.530	**52.97**	**29.77**

Selecting maximum ρ_0^+ identifies $K = 200$, which achieves the highest accuracy in this comparison at 29.77%. Minimum reverse KL instead selects $K = 100$, yielding 28.00% and incurring 1.77 percentage points of selection regret. Thus, the teacher closest to the student under initial reverse KL does not provide the most effective supervision at the evaluated training budget. Because all candidates share the same parameter count, this comparison illustrates the diagnostic value of ρ_0^+ beyond teacher size. Its sign information helps distinguish useful checkpoints that

Table 21: Downstream OPD training hyperparameters shared across SFT- and RLVR-adapted teachers.

Setting	SFT/RLVR Teacher
Student	Qwen3-4B-Base
Training data	DAPO-Math-17K
Objective	Sampled-token reverse KL policy gradient
Task rewards / auxiliary KL loss	Disabled / disabled
Optimizer	AdamW
Learning rate / warm-up	10^{-6} / 10 steps
LR schedule / Adam betas	Constant / $(0.9, 0.999)$
Weight decay / gradient clip	0.01 / 1.0
Prompt size	64
PPO epochs / clipping ratio	1 / 0.2
Rollout temperature / top-p	1.0 / 1.0
Prompt / response limit	2,048 / 8,192 tokens
Loss aggregation / entropy coefficient	Token mean / 0
Distillation coefficient / configured top-k	1 / 64
Loss / log-probability clamps	10 / -10
Configured epochs / step cap	1 / 200
Save / evaluation interval	20 / 20 steps
Eval temperature / top-p	0.6 / 0.95
Evaluation responses / top-k	4 / disabled
Student FSDP / inference dtype	bfloat16 / bfloat16
Student rollout / teacher TP	1 / 2
Student / teacher GPU memory fraction	0.70 / 0.85
Dynamic token budget per GPU	24,576
GPU type	8× NVIDIA B200

minimum divergence does not identify, supporting its use as a complementary measure of the initial student–teacher pairing in this setting.

D.6 Additional Results

Figure 13 shows the full downstream OPD trajectories for individual teacher checkpoints. With SFT adaptation, the runs remain closely aligned through approximately OPD step 140: more heavily adapted teachers converge to nearly the same accuracy as earlier checkpoints, with only small differences, before several runs decline afterward. This is consistent with SFT changing the teacher policy without materially increasing the amount of task capability available for distillation.

RLVR exhibits the opposite trend. Up to roughly OPD step 100, stronger RLVR teacher checkpoints generally produce progressively higher student accuracy. Beyond this point, however, the less-adapted checkpoints, particularly $K = 50$ and $K = 100$, become substantially less stable, while the more strongly adapted teachers retain higher performance. Among the later checkpoints, $K = 200$ reaches the strongest performance, slightly ahead of $K = 250$, with both outperforming $K = 150$, which in turn improves over the earlier checkpoints. These trajectories suggest a capability–stability trade-off in teacher adaptation: increasing RLVR initially strengthens the transferable learning signal, but the benefit saturates once teacher capability plateaus. In practice, this suggests stopping teacher RLVR near its capability plateau and selecting the teacher checkpoint using downstream OPD validation rather than simply using the final checkpoint.

E Downstream RLVR Initialization

In this section of Appendix, we detail the stage-ordering comparison in §4.3: SFT and OPD prepare a student for the same downstream RLVR procedure. We report the initialization and RLVR configurations separately, followed by additional results across student scales and model families.

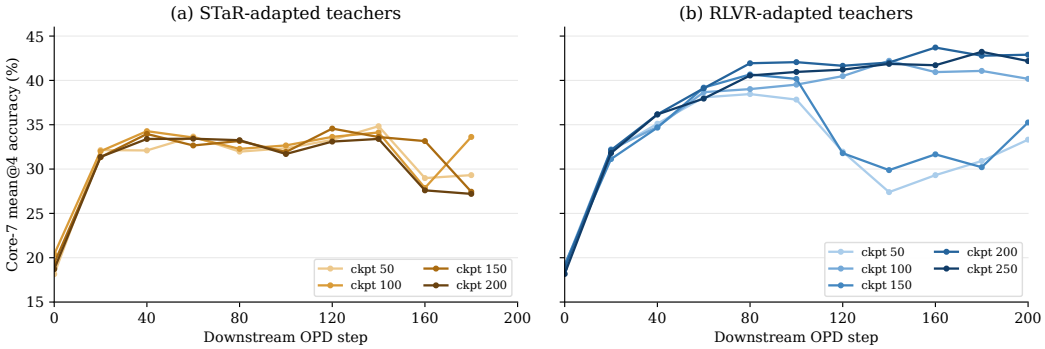


Figure 13: Downstream OPD performances across SFT and RLVR adapted teacher checkpoint sweeps.

E.1 Experimental Setting

For each student, the two initialization methods use the same starting model and the same fixed teacher. Qwen3 experiments pair Qwen3-Base students (1.7B and 4B) with Qwen3-14B and the Qwen2.5 experiments pair Qwen2.5-1.5B-Base with Qwen2.5-14B-Instruct. SFT learns from fixed, verified teacher-generated solutions, whereas OPD obtains teacher feedback on student-sampled responses. Both preparatory stages use OpenThoughts3-Math data. Their checkpoints then initialize GRPO on DAPO-Math-17K, with the downstream settings held fixed across the two arms within each student comparison. The initialization launchers specify a batch size of 64 and learning rate 10^{-6}. SFT uses one stored solution per training example, while OPD samples four responses per prompt following but the number of unique training samples each see is same following Li et al. (2026) and Zhu et al. (2026).

E.2 SFT Initialization

Table 23 gives the SFT configuration, which is shared by every student within a model family. The SFT data were built from the teacher that each family's students are later distilled from: Qwen3-14B for the Qwen3 students and Qwen2.5-14B-Instruct for the Qwen2.5 students, so that the comparison covers both a range of student sizes and a second architecture. We took the prompts from OpenThoughts3-Math, sampled one solution per prompt from the corresponding teacher, graded it against the gold answer, and kept only the prompts whose solution was correct. The construction script retains one verified teacher solution per solvable question and holds out 256 of these examples for SFT validation. It passes the full retained question pool, including those 256, to OPD as its prompt set. The two stages therefore see the same unique training prompts in the same order, and differ only in that SFT withholds the validation subset.

E.3 OPD Initialization

OPD needs no teacher rollouts. The student generates the trajectories, and the teacher, which is larger and more capable, scores every token of them, so the stage requires only a verifiable prompt set with gold answers. We use the same OpenThoughts3-Math questions as the SFT stage, in the same number and the same order, so that at every optimizer step the two stages have consumed the same unique prompts. Table 24 gives the OPD configuration. The student samples responses at temperature 1.0 and learns from the fixed teacher through the sampled-token reverse-KL policy-gradient objective of Equation 4. The OPD checkpoints shown in the figures use the same step sets K listed above for each student, so that the comparison covers several hand-off points.

E.4 Common Downstream RLVR Stage

Each selected checkpoint is exported to Hugging Face format and loaded as the initial policy for a separate GRPO run. Table 25 applies to both initialization arms. RLVR stage runs for one epoch over

Table 23: **SFT initialization settings.** The reported comparison on Qwen3-1.7B-Base, Qwen3-4B-Base, and Qwen2.5-1.5B students.

Setting	Qwen3 family	Qwen2.5 family
Supported base students	0.6B, 1.7B, 4B	1.5B
Teacher	Qwen3-14B	Qwen2.5-14B-Instruct
Training data	OpenThoughts3-Math	
Supervision	Fixed, verified teacher-generated solutions	
Unique prompts in batch	64	64
Microbatch per GPU	4	4
Learning rate	10^{-6}	10^{-6}
Maximum sequence length	6,145 tokens	6,145 tokens
Truncation	Right	Right
Training schedule	1 epoch; 250-step cap	1 epoch; 250-step cap
Save / validation interval	25 / 25 steps	25 / 25 steps
Training backend	FSDP	FSDP
Student GPUs	8	8
Sequence parallel size	2	2

Table 24: **OPD initialization settings.** All students use the same optimization and sampling settings, the teacher and initialization data follow the model family.

Setting	Qwen3 family	Qwen2.5 family
Supported base students	0.6B, 1.7B, 4B	1.5B, 3B
Teacher	Qwen3-14B	Qwen2.5-14B-Instruct
Training data	OpenThoughts3-Math	
Distillation objective	Sampled-token policy gradient reverse-KL	
Task rewards	Disabled	Disabled
Prompt / PPO minibatch size	64 / 64	64 / 64
Responses per prompt	4	4
Rollout temperature	1.0	1.0
Prompt / response limit	2,048 / 8,192 tokens	2,048 / 8,192 tokens
Learning rate	10^{-6}	10^{-6}
LR warm-up	10 steps	10 steps
Weight decay / gradient clip	0.01 / 1.0	0.01 / 1.0
Clipping ratio	0.2	0.2
Loss aggregation	Token mean	Token mean
Auxiliary KL / entropy loss	Disabled / 0	Disabled / 0
Configured distillation top-k	64	64
Loss / log-probability clamps	10.0 / -10.0	10.0 / -10.0
Training schedule	1 epoch; 250-step cap	1 epoch; 250-step cap
Save / evaluation interval	25 / 25 steps	25 / 25 steps
Student / teacher GPUs	4 / 4	4 / 4
Rollout / teacher tensor parallelism	1 / 1	1 / 1
Dynamic token budget per GPU	24,576	24,576

the processed DAPO-Math-17K data with no explicit step cap, so that every run covers the full dataset and each training prompt is seen by the optimizer exactly once. Evaluation samples four responses per question at temperature 0.6 and top-$p = 0.95$, with top-k filtering disabled on Qwen3 series and use temperature 0.7 for Qwen2.5 followig their suggested HuggingFace decoding parameters. Similar to §3, we report held-out `core-7` mean@4 accuracy on seven core reasoning benchmarks covering math and general reasoning: AIME24, AIME25, AMC23, MATH-500, Minerva, Olympiad-Bench, and GPQA.

Table 25: Common Downstream RLVR settings for SFT and OPD-initialized students. The reported horizons reflect the displayed curves, while the launchers specify an epoch-based training schedule.

Setting	Qwen3 family	Qwen2.5 family
Algorithm	GRPO	GRPO
Training data	DAPO-Math-17K	DAPO-Math-17K
Reward	Verifiable outcome reward (`ttrl_math.reward_func`)	
Prompt / PPO minibatch size	64 / 64	64 / 64
Responses per prompt	8	8
Rollout temperature	1.0	1.0
Prompt / response limit	2,048 / 8,192 tokens	2,048 / 8,192 tokens
Learning rate / warm-up	10^{-6} / 10 steps	10^{-6} / 10 steps
Weight decay / gradient clip	0.01 / 1.0	0.01 / 1.0
Lower / upper clipping ratio	0.20 / 0.28	0.20 / 0.28
Rollout importance correction	Token level; threshold 2.0	Token level; threshold 2.0
KL reward / KL loss	Disabled / disabled	Disabled / disabled
Entropy coefficient	0	0
Loss aggregation	Token mean	Token mean
Data shuffling	Disabled	Disabled
Training schedule	1 epoch; no explicit step cap	1 epoch; no explicit step cap
Reported RLVR horizon	260 steps (1.7B, 4B)	200 steps (1.5B)
Save / evaluation interval	20 / 20 steps	20 / 20 steps
Evaluation temperature / top-p	0.6 / 0.95	0.6 / 0.95
Evaluation samples per question	4	4
Student GPUs / rollout TP	8 / 1	8 / 1
Dynamic token budget per GPU	32,768	32,768

E.5 Generalization Across Student Scale, Model Family, and Handoff Checkpoint

We test whether the stage-ordering result in §4.3 generalizes beyond the main Qwen3-4B setting. Figure 14 repeats the OPD-vs. SFT initialization comparison for Qwen3-1.7B and Qwen2.5-1.5B. Within each student family, OPD and SFT are compared using the same student, teacher, RLVR data, RLVR recipe, and matched handoff-checkpoint grid. Across every matched handoff setting, the OPD-initialized run finishes above its SFT-initialized counterpart.

The final accuracy ranges are 26.66–27.53% versus 23.13–25.70% for Qwen3-1.7B, 38.53–43.11% versus 33.36–36.80% for Qwen3-4B, and 20.12–21.54% versus 18.69–19.61% for Qwen2.5-1.5B. Thus, the endpoint ordering is reproduced across all tested handoff checkpoints, two Qwen3 student scales, and the Qwen2.5 model family.

The strength of the control differs across settings. For Qwen3-4B, OPD and SFT start from overlapping accuracy ranges and SFT is slightly stronger on average before RLVR, yet every OPD-initialized run finishes higher. For the smaller students, OPD also begins RLVR at higher accuracy, so these experiments establish a consistent endpoint advantage but do not isolate the downstream RLVR gain from matched initial performance. The checkpoint sweep provides repeated evidence across independently chosen handoff states, but it is not a substitute for variation across random training seeds. We therefore use these experiments to support cross-model and cross-checkpoint reproducibility of the endpoint ordering rather than a claim of seed-level statistical significance.

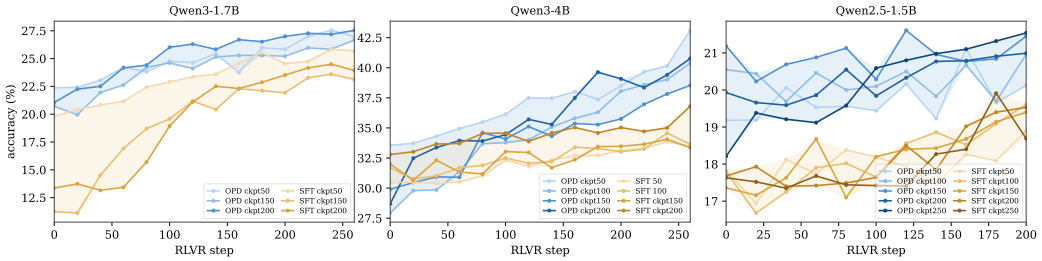


Figure 14: **The OPD initialization advantage reproduces across matched handoff checkpoints, student scales, and model families.** Held-out `core-7` mean@4 accuracy for Qwen3-1.7B-Base (left), Qwen3-4B-Base (center), and Qwen2.5-1.5B (right), using their corresponding 14B teachers. Within each student family, OPD and SFT use the same downstream RLVR recipe and matched handoff-checkpoint grid. Blue and orange curves denote OPD and SFT initialization. Shading shows the minimum–maximum range across handoff checkpoints within each arm, not variation across random seeds. The reported RLVR horizons are 260 steps for Qwen3 and 200 for Qwen2.5-1.5B.

F Case Study: Can Post-Training Composition Be Recursive?

We ask whether repeated SFT–OPD composition remains productive as the student evolves, and how the teacher schedule affects this process. Iterative self-training and self-play have been studied in STaR, ReST, and SPIN (Zelikman et al., 2022; Gulcehre et al., 2023; Chen et al., 2024). Here, we examine repeated composition of two stages while varying the teacher schedule. The teacher checkpoints are prepared in advance; the student carries its learned state from one round to the next.

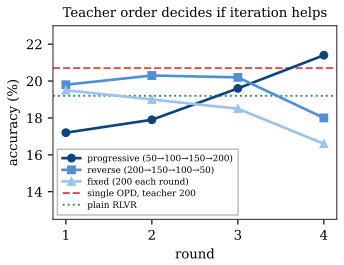


Figure 15: **Recursive composition under different teacher schedules.** Four SFT–OPD rounds on Qwen3-1.7B-Base, each using an RLVR-adapted Qwen3-14B checkpoint. Accuracy averages AIME 2024, AIME 2025, AMC 2023, and Minerva Math.

Starting from Qwen3-1.7B-Base, each round applies 200 SFT steps on trajectories from one RLVR-adapted Qwen3-14B checkpoint, followed by a short OPD stage from the same teacher. Handoffs use the observed OPD accuracy peak in each round. We compare progressive ($t_{50} \to t_{100} \to t_{150} \to t_{200}$), reverse ($t_{200} \to t_{150} \to t_{100} \to t_{50}$), and fixed ($t_{200}$ in every round) schedules. Progressive and reverse schedules reorder the same teacher checkpoints; the fixed schedule instead tests repeated use of the final teacher. A single 200-step OPD run from t_{200} and a plain RLVR run serve as references. The recursive schedules include additional SFT and OPD training, so total compute is not matched to these references. Figure 15 reports accuracy averaged over AIME 2024, AIME 2025, AMC 2023, and Minerva Math. Each schedule has one run.

Why this matters. Teacher post-training produces a sequence of candidate supervision policies. This permits a comparison between distilling directly from the final teacher and exposing the student to progressively more adapted checkpoints. The case study asks whether the sequence of student–teacher pairings matters when SFT and OPD are repeated.

Table 26: Diagnostics recorded at the start of OPD and four-benchmark accuracy for every round using teacher t_{200}. Reinforceable fractions are shown on a 0–1 scale.

Schedule	Round	rKL	ρ^+	Acc. (%)
Fixed	1	.42	.48	19.5
Fixed	2	.39	.46	19.0
Fixed	3	.38	.45	18.5
Fixed	4	.38	.44	16.6
Reverse	1	.43	.47	19.8
Progressive	4	.41	.44	21.4

Discussion. Under the fixed schedule, the reinforceable fraction falls from 0.48 to 0.44, while four-benchmark accuracy decreases from 19.5% to 16.6% (Table 26). Reverse KL also decreases, from 0.42 to 0.38, with little further change after round 2. Repeated exposure to the final teacher therefore does not sustain improvement in this run. The reinforceable fraction is also insufficient to

explain differences across schedules: progressive and fixed round 4 both report $\rho^+ = 0.44$, yet their accuracies are 21.4% and 16.6%, respectively.

The progressive schedule reaches the highest observed four-benchmark endpoint among the recursive runs; the reverse and fixed schedules do not sustain improvement across rounds (Figure 15). On the full seven-benchmark evaluation used in the main paper, however, the best recursive result remains 0.9 percentage points below single-stage OPD.

Recursive runs reach the $8{,}192$-token response limit, truncating trajectories used for OPD; the single-stage OPD reference does not exhibit the same saturation. This difference and the single run per schedule limit the interpretation of the observed ordering. The case study motivates testing teacher schedules under matched training budgets and response-length conditions; it does not establish that recursive composition improves over a single OPD stage.

G Further Discussions

G.1 Why Composition Matters

Our results suggest that post-training stages should be viewed as transformations of a *learning state*, rather than independent sources of capability. SFT trains on a fixed trajectory distribution, whereas RLVR and OPD train on states induced by the current policy. Consequently, an earlier stage changes not only the model parameters but also the data distribution and supervision encountered by the next stage. This makes post-training updates naturally order-dependent, which we can express schematically as $\mathcal{U}_B(\mathcal{U}_A(\pi)) \neq \mathcal{U}_A(\mathcal{U}_B(\pi))$.

Our experiments expose three manifestations of this non-commutativity. First, RLVR can produce a substantially stronger student while making subsequent OPD destructive, showing that capability and distillability are distinct. Second, changing the teacher through RLVR substantially changes what can be transferred through an otherwise identical OPD stage. Third, accuracy-matched SFT and OPD checkpoints respond very differently to subsequent RLVR, even when their aggregate entropies overlap. Together, these results motivate our central principle: a post-training stage should be evaluated not only by the capability it adds, but also by the learning interface it leaves for the next stage.

G.2 Limitations

Our conclusions are deliberately narrower than a universal post-training recipe:

- **Composition is tested more narrowly than OPD scaling.** Our OPD diagnostics span multiple student and teacher scales, architectures, objectives, and datasets, whereas the main composition study centers on a Qwen3-4B student and Qwen3-14B teacher. We therefore claim a controlled demonstration of stage interaction, not that the same ordering is optimal for every model or domain.
- **Some controls separate practical procedures rather than causal mechanisms.** In particular, RLVR substantially improves the teacher while our STaR-style control does not. This establishes that RLVR is the more effective teacher-preparation method under our protocol and that its added capability transfers through OPD; it does not establish that RLVR-induced capability is intrinsically more transferable at matched teacher accuracy.
- **Our diagnostics are predictive, not sufficient explanations.** ρ_0^+ separates productive and destructive student–teacher states where scalar reverse KL does not, and entropy alone does not explain the OPD-to-RLVR advantage. These results rule out simple scalar explanations, but they do not identify the complete mechanism governing downstream learnability.

These limitations do not change the primary conclusion: across controlled interventions, *the same stage behaves differently depending on the state created by the stage before it*.

G.3 Unsuccessful Attempts

Several negative results were particularly informative because they rule out seemingly natural alternatives.

- **Maximizing pre-OPD student capability.** RLVR produces a much stronger student than a short SFT warm-up, yet subsequent OPD can regress it below its starting accuracy. Thus, selecting a warm-up solely by pre-OPD accuracy can choose the worse downstream initialization.
- **Training each stage longer.** Additional SFT eventually stops improving downstream OPD, while continued OPD can reduce task accuracy even as teacher–student KL continues to improve. More optimization of an intermediate objective is therefore not necessarily better optimization of the complete pipeline.
- **Reusing the strongest teacher recursively.** In our recursive case study, repeatedly using the final teacher does not continue to improve the student, while progressive teacher refresh is the only tested ordering that improves across rounds. Because these runs are not fully replicated and exhibit trajectory-length truncation, we interpret this as evidence that supervision can become stale as the student changes, rather than as proof of an optimal recursive algorithm.

These failures reinforce the same pattern as the positive results: locally improving one stage can worsen the conditions faced by the next.