# UNPACKING THE COMPUTE EFFICIENCY OF STRONG-TO-WEAK OPD AGAINST GRPO

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

On-policy distillation (OPD) trains a student on its own rollouts under a frozen teacher's per-token supervision. Does this dense signal reduce the training FLOPs needed to reach a target accuracy compared with reinforcement learning? We systematically compare vanilla reverse-KL OPD with Group Relative Policy Optimization (GRPO) across task difficulty, student and teacher size, domain, and optimization choices, measuring held-out accuracy against training FLOPs. With an external post-trained teacher, OPD generally loses to GRPO in accuracy per training FLOP, even when the teacher is further trained with GRPO, and often suffers response-length inflation. Yet a different teacher changes the picture: in our math setting, OPD recovers the accuracy of the student's own GRPO checkpoint with $50\times$ fewer distillation FLOPs than direct GRPO needs to reach that accuracy. On a single domain, however, the GRPO checkpoint is already the desired model, so distilling it adds to total compute. Across domains, though, separate GRPO checkpoints hold different capabilities, and consolidating them into a single model has a distinct advantage. Multi-teacher OPD (DeepSeek-AI, 2026) consolidates per-domain GRPO experts into a single student that matches each expert's accuracy, with the distillation stage costing only $5.2\%$ (on 2 domains) and $8.0\%$ (on 4 domains) of the specialist training FLOPs. Counting specialist training and distillation together, this pipeline reaches joint GRPO's performance on the same data mix at $42.0\%$ fewer total FLOPs.

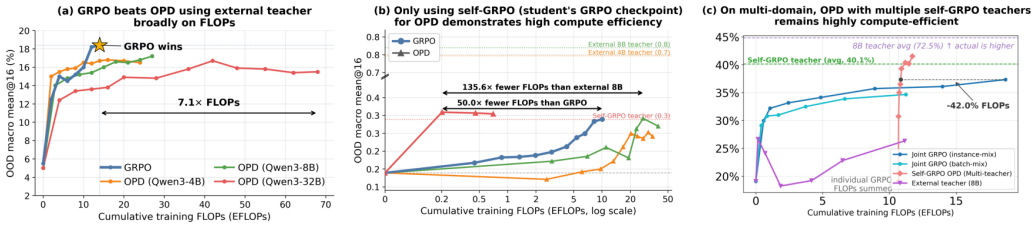



Figure 1: **OPD compute efficiency, at a glance.** (a) GRPO achieves higher performance at fewer FLOPs than OPD across external teacher sizes (up to 32B) on the OOD math evaluation (AIME24/25 and AMC23); the same ordering holds across the other settings. (b) On the OOD math eval with a `Qwen3-4B-Base` student, OPD with a *self-GRPO teacher* (a GRPO checkpoint of the student) is the only setup that matches GRPO performance with $50.0\times$ fewer FLOPs than GRPO, and $135.6\times$ fewer FLOPs than OPD with an external teacher `Qwen3-8B`. (c) On a multi-domain data mix with math and code, running multi-teacher OPD with self-GRPO teachers (per domain) matches joint GRPO at $42.0\%$ fewer total FLOPs. A single external 8B teacher substantially underperforms despite its stronger overall performance; the trained student demonstrates severe degenerate behavior of unstoppable repetition, and Chinese characters given an English prompt.

## 1 INTRODUCTION

On-policy distillation (OPD) (Agarwal et al., 2024) trains a student on its own rollouts, with each token scored by a frozen teacher. It combines the dense per-token supervision of classical knowledge distillation (Kim & Rush, 2016; Gu et al., 2024) with on-policy sampling that avoids the train/test



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.

mismatch of teacher-generated data. Dense teacher feedback suggests that OPD may learn from fewer rollouts than outcome-reward RL such as GRPO (Shao et al., 2024). Its promise is sharpest on hard problems: when the student rarely finds a correct answer, outcome-reward GRPO has little positive signal, while the teacher can still supervise every generated token. Yet that supervision requires a teacher forward pass on every training token, and longer responses raise the cost further. Fewer rollouts or optimizer steps therefore need not mean fewer FLOPs. **Can OPD reach the same accuracy as GRPO with fewer training FLOPs?** Existing efficiency reports use GPU-hours, gradient-step ratios, or partial FLOPs comparisons, leaving the matched-FLOPs comparison unresolved. To our knowledge, no prior work systematically compares OPD and GRPO at matched training FLOPs (§2; Appendix A).

In this work, we focus on systematically comparing reverse-KL OPD with GRPO by tracking held-out task performance against cumulative training FLOPs. Our experiments vary data difficulty, student and teacher sizes and initializations, training domain, thinking mode, distillation objective, and the RL framework. With off-the-shelf post-trained teachers, GRPO generally achieves higher accuracy at a fixed FLOP budget. To our surprise, this even holds in the subset that should intuitively favour OPD most: a small student that barely solves ($\texttt{mean@32}=0$) paired with a large teacher that solves well ($\texttt{mean@32}>0.6$), where GRPO's outcome reward is very weak and dense per-token supervision should matter most (§4.1). External-teacher OPD always pathologically inflates response length until it hits the training cap, increasing the FLOPs paid per step (Appendix E).

We next ask whether the teacher's training history changes the result. Recent work reports that GRPO-training the teacher before distillation can help the student learn better (Li et al., 2026). Building on this evidence, we compare external thinking models, further-GRPO-trained thinking models, and GRPO-trained base models at or above the student's size. Among the teachers tested, OPD consistently recovers the teacher's accuracy only when the teacher is a GRPO checkpoint of the student. In particular, in our math setting, OPD with one rollout per prompt matches that checkpoint's accuracy with $50\times$ fewer distillation-stage FLOPs than the eight-rollout GRPO training needs to reach the same accuracy.

However, distilling from the student's own GRPO checkpoint reveals a paradox: on a single domain, why distill when you already have the GRPO-trained model? The $50\times$ ratio counts only distillation: the GRPO run has already produced the desired single domain model, so distilling it back into the base student adds cost. Consider math and code: separate GRPO runs produce two specialists, but neither checkpoint can serve both domains. This is where multi-teacher OPD has a distinct compute-efficiency opportunity: it can consolidate the GRPO specialists into the multi-domain model neither provides, with only a small additional distillation cost. The decisive compute test is whether the cost of training those experts and consolidating them via distillation beats that of joint GRPO on the same data mix. Recent open-source models (Yang et al., 2026d; Team et al., 2026; DeepSeek-AI, 2026; Kimi Team, 2026) use this recipe, but leave that full-pipeline comparison unmeasured. Multi-teacher OPD combines the per-domain GRPO experts into a single student model that recovers their performance, with distillation adding only $5.2\%$ (two domains) and $8.0\%$ (four domains) to the specialists' training FLOPs. Counting both stages, the two-domain pipeline reaches joint GRPO's accuracy on the same data mix with $42\%$ fewer total FLOPs. On four domains, joint GRPO does not even reach the merged student's performance within the tested budget.

Our contributions identify where OPD falls short of GRPO and how distillation gains a distinct compute advantage with multi-teacher OPD:

- A comprehensive set of controlled experiments shows that dense token supervision alone does not make external-teacher OPD cheaper than GRPO (§4).
- Choose teachers by provenance. Among the teachers tested, a GRPO checkpoint of the student model itself transfers most reliably and $50.0\times$ more efficiently, but its distillation-stage savings must be weighed against the cost of training it (§5).
- Specialize, then consolidate. Train separate GRPO experts on different domains and combine them with routed multi-teacher OPD, in which scenario OPD clearly offers a more compute-efficient solution than joint GRPO by producing the multi-domain model that no single specialist alone supplies (§6).

## 2 RELATED WORK

**Recent OPD work rarely accounts for FLOPs when claiming advantage over GRPO.** Many recent methods modify the canonical reverse-KL objective for external-teacher OPD to outperform GRPO, but rarely evaluate their claimed advantage over outcome-reward methods like GRPO in terms of FLOPs. Most report sample efficiency (Ko et al., 2026), GPU-hours, or wall-clock (Zhang et al., 2026; Feng et al., 2026; Chen et al., 2026; Rang et al., 2026); Zhao et al. (2026) reports FLOPs but only among OPD variants, not against GRPO. Reported wall-clock or GPU-hour wins over GRPO also use different group sizes for the two methods (Zheng et al., 2026; Tian et al., 2026) or leave it unreported (Rang et al., 2026; Zhang et al., 2026; Chen et al., 2026), so the gap is not attributable to objective quality alone. We isolate this confound by controlling group size and characterizing external-teacher OPD through FLOPs, with a smaller-group-size ablation in Appendix I.

**Negative evidence and behavior of OPD** Several concurrent studies report evidence against a plain OPD compute win. Lee et al. (2026) find OPD slower than GRPO in wall-clock time; Liu et al. (2026) report external-teacher reverse-KL OPD undergoing response-length explosion while a reward-only GRPO baseline stays stable; and Yang et al. (2026a) find OPD improves the base student but stays below the GRPO-trained student, with OPD *after* GRPO lowering the average. Further studies document problems with external-teacher OPD—diversity collapse, prefix and support drift, low-information suffixes, length inflation (Kim et al., 2026; Fu et al., 2026a; Luo et al., 2026; Wang et al., 2026b; Li et al., 2026; Wang et al., 2026c; Yang et al., 2026c; Zhu et al., 2026b; Yang et al., 2026b; Xin et al., 2026; Ding et al., 2026; Nicolicioiu et al., 2026; Wang et al., 2026a).

**Concurrent OPD analyses also omit a compute-efficiency comparison with GRPO.** Li et al. (2026) find a small same-family RL-trained teacher yields consistent gains where a stronger, more divergent one does not, but without a GRPO baseline; Luo et al. (2026); Zhu et al. (2026c); Fu et al. (2026b) likewise do not consider efficiency. In contrast, we compare GRPO and conventional OPD in a controlled setting using FLOPs as the compute measure. In Appendix A, we conduct a broader, LLM-assisted survey to look for FLOPs comparison but find no qualifying work.

**Multi-teacher OPD in frontier model reports.** Several recent frontier-model reports use multi-teacher OPD to combine domain-expert teachers into a shared student, e.g., DeepSeek-V4 and Kimi K3 (Yang et al., 2026d; Team et al., 2026; DeepSeek-AI, 2026; Kimi Team, 2026). They present the recipe as an engineering choice; we provide a rationale for it (§6).

## 3 EXPERIMENTAL SETUP AND COMPUTE ACCOUNTING

**Comparison and compute accounting.** We compare held-out task performance against cumulative training FLOPs in our analysis. The main external-teacher comparisons match student initialization, prompts, rollout group size, and shared hyperparameters. Let $N_s$ and $N_t$ indicate the number of student and teacher parameters used in per-token matrix multiplications. GRPO costs $10N_s$ per generated token: $2N_s$ for generation, $2N_s$ for log-prob recomputation, and $6N_s$ for the update. OPD adds one frozen teacher forward pass with $2N_t$ cost. Each run pays for its generated tokens, excluding shared prompt context (Appendix B.1).

**GRPO and OPD Training Objectives.** Let $\pi_\theta$ be the student policy, $\pi_T$ a frozen teacher, and, for a prompt $q$, let $\{o_i\}_{i=1}^G$ be a group of $G$ student rollouts with binary outcome rewards $\{r_i\}_{i=1}^G \in \{0,1\}$. We denote $o_{i,t}$ for the $t$-th token of the $i$-th rollout and $q_{i,<t}$ for its prefix. GRPO (Group Relative Policy Optimization) maximizes the mean-subtracted advantage of the outcome reward, following Liu et al. (2025):

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\,\mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i\sum_{t=1}^{|o_i|}\log\pi_\theta(o_{i,t}\mid q, o_{i,<t})\right], \quad \hat{A}_i = r_i - \text{mean}(\{r_j\}).$$

Our OPD objective, following Li et al. (2026), is the closed-form reverse KL against $\pi_T$ over the student's top-$k$ support, computed on the student's own rollouts. At each position $(i,t)$ we pick the

student's top-$k$ token set $\mathcal{S}_k(o_{i,t}) \subset \mathcal{V}$ ($k = 16$), read the student and teacher probabilities on those $k$ tokens, and renormalize each to a distribution over $\mathcal{S}_k(o_{i,t})$:

$$\tilde{\pi}_\theta(v \mid o_{i,<t}) = \frac{\pi_\theta(v \mid o_{i,<t})}{\sum_{u \in \mathcal{S}_k} \pi_\theta(u \mid o_{i,<t})}, \quad \tilde{\pi}_T(v \mid o_{i,<t}) = \frac{\pi_T(v \mid o_{i,<t})}{\sum_{u \in \mathcal{S}_k} \pi_T(u \mid o_{i,<t})}, \quad v \in \mathcal{S}_k(o_{i,t}).$$

The loss is the closed-form reverse KL between the renormalized distributions:

$$\mathcal{L}_{\text{OPD}}(\theta) = \mathbb{E}_{q,\{o_i\}} \left[ \frac{1}{G} \sum_{i=1}^{G} \sum_{t=1}^{|o_i|} \sum_{v \in \mathcal{S}_k(o_{i,t})} \tilde{\pi}_\theta(v \mid o_{i,<t}) \Big[ \log \tilde{\pi}_\theta(v \mid o_{i,<t}) - \log \tilde{\pi}_T(v \mid o_{i,<t}) \Big] \right].$$

The outcome reward $r_i$ does *not* enter $\mathcal{L}_{\text{OPD}}$. OPD adds one frozen-teacher forward pass over every token it trains on; GRPO adds nothing. Our training hyperparameters are close to the ones in Li et al. (2026) and are detailed in Appendix B.3.

**Models and teachers** Students are Qwen3 models of 1.7B, 4B, and 8B parameters (Yang et al., 2025), mostly base checkpoints (e.g. `Qwen3-4B-Base`) with selected SFT cold starts for ablation. Teachers include external post-trained thinking models (`Qwen3-{4B, 8B, 14B, 32B}`), selected external models after further GRPO, and GRPO-trained base checkpoints at or above the student's size. We call the student's own GRPO checkpoint a self-GRPO teacher; multi-domain runs use one such checkpoint per domain.

**Data.** For *math* we train on DAPO-Math-17K (Yu et al., 2025) and DeepMath-103k (He et al., 2025). For *code* we train on LiveCodeBench v1–v5 (Jain et al., 2025) and a high-quality subset of CodeForces `open-r1/codeforces`. For the multi-domain experiments (§6) we add *MMLU-Pro* (knowledge) and *IFBench-RLVR* (Qin et al., 2024) (instruction following; rule-based constraint reward). Dataset construction details are in Appendices B.6–B.8.

**SFT cold start.** Our cold-started students are full fine-tunes of `Qwen3-{1.7B, 4B}-Base` on OpenThoughts-3 (Guha et al., 2025), following the SFT recipe from Thinking Machines Lab (Thinking Machines Lab, 2025). We sweep the SFT learning rate to pick the initialization we then post-train from (Tables 7a and 7b, Appendix B.5); full hypers are in Table 6.

**Evaluation.** All evaluations report `mean@`$k$ ($k$=16): the mean accuracy over $k$ independent samples per prompt. Math is evaluated on an out-of-distribution (*OOD*) macro, the unweighted average over AIME24 (30 prompts), AIME25 (30), and AMC23 (83), all held out from the math training data. Code is evaluated on LiveCodeBench v6 (Jain et al., 2025) (175 prompts). For the multi-domain experiments (§6), knowledge is evaluated on a 512-prompt held-out subset of MMLU-Pro (Wang et al., 2024), and instruction following on IFBench (Qin et al., 2024). Generation hyperparameters and length-cap conventions are in Appendix B.2.

## 4 External teachers are compute-inefficient for OPD

Our initial observation on DAPO-Math-17K contradicts the compute-efficiency claim. Training a `Qwen3-1.7B-Base` student against external `Qwen3-{4B, 8B, 32B}` teachers at matched settings, GRPO sits above every OPD run at every evaluated FLOP count on both the DAPO validation split and the OOD macro (Figure 7); larger teachers only increase the FLOPs of the OPD curve without lifting the student's performance.

### 4.1 Dataset surgery

One reading of the negative result is that the dataset may contain instances that the model can solve well and thus favour GRPO unintentionally. We therefore stratify DAPO-Math-17K by the joint difficulty of the student and a teacher model. This helps us understand how much OPD really provides a learning signal.

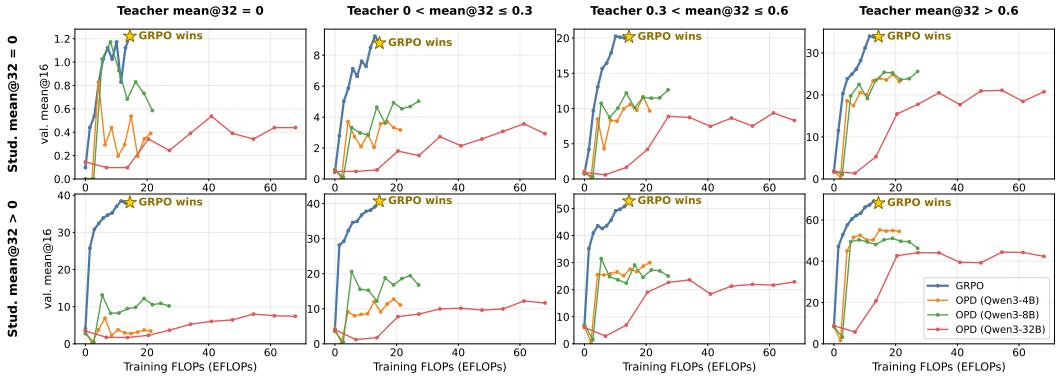



Figure 2: **GRPO vs OPD across difficulty levels.** We stratify DAPO-Math-17K by `mean@32` solve rate of the student (`Qwen3-1.7B-Base`) and teacher (`Qwen3-4B`). This yields eight subsets, each corresponding to a performance level by student and teacher; each subplot shows in-domain `mean@16` against cumulative training FLOPs. GRPO outperforms every OPD variant in every cell, including the cell most favourable to OPD (top-right: student barely solves, teacher performs very well). OOD results in Figure 8.

**Stratify the dataset by the performances of student and teacher.** We score every prompt in DAPO-Math-17K with `mean@32` under both the Qwen3-1.7B-Base student and the Qwen3-4B thinking teacher, then bin along both axes ($=0$ and $>0$ for the student; $=0$, $(0, 0.3]$, $(0.3, 0.6]$, $>0.6$ for the teacher), yielding eight cells with matched train- and validation-set sizes (Appendix D.1); OOD evaluation is kept as the AIME24/25 + AMC23 macro. We retrain GRPO and OPD in each subset, and evaluate in-domain and OOD separately. Intuitively, the cell most favourable to OPD on paper is student $= 0$, teacher $> 0.6$: GRPO's outcome reward is very weak, so the dense signal should matter most.

To our surprise, such intuition falls apart. Figure 2 shows the in-domain result: GRPO's final accuracy exceeds every OPD run's peak in every cell at a smaller FLOP budget throughout. Although an OPD run sometimes briefly reaches GRPO's early-training performance at lower compute early in the training, the advantage of GRPO over OPD widens as we use stronger and larger teachers. The OOD macro tells a slightly different but counter-intuitive story (Figure 8): OPD-4B and OPD-8B outperform GRPO in the three cells where the teacher itself performs badly (teacher `mean@32`$= 0$ or $(0, 0.3]$), the opposite of what the dense-signal motivation predicts; and GRPO wins five of the eight cells. Moreover, OPDs' OOD macro accuracy is nearly constant at $15\%$–$17.5\%$ across all eight regimes regardless of the training subset, while GRPO's varies with the data (Figure 8). We hypothesize an *extrapolation* component to OPD's learning (with external teacher) that is not tied to the training distribution; the observation is explored further in Appendix F. Repeating the same comparison on a Qwen3-1.7B-Base student with thinking mode on yields an even stronger pattern: GRPO wins over every OPD run on In-Domain and OOD performance (Appendix D.1).

### 4.2 GRPO'S COMPUTE EFFICIENCY ADVANTAGE IS ROBUST ACROSS ABLATIONS

The obvious objection to §4 is that we found a bad configuration; this section is our attempt to find a good one *among external teachers*, measured on the compute axis. We varied nine factors one at a time against the experiments in Figure 1. Broadly in every case GRPO dominates every external-teacher OPD run. Per-axis setup and results are deferred to Appendix D. In this section, we briefly discuss some highlights.

**Student initialization.** One hypothesis is that the `1.7B-Base` might be too weak to capture signal from the teacher. We therefore supervised-finetune `1.7B-Base` on OpenThoughts-3. With thinking mode off, GRPO still outperforms every OPD run (Figure 14). We conducted a learning rate tuning to obtain the best performing checkpoint for `1.7B-Base` and `Qwen3-4B-Base` and we turn on the thinking mode. However, a good student initialization (compared with a sub-optimal one) doesn't change the ordering (Figure 15). With SFT-ed `4B-Base` as the student, we also experiment

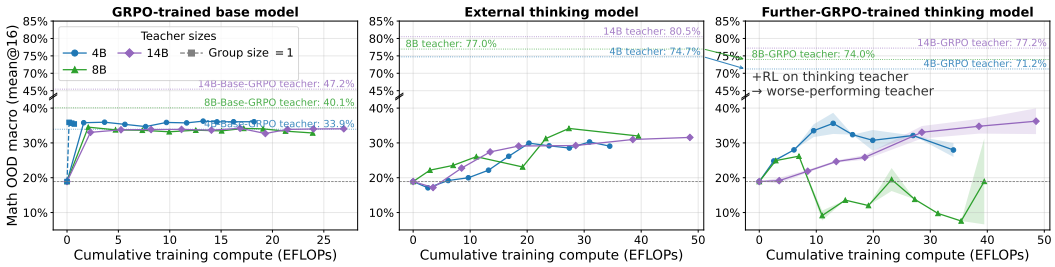



Figure 3: **Impact of running GRPO on teachers.** `Qwen3-4B-Base` student on DAPO-Math-17K; $x$ is cumulative training FLOPs, $y$ is Math OOD macro `mean@16`; dotted horizontals mark each teacher's own accuracy. We sweep teachers at 4B, 8B, and 14B across three types: GRPO-trained base (left), external thinking (centre), and GRPO on external thinking (right). **Left:** all three GRPO-trained base teachers quickly converge the student to similar performance; only the same-size 4B teacher is fully recovered, despite larger teacher having a higher performance ceiling; **Centre:** student performances climb much more slowly to the same level with GRPO-trained base; all students stay well below any teacher throughout training. **Right:** OPD with a further-GRPO-trained thinking teacher is more unstable than with a thinking teacher.

OPD with `QwQ-32B` — the generator of the OpenThoughts-1.2M data, OPD still underperforms GRPO (Figure 16). See more details in Appendix D.6.

**Training domain.** To understand whether our observation is dataset specific, we train `4B-Base` student on two code datasets (LiveCodeBench v1–v5 and a high-quality CodeForces set) with GRPO and OPD under two external teachers (8B and 32B). The ordering observed on math still holds on code: on LiveCodeBench v1–v5, GRPO reaches 46.4% LCBv6 accuracy at 0.6 EFLOP, whereas OPD-32B peaks at only 28.7% at 68.7 EFLOP and OPD-8B peaks at 36.7% at 45.7 EFLOP. Results on CodeForces are similar (Figure 19 left column). See more details in Appendix D.8.

**Additional ablations.** We also sweep the learning rate (Appendix D.2), batch size (Appendix D.3), reverse-KL estimator (closed-form top-$k$ vs sampled-token, Appendix D.4), the thinking mode (no-think vs think, Appendix D.5), the student model sizes (Appendix D.7), and the RL framework (`verl` and `tinker`, Appendix D.9). None changes the ordering. Across the ablations in this section, using an external teacher, even one from the same model family, for vanilla reverse-KL OPD is broadly less compute-efficient than running GRPO directly on the same student. In Appendix E, we discuss a prevailing response-length inflation behavior from OPD-trained student. In Appendix I, we also experimented with a reduced rollout group size.

In the next section (§5), following recent work that GRPO-trains the teacher before distilling, we examine whether that change to the external teacher alters what we observe here.

## 5 Efficient OPD only when the teacher is the student's own RL checkpoint

A highly cited work reports that GRPO-training the teacher before distillation can help OPD (Li et al., 2026), but it establishes this claim on a limited experiment: running GRPO on Qwen3-4B-Base leads to a better teacher than Qwen3-4B (Non-thinking) when the student is Qwen3-1.7B-Base. We vary the teacher along three dimensions: 1) initialization — a base model or an external thinking model, 2) whether it is used as-is or first GRPO-trained on the task, and 3) its size — same or larger.

**Setup.** We fix the student to `Qwen3-4B-Base` on DAPO-Math-17K and sweep the teacher across three types: (i) *GRPO-trained base* model: a base model (after pre-training only, e.g., `Qwen3-8B-Base`) is trained with GRPO on DAPO-Math-17K; (ii) *external thinking* model: the off-the-shelf post-trained Qwen3 checkpoints as used in §4.1 (e.g., `Qwen3-8B`); (iii) *further-GRPO-trained thinking* model: the external thinking model is further trained with GRPO on DAPO-Math-17K. All experiments in this section use thinking mode. Families (i) and (iii) share the same GRPO

recipe; the only difference is the response-length cap (7168 for base, 16 384 for thinking; full hyperparameters in Table 2). For each type, we cover the teacher sizes at 4B, 8B, and 14B. Every OPD run shares the same recipe; only the teacher varies.

### 5.1 Student's GRPO checkpoint is consistently the most efficient teacher

We present our results in Figure 3. With GRPO-trained base teachers, the student converges to similar performance by the first evaluated checkpoint ($\leq$ 2.5 EFLOPs) — the 4B, 8B, and 14B teachers reach 35.8%, 34.5%, and 33.0% — and each stays in a narrow band throughout training. The same-size 4B teacher fully recovers its own GRPO accuracy (33.9%, which GRPO reaches at 10.1 EFLOP) by 1.62 EFLOP, an $\sim 6\times$ compute reduction; with $G{=}1$ the same recovery costs only 0.2 EFLOP ($\sim 50\times$), matching the efficiency gains reported in the OPD literature (Thinking Machines Lab, 2025). Larger teachers are individually stronger, but their students plateau at an even slightly lower level — the extra capacity does not transfer. External thinking teachers improve the student much more slowly: the 8B teacher peaks at 34.2% (27.2 EFLOP), while the 4B and 14B teachers stay below 32%, all well below their teacher accuracy. Further-GRPO-trained thinking teachers give the most unstable curves: the 4B teacher's student peaks at 38.7% (step 80), near the best direct-GRPO accuracy (38.8%), then collapses to 26.0% by step 200; the 8B teacher's student falls below initialization to 6.6%; we repeat each run to confirm the instability. Only using GRPO-trained student as teacher avoids the pathological response-length inflation that dominates the other two families (Figure 23, Appendix E); Appendix H discusses a potential explanation via changes in the spectra of model weight during optimization.

**More OPD experiments with self-GRPO (the student's GRPO checkpoint) teacher** On math, using `Qwen3-4B-Base` as student and its GRPO checkpoint as teacher gave both compute-efficient improvement and full teacher recovery. In Appendix G we test this beyond that setting by (i) enlarging the student to `Qwen3-8B-Base`, (ii) switching the domain from math to code, and (iii) replacing the `Qwen3-*-Base` student (raw pretrained) with a post-trained `Olmo-3-7B-Instruct-SFT`. All three retain high compute-efficiency and full teacher recovery.

**An intriguing side observation in self-GRPO OPD** Across our self-GRPO experiments, the student's best accuracy often exceeds the teacher's: in Figure 3 the student reaches 35.8% versus the teacher's 33.9%. The most striking case is `Olmo-3-7B-Instruct-SFT` on math, where the self-GRPO teacher scores 42.0% but the student reaches 51% at 1.25 EFLOP (Figure 27)—a gap too large to attribute to randomness. Its cause is unknown and left to future work.

### 5.2 What predicts full teacher recovery before any training?

Section 5.1 identifies setups that are compute-efficient — using student's GRPO checkpoint as teacher, and setups that are not — external thinking and further-GRPO-trained thinking teachers. We ask whether this outcome can be predicted *before* the distillation run itself, using the initial top-$k$ token *overlap* between student and teacher proposed by Li et al. (2026) as a diagnostic, and scoring each run's outcome by its *normalized minimal teacher–student gap*.

**Overlap.** On the first training batch, we score each position with the fraction of the student's top-16 token IDs that also appear in the teacher's top-16, and average over all valid positions.

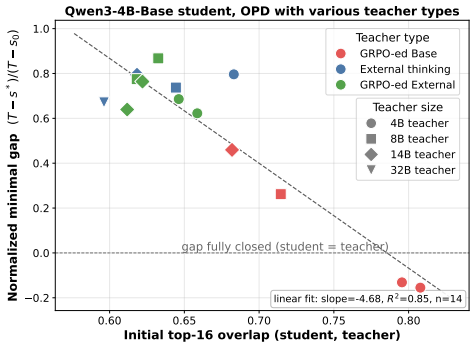



Figure 4: **Initial top-16 overlap highly correlates with the normalized minimal teacher–student gap**.

**Normalized minimal teacher–student gap.** To summarize how close each OPD run brings the student to its teacher, we look at the gap between teacher accuracy $T$ and the best student accuracy

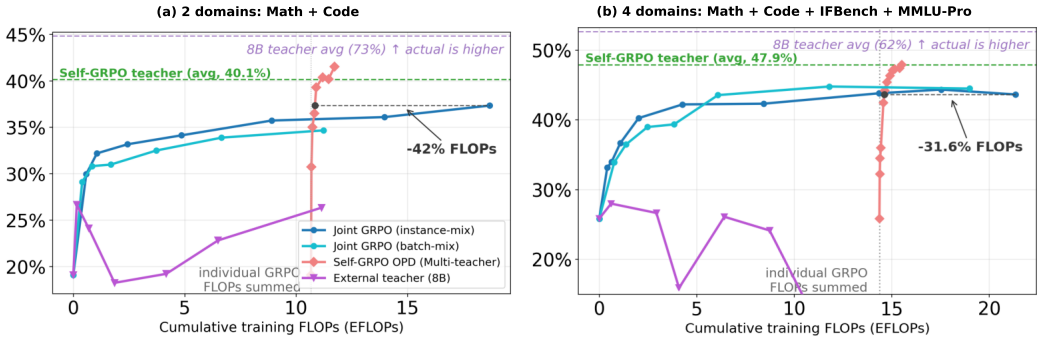



Figure 5: **Multi-teacher OPD beats joint GRPO on domain-averaged performance.** Qwen3-4B-Base student. **Left:** 2-domain average (math + code). **Right:** 4-domain average (math, code, instruction following, knowledge). Joint GRPO variants train on all domains simultaneously; multi-teacher OPD merges per-domain GRPO experts into a single student. In two plots, the cost of Self-GRPO OPD with multiple teachers starts from the total FLOPs used to obtain all self-GRPO teachers. A single external 8B-thinking teacher (purple) can improve over initial student on 2-domain setting, but collapse on 4-domain setting. Per-domain breakdowns are in Appendix J.

$s^*$ reached at any point during training, and divide it by the initial teacher–student accuracy gap at step zero $T - s_0$.

**Overlap highly correlates with the teacher–student gap.** Across all OPD runs in §5.1 and relevant runs in Figure 17, initial top-16 overlap correlates with a lower normalized minimal teacher–student gap (Figure 4). However, we want to note that there's a large gap between GRPO-trained `Qwen3-8B-Base` teacher (red square) and GRPO-trained `Qwen3-4B-Base` (red dot). With data point missing in between, our observation serves as a side analysis in a limited setting (i.e. student is fixed to be `Qwen3-4B-Base`) rather than a generalizable rule.

**Summary.** Across nine ablations of reverse-KL OPD, no external teacher transfers compute-efficiently and the student broadly loses to GRPO (§4, §5.1). Only the *self-GRPO teacher* — the student's own GRPO checkpoint — closes the gap with high compute-efficiency.

## 6 Multi-teacher OPD: A natural use case of strong-to-weak OPD

The finding of §5 remains a paradox in a single-task setting: *why distill when you already have the GRPO-trained model?* Across multiple tasks, however, the paradox can become an opportunity. Consider math and code: separate GRPO runs produce two specialists, but neither checkpoint can serve both domains. This is where multi-teacher OPD has a distinct compute-efficiency opportunity: it can consolidate the GRPO specialists into a single multi-domain model that neither specialist provides, with only a small additional distillation cost. *Would the cost of training those experts and consolidating them via distillation beat that of joint GRPO on the same data mix?* Recent open-source models (Yang et al., 2026d; Team et al., 2026; DeepSeek-AI, 2026; Kimi Team, 2026) use this recipe; we set out to understand the rationale behind it.

**Setup.** We consider 4 domains in total: math on DAPO-Math-17K, code on LiveCodeBench-v1–v5, instruction following on IFBench, and knowledge on MMLU-Pro. We experiment with 2-domain setting and 4-domain setting separately. From Qwen3-4B-Base we train one GRPO expert on each single domain, each with its own hyperparameters (Table 11). We build a data mix whose per-domain training instance count matches each individual expert, and we compare four methods on it: (i) *Joint GRPO (instance-mix)*, which fills every batch with instances drawn proportionally from all domains at one shared batch size; (ii) *Joint GRPO (batch-mix)*, which draws each batch from each domain (each domain has its own batch size, notably a smaller batch on code domain) per step and interleaves domains across steps; (iii) *Multi-teacher self-GRPO OPD*, which routes each instance to the self-GRPO teacher of the corresponding domain and computes the reverse KL; (iv) *External teacher OPD* with a single 8B-thinking teacher, the best-performing external teacher based on results from §4 and §5. All OPD runs use rollout group size $G{=}1$ unless otherwise noted.

We run each method on 2-domain (math + code) and 4-domain mixes to show some scaling behavior; see Appendix J for more details.

**Results.** Figure 5 shows that the single-domain success of self-GRPO OPD transfers to the multi-domain setting. On the 2-domain mix the two specialists cost $10.7\,\mathrm{EFLOP}$, and multi-teacher self-GRPO OPD reaches their ceiling in an additional $0.55\,\mathrm{EFLOP}$ ($\sim 5.2\%$); on the 4-domain mix the four specialists cost $14.4\,\mathrm{EFLOP}$, performance recovered in an additional $1.15\,\mathrm{EFLOP}$ ($\sim 8.0\%$). Single-domain and multi-teacher self-GRPO OPD are on par in compute-efficiency (Figures 31, 32). Even after accounting for the OPD and self-GRPO-teacher costs, both joint-GRPO variants (instance-mix and batch-mix) underperform on both multi-domain settings; specifically, multi-teacher self-GRPO OPD matches Joint GRPO(instance-mix) with $42.0\%$ fewer FLOPs on 2 domains and $31.6\%$ fewer FLOPs on 4 domains. *External-teacher OPD is worst*: on 2 domains it reaches $26.7\%$, drops, then finishes at a similar level; scaling to 4 domains, the student collapses on every domain ($0\%$ math, $0\%$ code, $7\%$ MMLU-Pro, $18\%$ IFBench) despite the 8B-thinking teacher being the strongest model overall, with severe response-length inflation at inference. The group-size ablation (Appendix J) shows $G{=}1$ reaches the specialist ceiling at substantially lower compute than $G{=}8$ (Section 5.1).

**Summary.** Multi-teacher OPD (route each instance to its domain's specialist to compute the reverse KL) turns the single-task paradox of §5 into a recipe: merge per-domain GRPO experts into one student, reach the specialists' ceiling for an extra $5.2\%$–$8.0\%$ of the FLOPs spent training them, and beat both joint GRPO and external-teacher OPD at matched compute.

**Benefit of multi-teacher OPD with self-GRPO teachers.** Mulit-teacher pipeline enjoys three benefits at once. *Better overall performance:* as we show in Figure 5, using multi-teacher OPD with self-GRPO teachers leads to better performance than joint GRPO. *Modular development:* different domains naturally require different RL infrastructure (code sandboxes, tool-use environments, symbolic verifiers) and different optimal hyperparameters (learning rate, rollout length, response-length cap); training each expert independently lets each team use its own stack and tune at its own schedule to get the best expert on each domain. *Parallelizable specialist training:* per-domain runs can naturally happen concurrently rather than sharing a single joint schedule.

**Relation to frontier practice.** Several recent frontier-model reports use multi-teacher OPD to combine domain-expert teachers into a shared student, including Nemotron-Cascade 2 (Yang et al., 2026d), DeepSeek-V4 (DeepSeek-AI, 2026), and Kimi K3 (Kimi Team, 2026). These reports describe the recipe as an engineering choice without matched-compute measurement. Our results provide the rationale behind their decision: Section 5 shows that OPD succeeds when the teacher is a GRPO checkpoint of the student, and Section 6 shows that merging per-domain experts trained under this condition via OPD outperforms training all domains jointly with GRPO or using an external teacher.

## 7 Conclusion

We give the first systematic, FLOPs-based characterization of strong-to-weak OPD against GRPO, and identify when OPD is a promising post-training tool. First, we find external teachers under vanilla reverse-KL broadly lose to GRPO across our nine-axis ablation (§4); and we widely observe pathological behavior from the trained student with external teachers. Second, through carefully studying the impact of GRPO on teachers, we realized the *self-GRPO teacher* (the student's own GRPO checkpoint) is the only setup where the student consistently learns compute-efficiently and fully recovers teacher performance (§5). It matches GRPO's performance at $50.0\times$ fewer FLOPs, replicated across domains, student sizes, and student families; the student's best performance can even substantially exceed the teacher's, calling for further study. Third, multi-teacher OPD with self-GRPO teachers is a cheap consolidation step (§6). It recovers every specialist for an extra $5.2\%$ (2 domains) or $8.0\%$ (4 domains) of specialist training FLOPs — cheaper than joint GRPO on the same data; and substantially cheaper than external-teacher OPD, which underperforms at 2 domains and collapses at 4. Together, these results suggest OPD is not a straightforward replacement for GRPO, but it is an efficient way to fold many same-origin GRPO-trained specialists into a single generalist student.

### AI USE STATEMENT

We used generative AI tools (large language models) to assist with drafting and editing portions of the manuscript text. All AI-assisted text was reviewed, revised, and verified by the authors. We did not use generative AI tools (e.g., image generation) to generate experimental results, design the method or produce experimental figures. LLM-generated code snippets used during development were reviewed and tested for correctness. We take responsibility for the final content of this work, including all text, claims, and artifacts.

### LIMITATION AND FUTURE WORK.

Our study focuses on the vanilla reverse-KL OPD objective with a closed-form top-$k$ estimator (§3). Recent work extends this baseline along axes we do not sweep at matched FLOPs: hybrid objectives that blend OPD with an outcome-reward term (Ko et al., 2026; Tan et al., 2026; Liu et al., 2026), alternative divergences (Luo et al., 2026; Chen et al., 2026), trajectory reshaping and per-token reweighting (Li et al., 2026; Rang et al., 2026), and async or speculative rollout pipelines (Feng et al., 2026; Ren et al., 2026). Any of these may narrow the FLOPs gap we identify. We encourage future work claiming an OPD advantage over GRPO to report training FLOPs as an implementation-invariant measure, so that improvements in objective, rollout scheduling, and infrastructure can be disentangled and directly comparable across papers.

### ETHICS STATEMENT

This work trains and evaluates open-weight language models on public mathematics and competitive-programming datasets; it involves no human subjects, no personal data, and no new data collection. We do not foresee direct negative societal impacts from the distillation methodology itself.

### REPRODUCIBILITY STATEMENT

The training and evaluation recipe for every experiment in this paper is documented in Section 3 and expanded in Appendix B.3. Shared hyperparameters common to all runs are listed in Table 2; the GRPO- and OPD-specific settings are in Tables 3 and 4, and the per-ablation overrides for the sweeps in Section 4.2 are in Table 5. All students, teachers, training corpora, and evaluation benchmarks are publicly released Qwen3 checkpoints and open datasets identified by version and citation in Section 3 and Appendices B.6–B.8. The framework versions and rollout engine are named in Section 3 (verl `0.7.0.dev` and `0.8.0.dev`, SGLang `0.5.2`), and the framework replication of Appendix D.9 additionally names the tinker library and its LoRA rank, learning rate, batch size, rollout count, and response cap for the matched 9B replication. We also release the prompts used for formating training instances in Appendix C. All models, datasets, and framework releases used in this work are public.

## REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.13649. Generalized knowledge distillation; the on-policy distillation formulation this work evaluates descends from it.

Xianwei Chen, Shimin Zhang, and Jibin Wu. f-opd: Stabilizing long-horizon on-policy distillation with freshness-aware control, 2026. arXiv:2605.17862.

DeepSeek-AI. Deepseek-v4. `https://huggingface.co/deepseek-ai`, 2026. Open-weights model release (V4-Flash and V4-Pro); no arXiv technical report at time of writing.

Tianyu Ding, Jianhong Xin, and Juan Pablo De la Cruz Weinstein. Keep policy gradient in charge: Sibling-guided credit distillation for long-horizon tool-use agents, 2026. arXiv:2606.12634.

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, 2026. arXiv:2607.05394.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes, 2026a. arXiv:2603.25562.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking on-policy distillation of large language models ii: One training example, 2026b. URL https://arxiv.org/abs/2609.04172.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.08543. Reverse-KL distillation on student-generated text.

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. *arXiv preprint arXiv:2506.04178*, 2025.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations (ICLR)*, 2025.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms?, 2026. arXiv:2603.24472.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. arXiv:1606.07947.

Kimi Team. Kimi k3: Open frontier intelligence, 2026. arXiv:2607.24653.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation, 2026. arXiv:2603.11137.

Byung-Kwan Lee, Ximing Lu, Shizhe Diao, Minki Kang, Saurav Muralidharan, Karan Sapra, Andrew Tao, Pavlo Molchanov, Yejin Choi, Yu-Chiang Frank Wang, and Ryo Hachiuma. Zone of proximal policy optimization: Teacher in prompts, not gradients, 2026. arXiv:2606.18216.

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

Xinyu Liu, Kechen Jiao, Chunyang Xiao, Runsong Zhao, Junhao Ruan, Bei Li, Jiahao Liu, Qifan Wang, Xin Chen, Jingang Wang, Chenglong Wang, Tong Xiao, and JingBo Zhu. Teacher-guided policy optimization for on-policy reasoning distillation under large policy divergence, 2026. arXiv:2605.13230.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Second Conference on Language Modeling*, 2025.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying opd: Length inflation and stabilization strategies for large language models, 2026. arXiv:2604.08527.

Andrei Liviu Nicolicioiu, Mohammad Pezeshki, and Aaron Courville. On-policy self-distillation with sampled demonstrations reduces output diversity, 2026. arXiv:2606.26091.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*, 2024.

Miao Rang, Zhenni Bi, Hang Zhou, Kai Han, Xuechun Wang, An Xiao, Xinghao Chen, Yunhe Wang, and Hanting Chen. Near-policy: Accelerating on-policy distillation via asynchronous generation and selective packing, 2026. arXiv:2605.05940.

Haolin Ren, Ziyang Huang, Chenhao Yuan, Jun Zhao, and Kang Liu. Trace-based on-policy distillation for masked diffusion language models, 2026. arXiv:2607.16872.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys)*, 2025. The verl implementation used in this work.

Qitai Tan, Zefang Zong, Mo Li, Yipeng Shi, Yang Li, and Peng Chen. Atod: Annealed turn-aware on-policy distillation for multi-turn agentic tasks, 2026. arXiv:2606.27814.

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Lei Li, Liang Zhao, Linghao Zhang, Peidian Li, Qianli Chen, Shaohui Liu, Shihua Yu, Shijie Cao, Shimao Chen, Shouqiu Yu, Shuo Liu, Tianling Zhou, Weijiang Su, Weikun Wang, Wenhan Ma, Xiangwei Deng, Bohan Mao, Bowen Ye, Can Cai, Chenghua Wang, Chengxuan Zhu, Chong Ma, Chun Chen, Chunan Li, Dawei Zhu, Deshan Xiao, Dong Zhang, Duo Zhang, Fangyue Liu, Feiyu Yang, Fengyuan Shi, Guoan Wang, Hao Tian, Hao Wu, Heng Qu, Hongfei Yi, Hongxu An, Hongyi Guan, Xing Zhang, Yifan Song, Yihan Yan, Yihao Zhao, Yingchun Lai, Yizhao Gao, Yu Cheng, Yuanyuan Tian, Yudong Wang, Zhen Tang, Zhengju Tang, Zhengtao Wen, Zhichao Song, Zhixian Zheng, Zihan Jiang, Jian Wen, Jiarui Sun, Jiawei Li, Jinlong Xue, Jun Xia, Kai Fang, Menghang Zhu, Nuo Chen, Qian Tu, Qihao Zhang, Qiying Wang, Rang Li, Rui Ma, Shaolei Zhang, Shengfan Wang, Shicheng Li, Shuhao Gu, Shuhuai Ren, Sirui Deng, Tao Guo, Tianyang Lu, Weiji Zhuang, Weikang Zhang, Weimin Xiong, Wenshan Huang, Wenyu Yang, Xin Zhang, Xing Yong, Xu Wang, Xueyang Xie, Yilin Jiang, Yixin Yang, Yongzhe He, Yu Tu, Yuanliang Dong, Yuchen Liu, Yue Ma, Yue Yu, Yuxing Xiang, Zhaojun Huang, Zhenru Lin, Zhipeng Xu, Zhiyang Chen, Zhonghua Deng, Zihan Zhang, and Zihao Yue. Mimo-v2-flash technical report, 2026. URL https://arxiv.org/abs/2601.02780.

Thinking Machines Lab. On-policy distillation. https://thinkingmachines.ai/blog/on-policy-distillation/, 2025. Non-peer-reviewed web report. Source of the "7-10× fewer gradient steps ⇒ 50-100× compute efficiency" claim this paper tests; it reports no method-separated GPU-hours or wall clock.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. Vicur: Visual cues as recoverable privilege for multimodal on-policy distillation, 2026. URL https://arxiv.org/abs/2606.05718.

Chen Wang, Zhaochun Li, Jionghao Bai, Yining Zhang, Hexuan Deng, Ge Lan, and Yue Wang. Distilled reinforcement learning for llm post-training, 2026a. arXiv:2607.17247.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, and Honggang Qi. Skill-sd: Skill-conditioned self-distillation for multi-turn llm agents, 2026b. arXiv:2604.10674.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. Tcod: Exploring temporal curriculum in on-policy distillation for multi-turn autonomous agents, 2026c. arXiv:2604.24005.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuanhan Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

Haoran Xin, Anhao Zhao, Ying Sun, Jin Li, Xiaoyu Shen, and Hui Xiong. Escaping the kl agreement trap in on-policy distillation, 2026. arXiv:2606.09471.

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.

Fan Yang, Rui Meng, and Yuxin Wen. Why does feedback-augmented self-distillation fail to improve retrieval-interleaved search agents?, 2026a. arXiv:2607.17558.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, and Gang Chen. Oprd: On-policy representation distillation, 2026b. arXiv:2606.06021.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-opd: Efficient and reliable on-policy distillation for long-horizon reasoning, 2026c. arXiv:2605.07804.

Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, Renjie Pi, Grace Lam, Nayeon Lee, Alexander Bukharin, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nemotron-cascade 2: Post-training llms with cascade rl and multi-domain on-policy distillation, 2026d. arXiv:2603.19220.

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

Yulin Zhang, He Cao, Zihao Jiang, Chenyi Zi, Zhipeng Zhou, Zijing Liu, Yu Li, Jia Li, and Ziqi Gao. Proteinopd: Towards effective and efficient preference alignment for protein design, 2026. arXiv:2605.10189.

Anhao Zhao, Haoran Xin, Yingqi Fan, Junlong Tong, Wenjie Li, and Xiaoyu Shen. Decoupling kl and trajectories: A unified perspective for sft, dagger, offline rl, and opd in llm distillation, 2026. URL https://arxiv.org/abs/2605.16826.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. Scope: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting, 2026. arXiv:2604.10688; runs $G{=}8$ GRPO vs $G{=}1$ OPD (mismatched rollout group size).

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hanqing Zhu, Wenyan Cong, Zhizhou Sha, Sagnik Mukherjee, Xinyuan Song, David González-Martínez, Xiaoxia Wu, Yuandong Tian, Shiwei Liu, David Z. Pan, and Zhangyang Wang. Iso: An rlvr-native optimization stack. *arXiv preprint arXiv:2607.19331*, 2026a.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes, 2026b. arXiv:2605.11182.

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

# A LITERATURE SURVEY: HOW WE FILTER TO ZERO FLOPS-BASED WINS

**Corpus.** We take the deduplicated union of three public OPD bibliographies: `chrisliu298/awesome-on-policy-distillation`, `thinkwee/awesomeopd`, and `lingyunyang-opd`. Identity is resolved by arXiv ID, DOI, and normalized-title merge. Restricted to papers whose first arXiv version was posted between 2025-10-06 and 2026-08-06, the corpus is 406 unique preprints.

**Reading.** Every paper is read end-to-end by `Claude Opus 4.8`, which returns a schema-enforced structured record; every classification-bearing field is required to be backed by a verbatim quotation with a page locator copied from the PDF. Papers clearing the coarse rule (a distinct teacher plus any GRPO mention) are then re-read by a second, independent Claude Opus 4.8 invocation that does not see the first pass's verdict. No keyword-only filters make classification calls; every field is grounded in PDF-verified quotation.

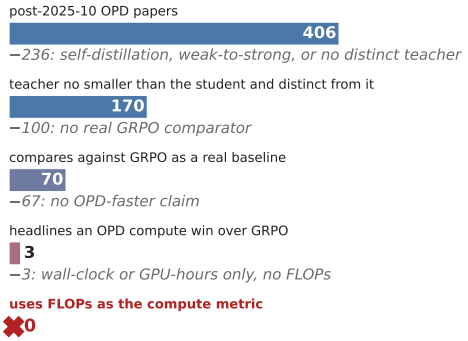



Figure 6: **Filtering of the 406-paper corpus to zero FLOPs-based OPD compute wins over GRPO.** Each bar is the surviving count after applying the criterion labelled above it; the italic drop text below each bar names what the next filter removes.

**Filter criteria.** Starting from 406 papers, we apply four criteria in strict precedence. Each row is the surviving count after the criterion is applied.

- Corpus of post-2025-10 OPD preprints: **406**.
- Teacher no smaller than the student and distinct from the student: the teacher is a distinct model that is larger than or the same size as the student. Excludes self-distillation and weak-to-strong (smaller teacher) setups. **170** (drops 236: 146 self-distillation, 2 smaller (weak-to-strong), 8 unclear or not-applicable, plus 80 papers with no distinct teacher or no resolvable teacher role).
- Compares against GRPO as a real baseline: canonical GRPO or a named variant (DAPO, GRPO++, REINFORCE++, or a paper-specific modification) is run with reported numbers, verified from the PDF; a text mention without a GRPO experiment does not count. **70** (drops 100: papers with an eligible teacher but no real GRPO baseline).
- Reports an OPD compute advantage over GRPO: the paper's abstract, main table, or discussion explicitly headlines an OPD wall-clock, GPU-hour, or FLOPs win over the GRPO baseline. **3** (drops 67: papers that report a compute metric but make no OPD-faster claim, or where GRPO wins). The three surviving papers are REOPOLD (Ko et al., 2026), Near-Policy Distillation (Rang et al., 2026), and $f$-OPD (Chen et al., 2026).
- Uses FLOPs as the compute metric. Wall-clock and GPU-hours entangle the training objective with implementation and hardware; FLOPs is the only implementation-invariant compute metric. **0** (drops 3). None of the three surviving headline claims reports a FLOPs comparison:
  - REOPOLD (Ko et al., 2026) reports GPU-hours, wall-clock, optimizer steps, and rollouts at a matched 300-step budget for RKL OPD vs canonical GRPO. Per-step compute differs substantially between the two methods (OPD adds one teacher forward pass per training token), so the reported wall-clock speedup does not translate into a matched-FLOPs claim.
  - Near-Policy Distillation (Rang et al., 2026) reports wall-clock and optimizer steps for its async-KD acceleration versus synchronous OPD; GRPO enters only as a downstream RL stage compared against SFT and NPD initializations, not as a matched-FLOPs OPD-vs-GRPO comparison. No FLOPs are reported.
  - $f$-OPD (Chen et al., 2026) reports throughput, resolve rate, and GPU-hours on SWE-bench for vanilla sync/async OPD versus canonical GRPO. Throughput and resolve

rate do not capture total compute, and GPU-hours are implementation-dependent; no FLOPs comparison appears.

**Conclusion.** No paper in the surveyed corpus establishes a FLOPs-based compute advantage for OPD with a same-size-or-larger teacher over GRPO. This is the gap the paper fills.

# B Experiment setup details

## B.1 Compute FLOPs Accounting.

We account for training compute with a single per-token forward/backward model, applied identically to both methods. Let $N_s$ be the number of student parameters that enter a matrix multiplication per token, and $N_t$ the corresponding count for the teacher. Charging a forward pass at 2 FLOPs per parameter per token and a combined forward and backward pass at 6, the student is used three times per token: generation ($2N_s$), log-probability recompute ($2N_s$), and the policy update ($6N_s$), for $10\,N_s$ in total. OPD additionally runs one frozen-teacher forward pass at $2\,N_t$, and GRPO runs none. The per-token cost is therefore

$$f_{\mathrm{GRPO}} = 10\,N_s, \qquad f_{\mathrm{OPD}} = 10\,N_s + 2\,N_t, \tag{1}$$

We count only generated tokens and not the shared prompt context for each run.

## B.2 Evaluation: generation hyperparameters and length caps

Table 1 lists the generation hyperparameters used for evaluation. All evaluations sample $n$=16 responses per prompt. Different domains use different response-length caps.

Table 1: Evaluation generation hyperparameters by thinking mode and domain.

| Thinking mode | Domain | Temperature | top-$p$ | top-$k$ | Response cap (tokens) |
|---|---|---|---|---|---|
| Off | Math | 0.6 | 0.95 | – | 7168 |
| On | Math | 0.6 | 0.95 | 20 | 30 720 |
| On | Code (LCBv6) | 0.6 | 0.95 | 20 | 31 744 |
| On | MMLU-Pro | 0.6 | 0.95 | 20 | 7168 |
| On | IFBench | 0.6 | 0.95 | 20 | 30 720 |

Every table and figure states its generation budget (response-token cap), because OPD rollouts saturate length caps and the budget therefore changes conclusions (§E). Where a budget must be matched post hoc we truncate responses to the token prefix and re-grade with the same verifier the run used. This is exact because generation is causal.

## B.3 Training hyperparameters

Tables 2–4 list the training-recipe values widely used in our experiments: shared hyperparameters, GRPO-specific configuration, and OPD-specific configuration. Table 5 records the appendix sections that document deviations along each ablation axis.

## B.4 Infrastructure.

Training uses verl (Sheng et al., 2025) with SGLang (Zheng et al., 2024) `0.5.2` rollout on a single 8-GPU H200 node; the two framework generations we compare are labelled *old/custom* (verl `0.7.0.dev` fork with our own OPD implementation) and *latest/native* (verl `0.8.0.dev` with native OPD implementation).

## B.5 SFT cold-start recipe and learning-rate sweep

Our cold-started students are full fine-tunes of Qwen3-{1.7B, 4B}-Base on OpenThoughts-3 (Guha et al., 2025). We follow the SFT recipe from step 1 of the Thinking Machines Lab OPD blog

Table 2: Shared training hyperparameters. Both GRPO and OPD use identical values for every setting in this table.

| Setting | Value |
|---|---|
| Framework | `verl 0.7.0.dev` (custom fork; see §D.9 for the native replication) |
| Rollout inference | SGLang `0.5.2` |
| Rollout group size $G$ | 8 rollouts per prompt |
| Batch size (prompts) | 64 |
| Response-length cap | 7168 tokens (unless noted otherwise) |
| Rollout sampling | temperature $= 1.0$, top-$p = 1.0$, top-$k$ disabled |
| Data shuffling | disabled |
| Gradient clipping | max norm $= 1.0$ |
| Optimizer | Adam ($\beta_1$=0.9, $\beta_2$=0.95, $\epsilon$=$1 \times 10^{-8}$) |
| Weight decay | 0 |
| Learning rate | $2 \times 10^{-6}$ (constant, no warmup) |
| Optimizer steps | 200 |
| Precision | bf16 rollout inference; optimizer in fp32 |
| Hardware | 8×H200, single node |
| Sharding | FSDP |
| Thinking mode | on |
| KL-to-reference term | disabled (neither method pays for a reference-policy forward pass) |

Table 3: GRPO-specific configuration.

| Setting | Value |
|---|---|
| Reward | binary outcome (0/1); math: exact match; code: pass/fail |
| KL-to-reference | disabled ($\beta_{\text{KL}}$=0; reference-policy KL term ablated, §3) |
| Policy-update epochs | 1 (single pass over each rollout batch) |
| Clip ratio | asymmetric (DAPO-style): $\epsilon_{\text{low}}$=0.2, $\epsilon_{\text{high}}$=0.28 $\Rightarrow$ [0.8, 1.28] |
| Loss aggregation | token-mean |
| Teacher forward pass | none (GRPO needs no teacher) |

post (Thinking Machines Lab, 2025) as closely as `verl` allows. Table 6 lists the hyperparameters we use. Tables 7a and 7b report the SFT learning-rate sweep we ran to pick the initialization: for both students we sweep lr $\in \{1 \times 10^{-6}, 1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}\}$ and evaluate at step 3000 under the paper's default decoding, then use the best learning rate for downstream RL and OPD runs.

The SFT learning-rate sweep for the cold-start initialization is reported in Tables 7a and 7b: Qwen3-1.7B-Base and Qwen3-4B-Base, full fine-tuned on OpenThoughts-3 for a single pass ($\approx 3000$ optimizer steps at global batch 128), evaluated on the paper's OOD macro (AIME24 + AIME25 + AMC23, `mean@16` at temperature 0.6, top-$p$ 0.95, top-$k$ 20, thinking enabled, 31 744-token cap). The *Base ref* column is the external Base model before SFT under the same decoding.

### B.6 High-quality CodeForces training set

We carve a high-quality CodeForces training set from `open-r1/codeforces` (Hugging Face). Starting from the 6594-problem bank, we keep only the 1213 problems whose `official_tests_complete` flag is true (the release carries the full official test set rather than a truncated prefix, so a wrong solution cannot pass by chance), and drop 6 problems whose graded tests are already shown as worked examples in the statement. The remaining 1207 problems split 1107 train / 100 val by a random shuffle.

### B.7 MMLU-Pro: 5-shot knowledge prompts

We use the MMLU-Pro test set (Wang et al., 2024) (11 904 multiple-choice questions across 14 disciplines) as our knowledge-domain training source. Each prompt prepends a fixed block of 5 randomly sampled in-context examples from the MMLU-Pro validation split, drawn once and shared

Table 4: OPD-specific configuration for the headline matched pair.

| Setting | Value |
|---|---|
| Teacher temperature | 1.0 |
| Support set | student top-$k$ tokens, $k$=16 (closed-form reverse KL over student's top-$k$ support) |
| Divergence | reverse KL, $\mathrm{KL}(\tilde{\pi}_\theta \,\|\, \tilde{\pi}_T)$ on renormalized top-$k$ |
| Outcome reward | not used (vanilla reverse-KL OPD; $r_i$ does not enter $\mathcal{L}_{\mathrm{OPD}}$) |
| Policy-update epochs | 1 (single pass over each rollout batch) |
| Loss aggregation | token-mean |

Table 5: Deviations from the shared recipe by ablation axis. Each row names the appendix section documenting the change.

| Ablation axis | Key deviation | Section |
|---|---|---|
| Dataset difficulty | no-think mode | §D.1 |
| Learning rate | swept $\{2 \times 10^{-6}$–$5 \times 10^{-5}\}$, no-think mode | §D.2 |
| Batch size (12800-prompt budget) | $\{64, 128, 256\}$ at $\{200, 100, 50\}$ steps, no-think mode | §D.3 |
| Reverse-KL estimator | closed-form top-$k$ vs sampled-token, no-think mode | §D.4 |
| Thinking mode | OpenThoughts-3-SFT init, no-think and think teachers | §D.5 |
| Student initialization | Qwen3-4B-Base, OpenThoughts-3-SFT variants | §D.6 |
| Model sizes | Qwen3-4B-Base and Qwen3-8B-Base students | §D.7 |
| Training domain | LCBv1–v5, CodeForces, response cap 16 384 | §D.8 |
| Framework | `verl 0.8.0.dev` with native OPD, `tinker` | §D.9 |
| Rollout size | 1 | §I |

across all prompts. We hold out 512 prompts (stratified by discipline) for evaluation and train on the remaining. The reward is binary: 1 if the model's extracted answer matches the ground-truth option, 0 otherwise.

We use the MMLU-Pro test set rather than a separate training corpus because the multi-domain experiments (§6) are designed to test whether OPD can recover multiple specialist teachers simultaneously, not to produce a competitive MMLU-Pro score.

### B.8 IFBench-RLVR: instruction following with rule-based reward

For instruction following we use IFBench-RLVR (Qin et al., 2024), which pairs natural-language instructions with programmatically verifiable constraints (e.g. word count, formatting, keyword inclusion). The training set contains 94 873 prompts. The reward is the fraction of constraints satisfied per response, thresholded to binary (1 if all constraints pass, 0 otherwise). Evaluation uses the IFBench test split.

## C Prompt templates

We list the prompt templates used for each training and evaluation dataset in this paper. Placeholders ($\{\ldots\}$) denote per-row substitutions.

**DAPO-Math-17K (math training) and AIME24/25, AMC23 (math eval)**

```
<System Prompt>
Please reason step by step, and put your final answer within \boxed{}.

<User Prompt>
{problem statement}
```

Reward: rule-based extractor over the value inside `\boxed{}`; a response is scored 1 if the extracted string matches the ground-truth answer verifier and 0 otherwise.

Table 6: SFT hyperparameters. We follow the SFT recipe from the Thinking Machines Lab OPD blog post (Thinking Machines Lab, 2025) as closely as `verl` allows.

| Hyperparameter | Value |
|---|---|
| Dataset | OpenThoughts-3, 1.2M-example single-turn subset |
| Prompts trained on | $128 \times 3000 = 384\,000$ |
| Loss support | response tokens only (all assistant messages) |
| Global batch size | 128 |
| Max sequence length | 16 384 |
| Total epochs | 1 (single pass) |
| Learning-rate schedule | linear decay to 0, no warmup |
| Learning rate | swept in $\{1 \times 10^{-6}, 1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}\}$ (see Table 7) |
| Adam $(\beta_1, \beta_2, \epsilon)$ | $(0.9,\ 0.95,\ 1 \times 10^{-8})$ |
| Weight decay | 0 |
| Fine-tuning mode | full (no LoRA) |
| Precision / infra | FSDP, 8×H200 (single node) |

Table 7: SFT learning-rate sweep for the cold-start initialization. `mean@16` (%) is reported. Best per row in bold.

(a) Qwen3-1.7B-Base

| Benchmark | $1 \times 10^{-6}$ | $1 \times 10^{-5}$ | $1 \times 10^{-4}$ | $1 \times 10^{-3}$ | Base ref |
|---|---|---|---|---|---|
| AIME24 | 4.8 | 1.3 | **11.3** | 0.0 | 3.5 |
| AIME25 | 3.1 | 2.7 | **13.8** | 0.0 | 2.7 |
| AMC23 | 26.4 | 13.6 | **37.5** | 0.3 | 27.0 |
| OOD macro | 11.4 | 5.9 | **20.8** | 0.1 | 11.1 |

(b) Qwen3-4B-Base

| Benchmark | $1 \times 10^{-6}$ | $1 \times 10^{-5}$ | $1 \times 10^{-4}$ | $1 \times 10^{-3}$ | Base ref |
|---|---|---|---|---|---|
| AIME24 | 7.9 | 20.8 | **47.1** | 0.0 | 10.0 |
| AIME25 | 5.0 | 16.9 | **40.0** | 0.0 | 7.7 |
| AMC23 | 31.6 | 49.3 | **73.0** | 2.5 | 39.0 |
| OOD macro | 14.8 | 29.0 | **53.4** | 0.8 | 18.9 |

**LiveCodeBench v1–v5 and CodeForces (code training), LCB v6 (code eval)**

```
<User Prompt>
You are a coding expert. You will be given a coding problem, and you need
to write a correct Python program that matches the specification and
passes all tests. The time limit is 1 second. You may start by outlining
your thought process. In the end, please provide the complete code in a
code block enclosed with ``` ```.

{problem statement}
```

Reward: sandboxed unit-test execution against the row's `ground_truth` `inputs`/`outputs`; a response is scored 1 if every test case passes within the time limit and 0 otherwise.

**MMLU-Pro 5-shot (knowledge training and eval)**

```
<System Prompt>
Given a question and up to ten options, please select the right answer.
Respond in the following format:
<reasoning>
...
</reasoning>
<answer>
```

```
...
</answer>

For the answer, only output the letter corresponding to the correct
option (one of A, B, C, D, E, F, G, H, I, or J), and nothing else. Do
not restate the answer text. For example, if the answer is "A", just
output:
<answer>
A
</answer>

<User Prompt>
Here are 5 worked examples showing the required format.

=== Example 1 ===
{example 1 question}
{options 1 (A ... J)}

<reasoning>
{worked reasoning 1}
</reasoning>
<answer>
{letter 1}
</answer>

... (five worked examples in total, drawn once from the MMLU-Pro
validation split and shared across all training prompts) ...

=== Now answer this ===
{target question}
{target options}
```

Reward: exact letter match; a response is scored 1 if the string inside `<answer>...</answer>` equals the ground-truth letter and 0 otherwise.

**IFBench-RLVR (instruction-following training)**

```
<User Prompt>
{base prompt} {constraint sentence}

Example row (instruction_id = paragraphs:paragraphs):
John goes to the store to buy Slurpees and gives them $20. Slurpees cost
$2 each and he got $8 in change. How many Slurpees did he buy? There
should be 2 paragraphs. Paragraphs are separated with the markdown
divider: ***
```

Reward: rule-based; a response is scored 1 if it satisfies all the constraints, and 0 otherwise.

# D External robustness: per-ablation detail

This section gives the detailed description per ablation of §4.2.

## D.1 Training-data difficulty

This subsection documents the data construction and Qwen3-1.7B-Base results behind the difficulty grid in §4.1. The main in-domain and OOD findings are discussed there; here we provide construction details, the full per-cell OOD breakdown, and the 1.7B-specific observations that inform the broader study.

**Data construction.** We score every prompt in DAPO-Math-17K with `mean@32` under both the Qwen3-1.7B-Base student and the Qwen3-4B thinking teacher, then bin along both axes ($=0$/$>0$ for the student; $=0$, $(0, 0.3]$, $(0.3, 0.6]$, $>0.6$ for the teacher), yielding eight cells with matched train/in-domain-validation splits; for every subset, we keep the training set size to be 442 and validation set to be 128. We use a Qwen3-1.7B-Base student in no-think mode for the full eight-cell sweep; the OPD recipe in Yang et al. (2025) distills in either `/think` or `/no_think` mode, so no-think is a supported regime, and the 1.7B pattern is if anything more severe than the 4B headline.

**OOD performance of Qwen3-1.7B-Base.** The most striking feature of the 1.7B difficulty grid is the behaviour on the OOD macro (AIME24/25 + AMC23; Figure 8). While GRPO's OOD accuracy varies substantially across the eight cells, tracking which training subset it was given, OPD's OOD macro is nearly constant at 23.3–25.6 regardless of the training cell. This constancy holds across all three teacher sizes (4B, 8B, 32B) and persists even in the degenerate cell where both student and teacher score zero on the training prompts. In-domain, GRPO wins every cell over the best OPD results by a noticeable margin. On OOD, GRPO wins five of the eight cells, but OPD-4B and OPD-8B finish above GRPO in the three cells where the teacher itself performs badly (the entire teacher $= 0$ column plus the student-zero/low-teacher cell). This reversal is the opposite of what the dense-signal motivation predicts: OPD gains its OOD advantage precisely where the teacher has the least task-solving information to contribute, suggesting a distributional component in OPD's learning that is not captured by the quality of the teacher's per-token signal (see Appendix F).

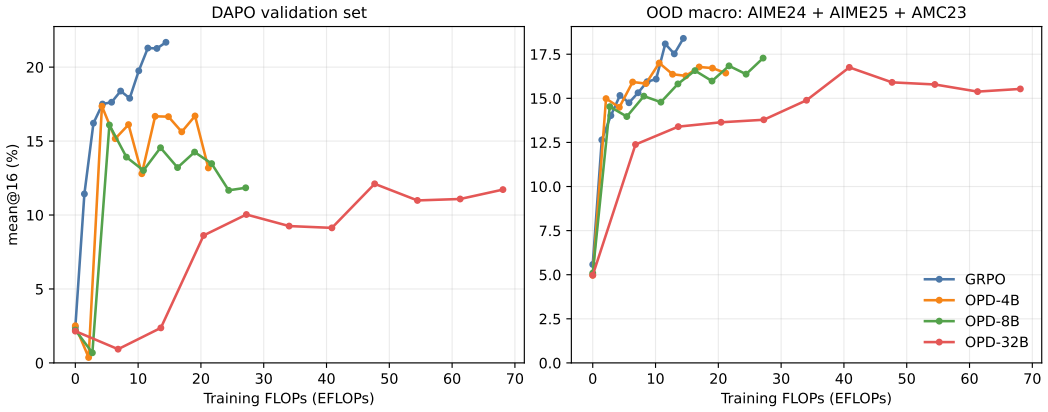



Figure 7: **GRPO dominates OPD across compute on the unstratified training set.** Qwen3-1.7B-Base student, DAPO-Math-17K, no-think, learning rate $2 \times 10^{-6}$, batch 64, 8 rollouts per prompt, 200 steps, evaluated every 20 optimizer steps against realised-token training FLOPs. **Left:** `mean@16` on a held-out DAPO validation split. **Right:** OOD macro over AIME24, AIME25, AMC23. Every OPD teacher curve plateaus below GRPO at multiples of GRPO's compute.

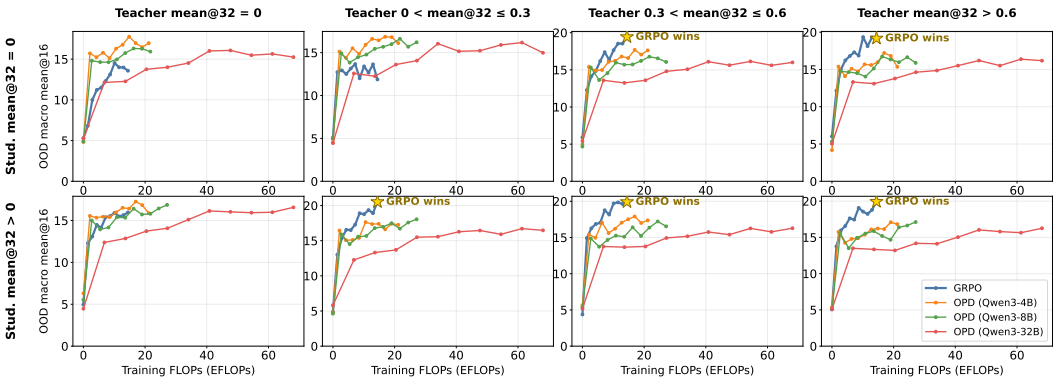



Figure 8: **OOD macro (`mean@16` on AIME24 + AIME25 + AMC23) across training task difficulty level**, one subplot per (student, teacher) `mean@32` cell, companion to Figure 2. Qwen3-1.7B-Base student, Qwen3-{4B, 8B, 32B} teachers, no-think, learning rate $2 \times 10^{-6}$, batch 64, 32 rollouts per prompt, 200 steps. GRPO wins five cells (marked "GRPO wins"); in the three low-teacher-difficulty cells (the entire teacher $= 0$ column plus the (student $= 0$, teacher $\in (0, 0.3]$) cell) OPD-4B and OPD-8B finish above GRPO on the OOD macro even though GRPO won the same cells in-domain (Figure 2). OPD-32B is the worst run on OOD in every cell.

**Difficulty study with Qwen3-1.7B.** Figure 9 repeats the difficulty study with Qwen3-1.7B (in no-think mode) instead of Qwen3-1.7B-Base, using a finer student-side stratification (4 student bins

$\times$ 3 teacher bins). We keep the amount of training/validation sets the same as what's used for Qwen3-1.7B-Base. The performance gap between GRPO and OPD is even larger. In all training sets, GRPO performance increases on both in-domain and out-of-domain. In contrast, OPD with external teachers makes student performance even worse.

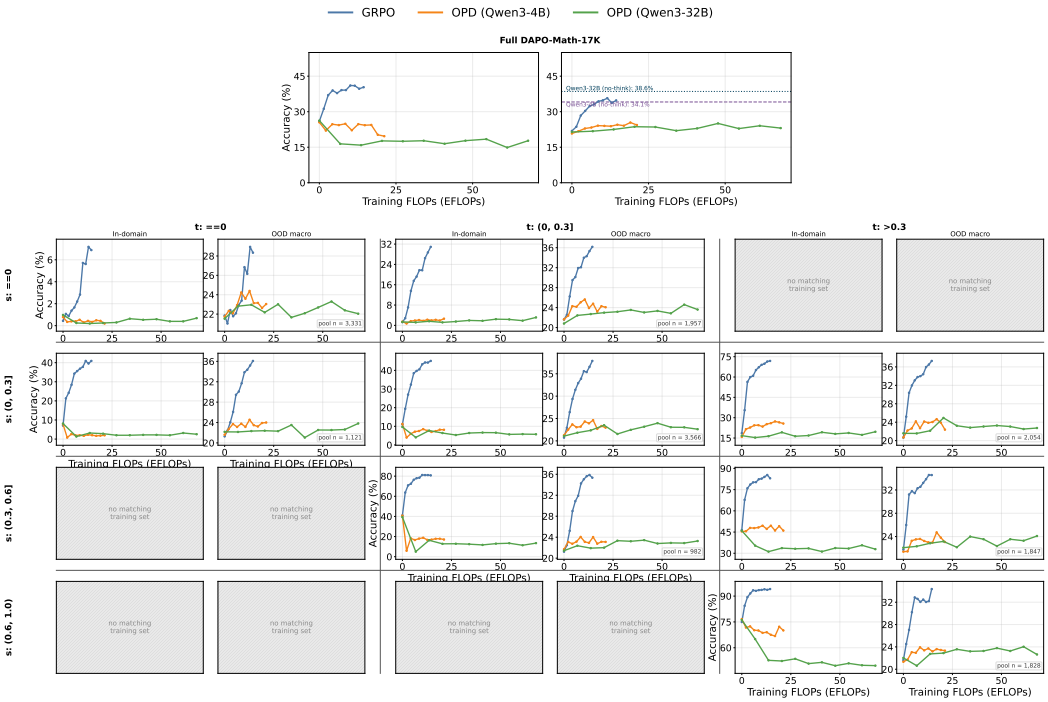



Figure 9: **Difficulty grid with Qwen3-1.7B, no-think.** Each cell shows in-domain validation (left) and OOD macro (right) against training FLOPs. Student bins: `mean@32` $= 0$, $(0, 0.3]$, $(0.3, 0.6]$, $(0.6, 1.0)$; teacher bins: $= 0$, $(0, 0.3]$, $> 0.3$. Grey cells lack enough training data. Top row: unstratified DAPO-Math-17K with teacher reference lines (Qwen3-32B no-think: $38.6\%$, Qwen3-4B no-think: $34.1\%$). Although we only use 442 instances for training, we label the total examples in each subset in the subplot – "pool n=".

### D.2 Learning rate

We swept the learning rate over five values ($2 \times 10^{-6}$, $5 \times 10^{-6}$, $1 \times 10^{-5}$, $2 \times 10^{-5}$, $5 \times 10^{-5}$) for GRPO and for all three OPD teacher sizes on DAPO-Math-17K, no-think (Figure 10). OPD is more lenient: OPD-4B's best-over-training OOD macro spans $4.3$ points across rates ($14.1$–$18.4$), whereas GRPO's spans $10.3$ points ($8.6$–$18.9$). GRPO is more sensitive to the learning rate, but the best GRPO configuration outperforms the best OPD in both accuracy and compute efficiency: GRPO at lr $2 \times 10^{-6}$ reaches $18.9\%$ OOD macro at $7.8\,\text{EFLOP}$, while OPD's best (OPD-4B at lr $5 \times 10^{-6}$) reaches $18.4\%$ at $14.3\,\text{EFLOP}$, requiring $1.8\times$ the compute for a lower score.

### D.3 Batch size

Batch size is another key hyperparameter for model tuning: it controls the noise level of gradient estimates and the number of update steps per training instance. To isolate its effect while keeping the total training data fixed, we ran three batch-size configurations – 64 for 200 steps, 128 for 100 steps, and 256 at 50 steps – for GRPO and for OPD under each of the three teacher sizes, all on DAPO-Math-17K with a Qwen3-1.7B-Base student in no-think mode at learning rate $2 \times 10^{-6}$ (Figure 11). Smaller batches help both methods: GRPO climbs from $19.5$ to $20.9$ to $24.0$ (in-domain validation mean@16, %) as the batch shrinks from 256 to 128 to 64; OPD-4B follows the same staircase from $16.0$ to $16.2$ to $17.3$. The gain is roughly parallel, so the relative ordering GRPO $>$ OPD holds at

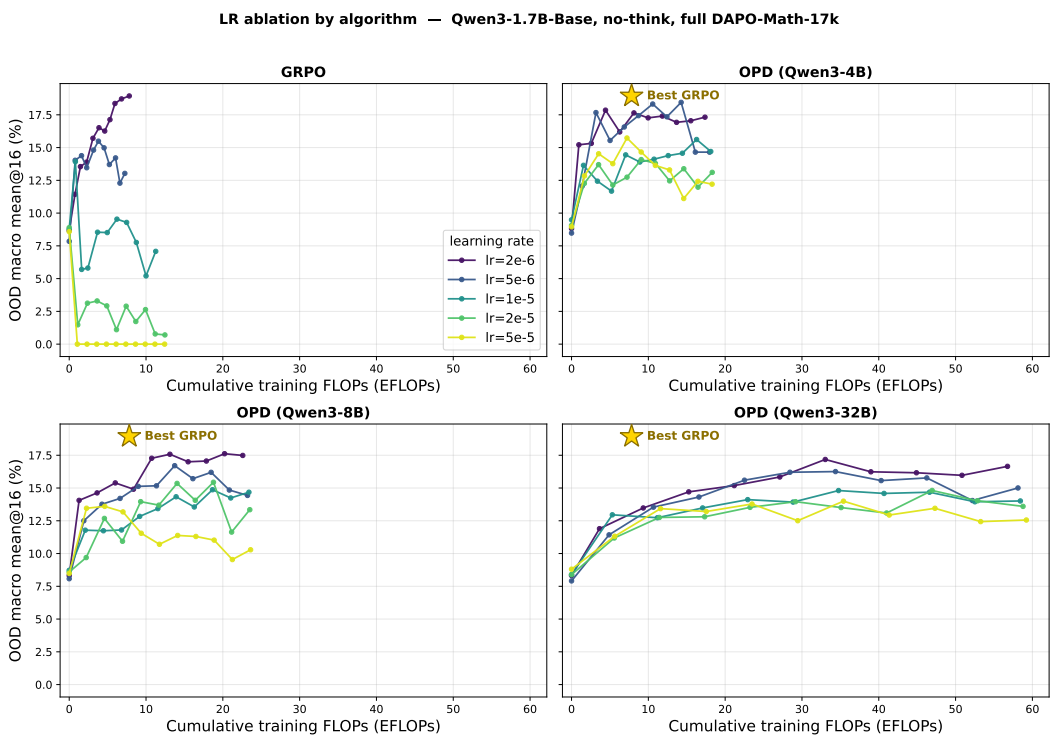



Figure 10: **Learning-rate sweep.** OOD macro (`mean@16`) versus training FLOPs for five learning rates on DAPO-Math-17K, Qwen3-1.7B-Base, no-think. Performance of all runs decreases when learning rate increases. OPD (three teacher sizes) is more stable across rates than GRPO, yet GRPO's best rate ($2 \times 10^{-6}$) still outperforms every OPD configuration.

every batch size. For the two larger teachers (8B, 32B) the pattern is noisier but still favours small batch, and at no batch size does any OPD run overtake GRPO.

### D.4 Reverse-KL estimator variant

The default OPD loss in our experiments uses the *closed-form top-$k$* reverse KL: at each token position, the student's top-$k$=16 tokens define a support set, both the student and teacher distributions are renormalized onto that set, and the KL is computed exactly over the $k$ tokens (§3). An alternative is the *sampled-token* estimator, which evaluates the reverse KL at only the single token the student actually produced. The $k_1$ estimator replaces the inner sum over the top-$k$ support with the importance-weighted log-ratio at the single token $o_{i,t}$ sampled from the rollout policy $\pi_{\text{old}}$:

$$\mathcal{L}_{\text{OPD}}^{k_1}(\theta) = \mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\text{old}}(o_{i,t} \mid q, o_{i,<t})}\Big[\log \pi_{\text{old}}(o_{i,t} \mid q, o_{i,<t}) - \log \pi_T(o_{i,t} \mid q, o_{i,<t})\Big]\right].$$

The ratio $\pi_\theta/\pi_{\text{old}}$ corrects for off-policy sampling, and the per-token signal $\log \pi_{\text{old}} - \log \pi_T$ is negative where the teacher assigns higher probability than the rollout policy. This estimator needs only the student, rollout, and teacher log-probabilities at the sampled token, avoiding the top-$k$ renormalization entirely.

Figure 12 compares the two estimators under the same recipe (Qwen3-1.7B-Base, no-think, DAPO-Math-17K, lr $2 \times 10^{-6}$, batch 64, 200 steps). The two variants trace essentially the same accuracy curves across all three teacher sizes and across difficulty bands, so the choice of reverse-KL support set is not what determines OPD's loss to GRPO. This is consistent with the observation by Kimi Team (2026) that the estimator choice has minimal impact in practice.

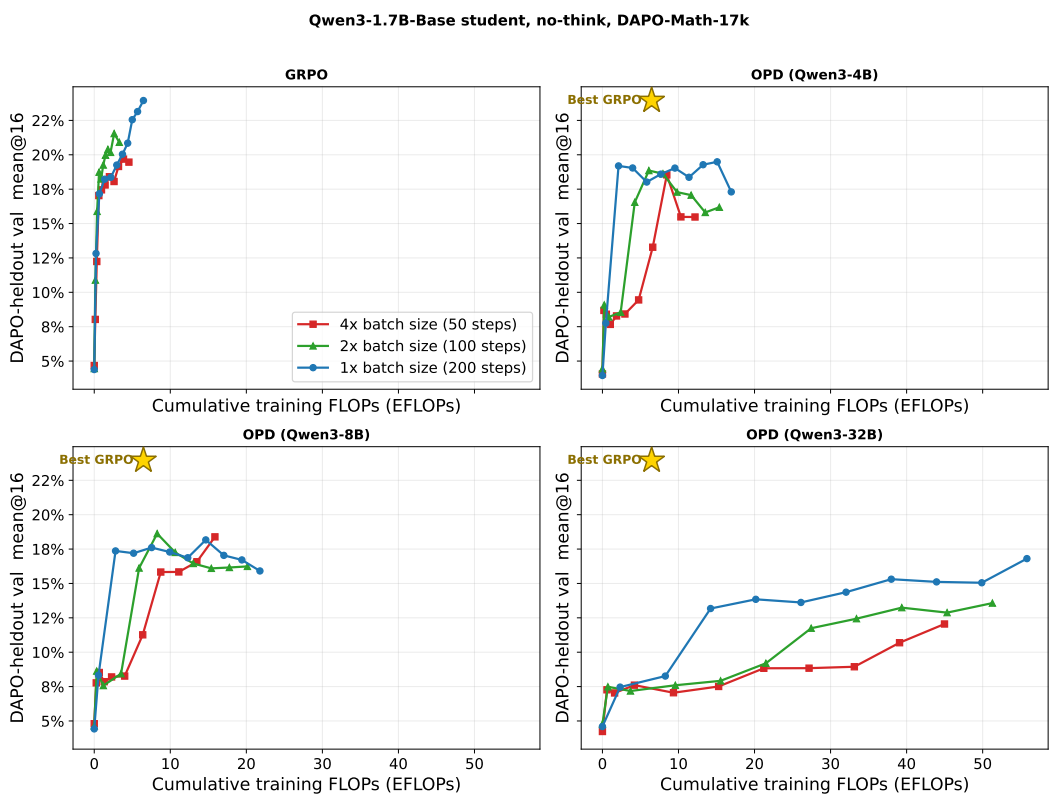



Figure 11: **Budget-controlled batch-size comparison.** We control the total number of training prompts to be 12 800 across all three configurations: batch size 64 for 200 steps, 128 for 100 steps, and 256 for 50 steps ($64 \times 200 = 128 \times 100 = 256 \times 50 = 12\,800$). Qwen3-1.7B-Base student, no-think, DAPO-Math-17K, learning rate $2 \times 10^{-6}$; the $x$-axis (prompts seen) aligns all three configurations. Smaller batches help both GRPO (top left) and every OPD run roughly in parallel; at no batch size does any OPD run overtake GRPO.

### D.5 Thinking mode

Figure 13 isolates the effect of the teacher's thinking mode. All runs use the same student (Qwen3-1.7B-Base with OpenThoughts-3-SFT initialization) and the same data (DAPO-Math-17K); only the teacher decoding mode changes. In no-think mode (left panel), GRPO reaches roughly 19% OOD macro and the three OPD variants (4B, 8B, and 32B teachers) cluster in a narrow 16%–17.5% band, consistent with the main-text pattern. Switching to think mode (right panel) costs GRPO about 2.1 points, but the effect on OPD varies sharply with teacher size: OPD-8B loses a comparable 2.1 points and stays within 1.5 points of GRPO, whereas OPD-4B drops 8.2 points and OPD-32B drops 7.7 points, leaving both at roughly half the GRPO level.

Two observations follow. First, thinking mode hurts OPD more than GRPO across most teacher sizes, consistent with the hypothesis that long chains of thought amplify the distribution mismatch between student rollouts and teacher token-level supervision. Second, teacher capacity does not rescue the gap: the 32B teacher fares worse than the 8B teacher in think mode, not better, so the fragility is not a matter of teacher scale. We note that no-think and think runs differ in decoding temperature and prompt format, so the two panels are not a controlled comparison with each other; the GRPO-vs-OPD contrast within each panel is controlled.

### D.6 Student initialization

A natural hypothesis is that the base model must first learn to produce rollouts in the right format before OPD can distill from them effectively: if the student's on-policy outputs are too far from the

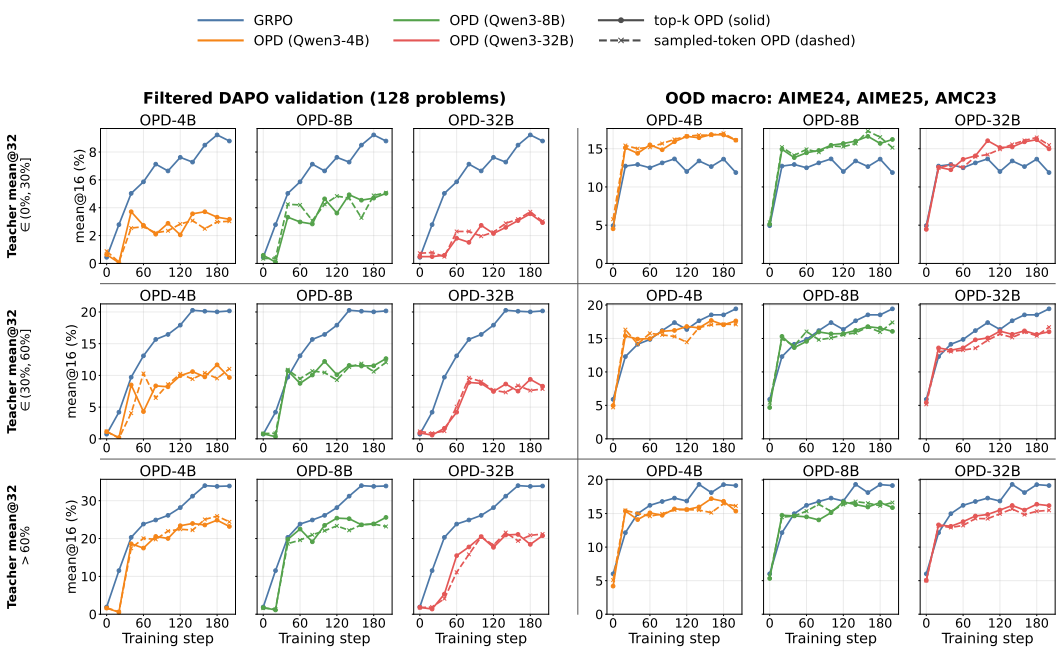



Figure 12: **Reverse-KL estimator ablation: sampled-token vs top-$k$ token.** Two OPD loss variants on the same recipe, per teacher size and per student-zero difficulty band like Figure 8 and Figure 2. Student is set to be Qwen3-1.7B-Base, no-think, DAPO-Math, lr $2 \times 10^{-6}$; same evaluation; everything else fixed. The two estimators trace essentially the same accuracy curves throughout training.

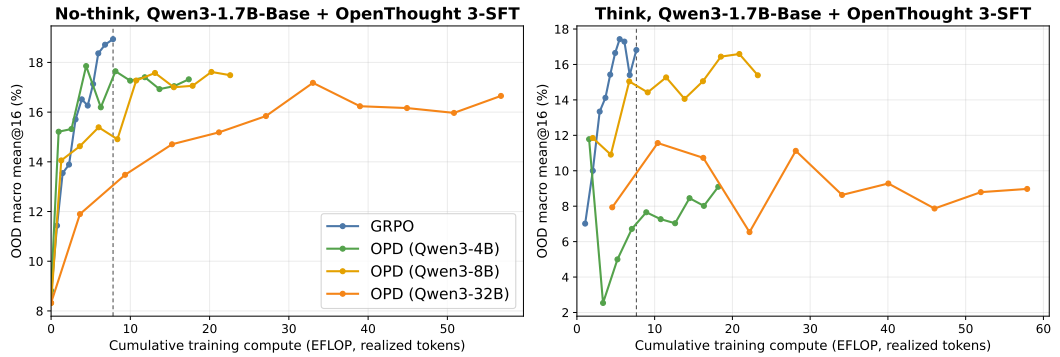



Figure 13: **Thinking-mode ablation (Qwen3-1.7B-Base, OpenThoughts-3-SFT init, DAPO-Math-17K).** OOD macro (`mean@16`) versus training FLOPs. Left: no-think teachers. Right: think teachers. GRPO loses about 2 points across modes; OPD-4B and OPD-32B lose 7–8 points each; OPD-8B is comparably robust to GRPO.

teacher's distribution, token-level supervision may have little to work with. SFT cold-starting on high-quality traces (e.g. OpenThoughts-3) could therefore be important for OPD to succeed. We test this hypothesis at two student scales.

**1.7B-Base, no-think mode.** Figure 14 compares Qwen3-1.7B-Base initialized from the external pretrained checkpoint against the same model initialized from an OpenThoughts-3-SFT cold start (lr $1 \times 10^{-5}$, step 3000). Both runs use no-think mode with matched hyperparameters (lr $2 \times 10^{-6}$, $k$=32 rollouts, batch 64, 200 steps). The SFT initialization provides a head start at step 0, but GRPO overtakes all OPD variants under both initializations. The SFT cold start does not change the conclusion.

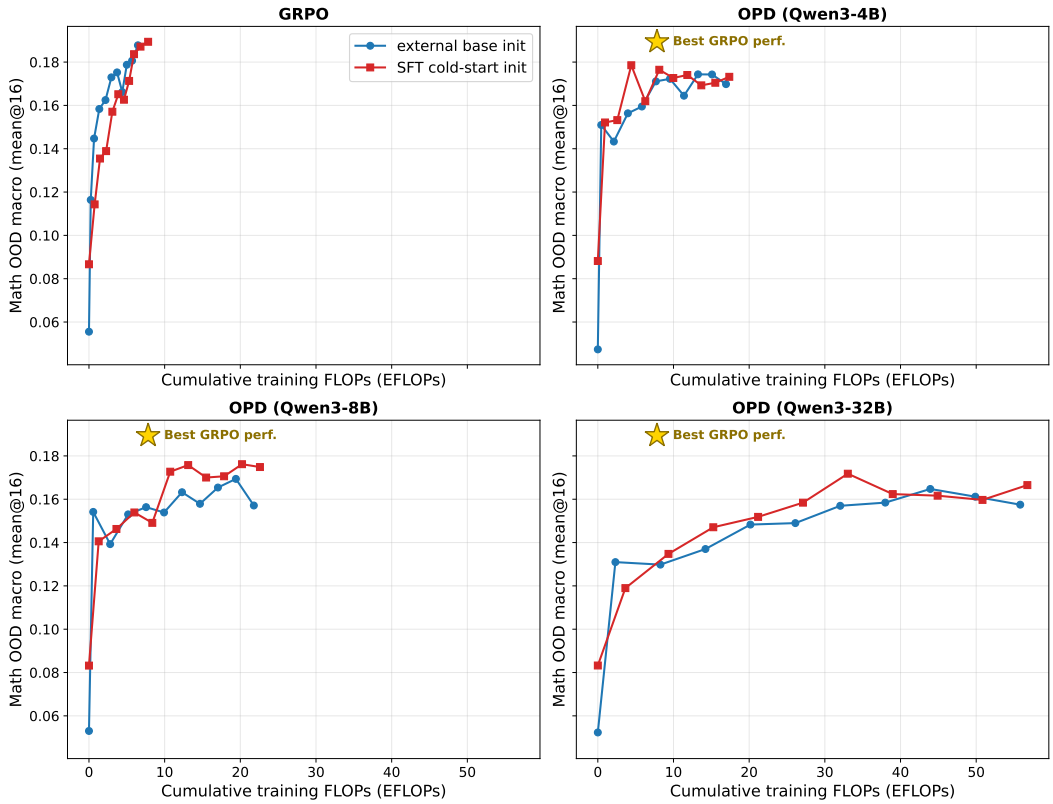



Figure 14: **Cold-start init for Qwen3-1.7B-Base, no-think mode.** OOD macro `mean@16` (AIME24 + AIME25 + AMC23) over training steps. All runs share: no-think, $k$=32 rollouts, lr $2 \times 10^{-6}$, DAPO-Math-17K, batch 64, 200 steps, 7168-token cap. **Left:** external base init. **Right:** OpenThoughts-3-SFT cold-start init (lr $1 \times 10^{-5}$, step 3000). The SFT init provides a head start at step 0 but GRPO overtakes all OPD variants under both initializations.

**1.7B-Base, think mode.** We next test two SFT cold starts with thinking enabled: lr $1 \times 10^{-5}$ (the same checkpoint as above) and lr $1 \times 10^{-4}$ (the optimal SFT learning rate). Figure 15 plots OOD macro against training FLOPs for both. Under both cold starts the within-panel ordering holds: GRPO sits above every OPD run per FLOP. Switching to a stronger SFT initialization and enabling thinking does not change the ranking.

**4B-Base with SFT cold start.** We scale the test to Qwen3-4B-Base with thinking enabled, initialized from OpenThoughts-3-SFT (lr $1 \times 10^{-4}$, the optimal learning rate for this model), and increase the response cap from 7168 to 16 384 tokens. In addition to the standard OPD-32B teacher, we include QwQ-32B, the model whose rollouts generated the OpenThoughts corpus used for the SFT cold start, to accommodate teacher similarity with the SFT-initialized student. Figure 16 shows GRPO outperforming both OPD teachers across all four evaluation benchmarks. The QwQ-32B teacher does not recover the gap, and both OPD variants decline over training while GRPO continues to improve.

### D.7 Student model sizes

The negative result in §4 is established on Qwen3-1.7B-Base. A natural concern is that the 1.7B student is simply too small to benefit from a strong teacher and that the gap would close at larger student scale. We test this by running the same external comparison at Qwen3-4B-Base and Qwen3-8B-Base (Figure 17). At both scales, no external OPD run reaches its teacher's accuracy within the compute budget, and a larger teacher does not improve compute efficiency: the 32B teacher curves

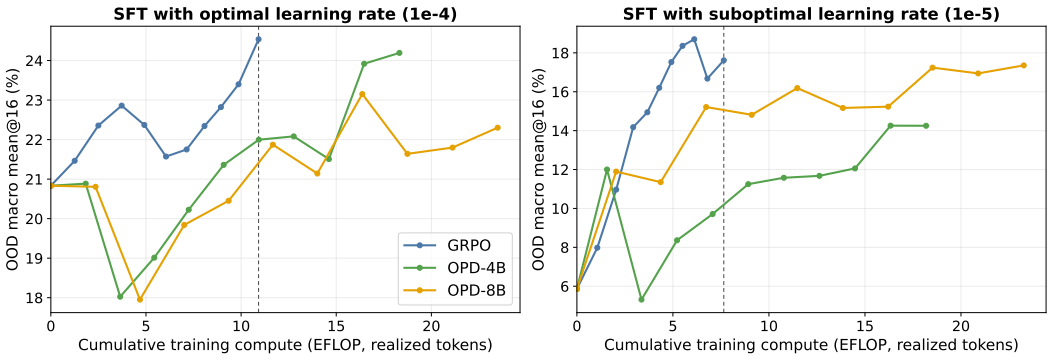



Figure 15: **Cold-start init for Qwen3-1.7B-Base, think mode.** OOD macro `mean@16` (AIME24 + AIME25 + AMC23) against cumulative training FLOPs. Both panels share: thinking enabled, DAPO-Math-17K, batch 64, 8 rollouts, 7168-token training cap, 30 720-token evaluation cap. **Left:** SFT init at lr $1 \times 10^{-5}$. **Right:** SFT init at lr $1 \times 10^{-4}$ (the optimal SFT rate). The two panels use different RL learning rates ($2 \times 10^{-6}$ vs. $1 \times 10^{-6}$), so the panel-to-panel contrast is uncontrolled; the GRPO-vs-OPD contrast within each panel is controlled.

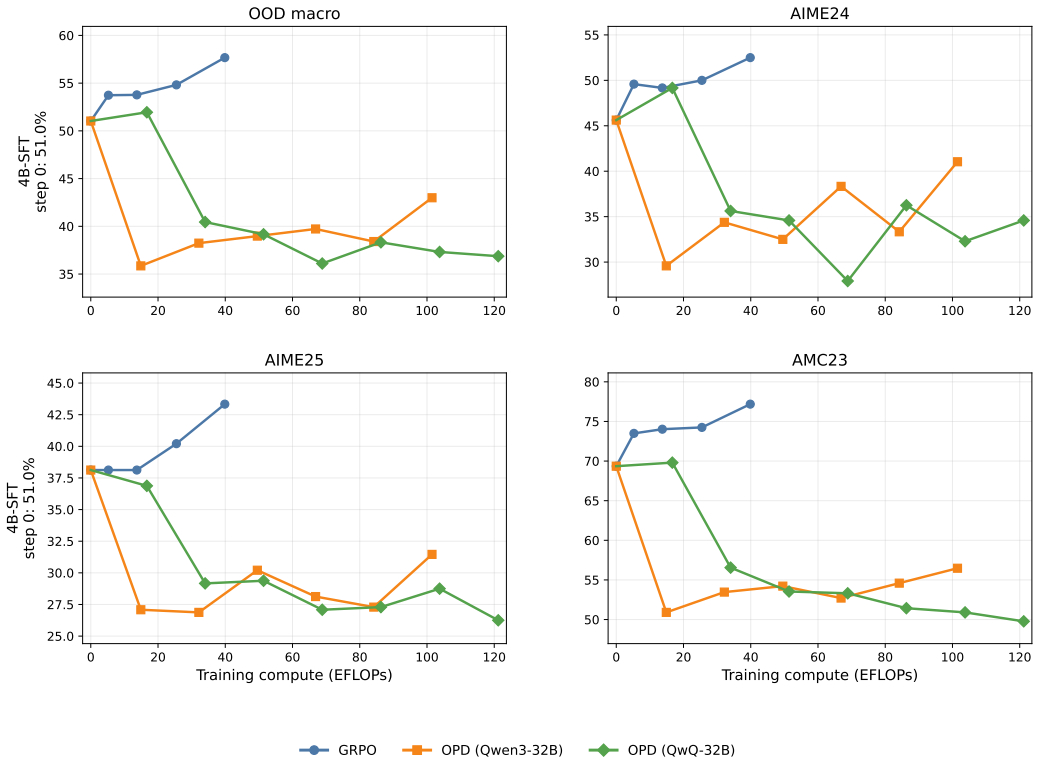



Figure 16: **Cold-start init for Qwen3-4B-Base-SFT, compute axis.** OOD macro and its three components (AIME24, AIME25, AMC23) against cumulative training FLOPs. The student is Qwen3-4B-Base initialized from OpenThoughts-3-SFT (optimal learning rate $=1 \times 10^{-4}$). Teachers: Qwen3-32B (OPD-32B) and QwQ-32B (OPD-QwQ-32B, the generator of the SFT corpus). GRPO outperforms both OPD variants across all panels; the distribution-matched QwQ-32B teacher does not close the gap.

sit close to (and sometimes below) the same-size teacher curves. The negative result is not a small-student artifact.

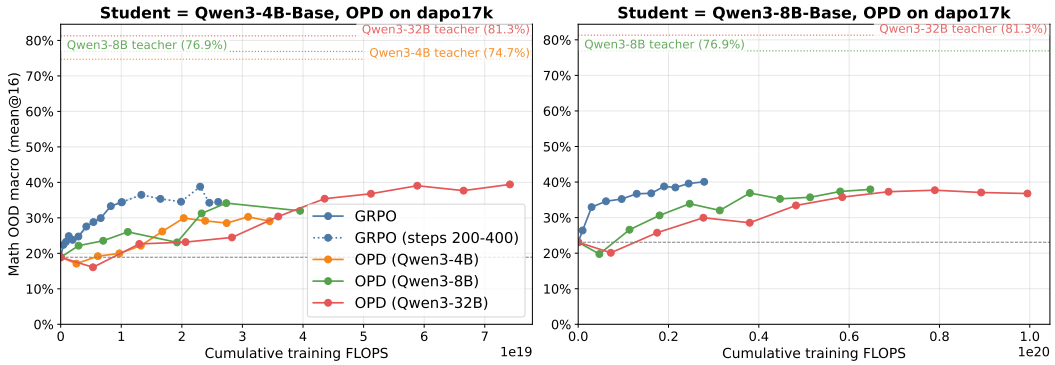



Figure 17: **External thinking teachers, 4B and 8B students.** Two panels, one per student size. Curves are training trajectories on DAPO-Math-17K for direct GRPO and for OPD against external Qwen3-{4B, 8B, 32B} thinking teachers; $x$ is cumulative training FLOPs, $y$ is OOD macro `mean@16` (AIME24 + AIME25 + AMC23). GRPO curves: solid up to optimizer step 200, dotted for steps 200–400 (the OPD runs trained for at most 200 steps, so the solid segments are the matched-step comparison). No OPD run reaches its teacher's accuracy within the compute budget, and a larger external teacher does not close the gap. The same-family GRPO-trained teachers of §5 are omitted here to isolate the external ablation.

### D.8 Training domain

**Math.** Our default math training set is DAPO-Math-17K. We also run on DeepMath-103k, the dataset used by Thinking Machines Lab (Thinking Machines Lab, 2025) for their OPD experiments. Swapping DAPO-Math for DeepMath does not change the ordering; it widens it, with GRPO ahead by 11.5 points of `mean@16` on DeepMath versus 7.7 on DAPO (Figure 18).

**Code.** We train the same Qwen3-4B-Base student on two code datasets (LiveCodeBench v1–v5 and a high-quality CodeForces set we carve from `open-r1/codeforces`; Appendix B.6) with GRPO and OPD under external 8B and 32B teachers. The ordering observed on math holds on code: GRPO outperforms both OPD variants on LCBv6 at every evaluated compute budget on both training sets (Figure 19, left column).

### D.9 Replication across RL frameworks

We replicated the comparison across three RL stacks. The first two are two generations of our `verl`-based training pipeline. Our headline infrastructure uses a slightly older `verl` generation in which we implement the reverse-KL OPD objective ourselves. We also re-ran the comparison in the latest `verl` release, which natively supports reverse-KL OPD with a sampled-token estimator. Hyperparameters are controlled between the two stacks; the only difference is the reverse-KL estimator (closed-form top-$k$ in our custom implementation against sampled-token in the native one). Figure 21 shows both `verl` generations: GRPO stays above every OPD run at every evaluated step under either implementation.

**tinker replication at 9B scale.** The third replication uses the **`tinker`** training library from Thinking Machines Lab (Thinking Machines Lab, 2025), at a different scale and student family. From a shared rank-128 SFT checkpoint of Qwen3.5-9B-Base (as student/initialization) and a Qwen3.5-9B teacher (OPD run) we ran GRPO and OPD for 40 optimizer steps at matched settings (learning rate $1 \times 10^{-4}$, 64 prompts per batch, 4 rollouts per prompt, 16 384-token response cap, DeepMath data); the choice of hyperparamters are detailed in TML's tutorial page [1] or default values in the `tinker` codebase. At step 40, GRPO reaches AIME'24 `mean@8` $= 0.542$ while OPD collapses to `mean@8` $= 0.013$; 98.75% of OPD's evaluation rollouts saturate the 16 384-token cap against 26.3% for GRPO (Figure 20). This is the same length-inflation pathology documented on `verl` in

[1] `https://tinker-docs.thinkingmachines.ai/cookbook/recipes/distillation/`

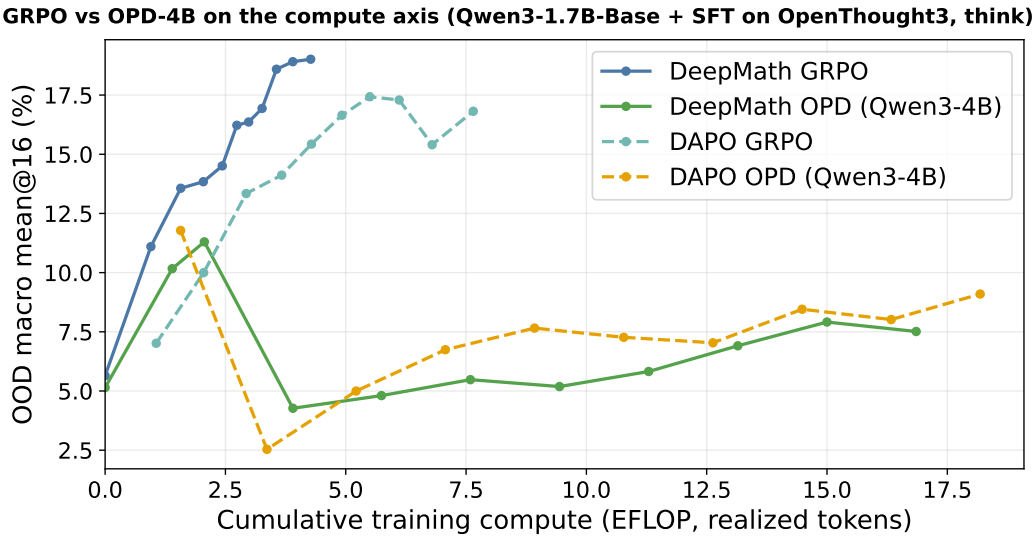



Figure 18: Training domain on the compute axis: DAPO-Math-17K and DeepMath-103k, GRPO vs OPD-4B on a Qwen3-1.7B-Base (after SFT) student with thinking mode on. Swapping the training set does not change the ordering per FLOP, both OPD runs end below their GRPO twin and pay more compute to get there.

Appendix E, now reproduced on a different framework, student family (Qwen3.5), base scale (9B), and cold start (rank-128 LoRA).

Using a controlled suite of GRPO and OPD within each framework, we show that GRPO is more compute-efficient than OPD with an external teacher across all three stacks.

# E LENGTH AS A LENS TO COMPARE MODEL BEHAVIORS

## E.1 LENGTH INFLATION EXCEPT SELF-GRPO TEACHER

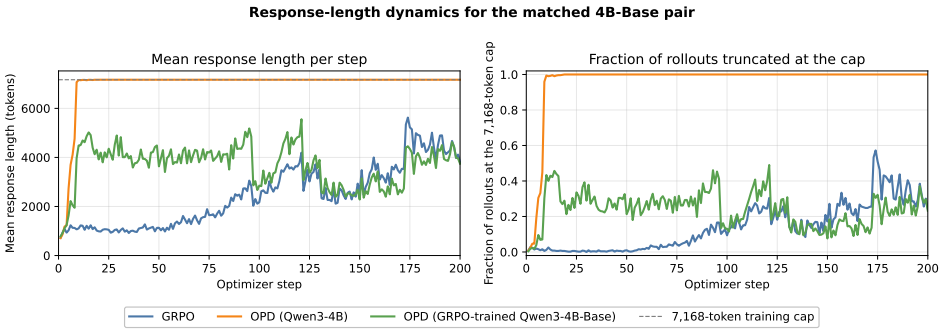



Figure 22: Response-length dynamics for the matched 4B-Base pair; the external teacher is an external thinking model Qwen3-4B, the same size as the student. **Left:** mean response length per step (dashed line: 7168-token cap). **Right:** fraction of rollouts truncated at the cap. OPD reaches the cap by step 9 and holds every sequence there for 182 of 200 steps, whereas GRPO (3813 tokens) and the self-GRPO-teacher run OPD-4B-GRPO (3741 tokens) never fill it.

With an external thinking teacher, OPD response lengths inflate rapidly until every rollout saturates the response cap. Here the teacher is an external thinking model Qwen3-4B —the *same size* as the Qwen3-4B-Base student—so the collapse cannot be attributed to the student lacking the capacity to represent a much larger teacher. OPD nonetheless drives the mean response length past 7000 tokens

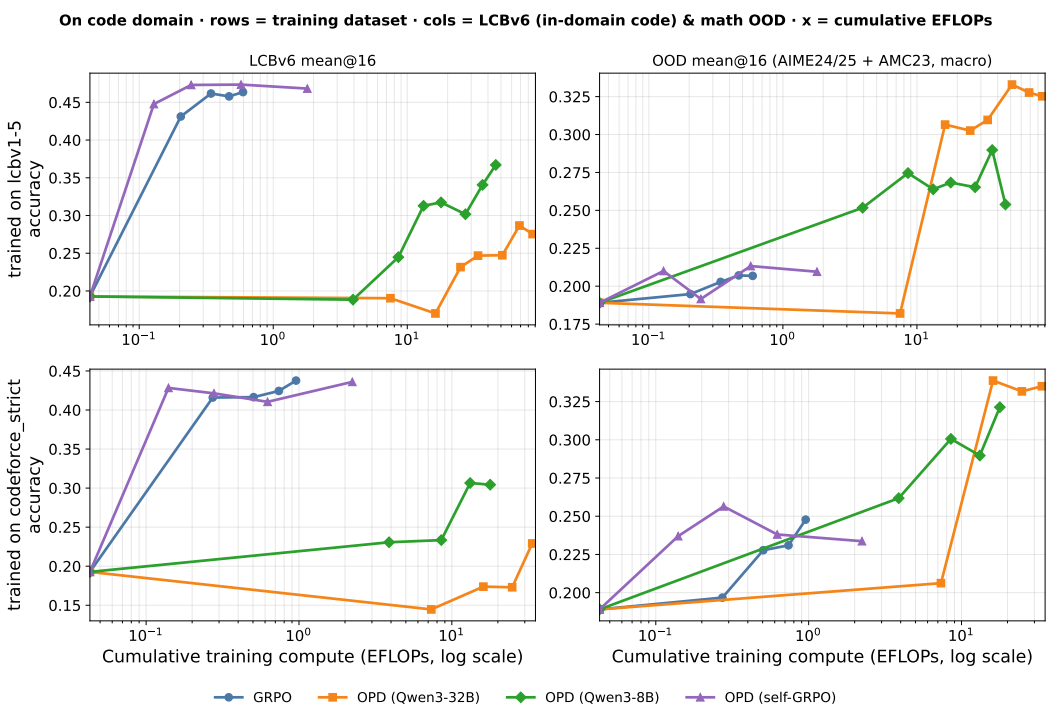



Figure 19: **Code training: in-domain collapse, out-of-domain math extrapolation.** Qwen3-4B-Base student trained on LiveCodeBench-v1–v5 (top row) and strict CodeForces (bottom row) with GRPO or external OPD-32B, lr $2 \times 10^{-6}$, 16 384-token response cap; $x$-axis is cumulative training compute (EFLOPs). **Left column:** LCBv6 `mean@16` (in-domain code). OPD-32B loses to GRPO at every evaluated compute budget on both training sets. **Right column:** math OOD macro `mean@16` on AIME24 + AIME25 + AMC23. Despite training only on code, OPD-32B reaches $\approx$31%–34% math OOD, beating the code-trained GRPO by roughly 10 points on both training sets. Additional evidence for OPD's cross-domain extrapolation mechanism.

by step 9 (with 96% of rollouts already truncated) and then holds every sequence in the batch at the 7168-token cap (clip ratio $\geq 0.999$) from step 19 onward (Figure 22). The same student trained with no teacher (GRPO) or self-GRPO teacher never fills the cap: it holds near 1000–1500 tokens for the first $\approx$55 steps and ends at a mean length of 3813 tokens (53% of the budget).

This pathology is not inherent to dense per-token supervision. When the teacher is a GRPO checkpoint of the student (§5), OPD does not exhibit length inflation and recovers the teacher at 50–74$\times$ fewer FLOPs than GRPO. The problem is not the dense signal itself but the wrong dense signal: an external teacher whose token distribution is far from the student's. However, it is difficult to characterize a priori what makes a given model a good teacher. Given an arbitrary LLM, there is currently no way to predict whether it will function as a stable OPD teacher without running the distillation. Developing such a diagnostic remains an open problem.

**Additional results with other teacher provenance** Figure 23 plots mean response length (tokens) on the OOD math evaluation across training steps for three teacher families — GRPO-trained base models, external thinking models, and further-GRPO-trained thinking models. With an external thinking teacher (centre panel), the OPD-trained student's responses balloon to $\approx$28 000–30 000 tokens by step 40 and stay there, while the teachers themselves respond with only $\approx$12 000–13 000 tokens on the same prompts (4B teacher: 12 708; 8B teacher: 13 040; 14B teacher: 11 578). The student is therefore emitting $\approx$2.3–2.6$\times$ longer responses than the teacher whose token distribution it is trying to match, an outright divergence from teacher behavior rather than mere length matching. The same pattern holds with a Thinking + GRPO teacher (right panel: student $\approx$22 000–30 000 vs teacher $\approx$9000–10 000). With a GRPO-trained base teacher (left panel), student response length is much shorter but still mismatch with teachers'. When teacher is GRPO-trained 14-base, student

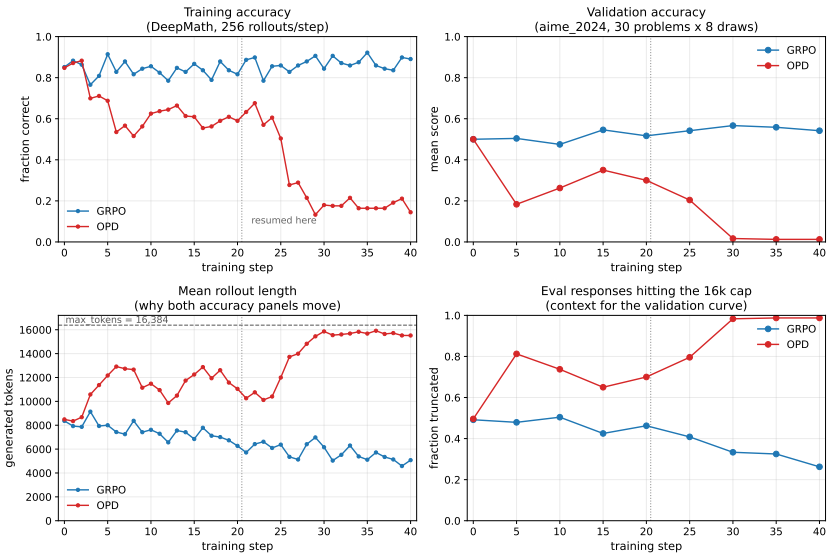



Figure 20: **tinker replication, Qwen3.5-9B-Base.** GRPO vs OPD on DeepMath from a shared rank-128 SFT checkpoint, teacher Qwen3.5-9B for OPD. AIME'24 accuracy over 40 optimizer steps. GRPO reaches `mean@8` $= 0.542$; OPD collapses to $= 0.013$ with $98.75\%$ of evaluation rollouts saturating the 16 384-token cap. The within-framework ordering (GRPO $>$ OPD) and the length-saturation mechanism are the same as the two `verl` replications above.

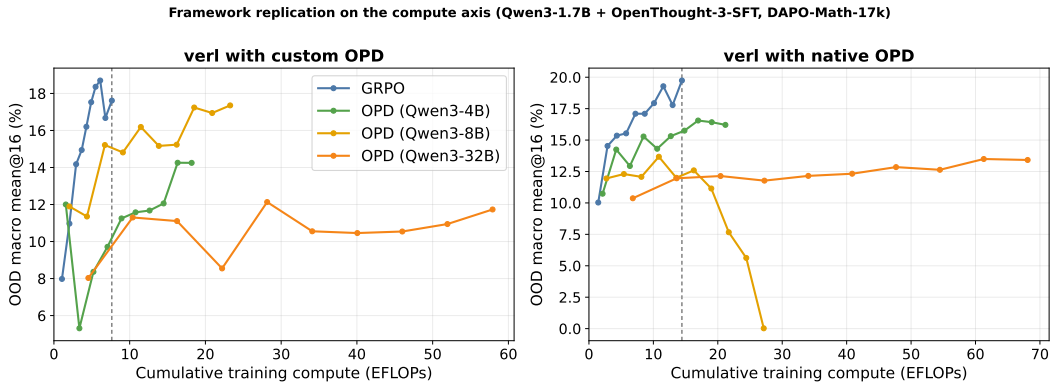



Figure 21: **Replicate GRPO $>$ OPD ordering on different frameworks.** The compute ordering survives the change of framework: GRPO is above every OPD run per FLOP in both the old/custom and latest/native stacks. On latest/native stack, we follow their default choice to use sampled-token reverse KL estimator. Other than the difference we specified, the hyperparamters and configurations match between left and right plots.

output response with ∼9K length while teacher outputs 4.7K; when teacher is GRPO-trained 8B-base, student's output lengths is around 12K outnumbering teacher's 8.9K; Only when teacher is the student's GRPO checkpoint, we observe student and teacher response length matches.

### E.2 Length inflation co-occurs with other pathology modes in OPD with external thinking models

We first isolate the effect of teacher choice in the simplest setting: math-only training. All three runs below share the student (Qwen3-4B-Base), the training data (DAPO-math-17K), the training cap (16,384 tokens), the eval cap (30,720 tokens), and the decoding configuration (temperature 0.6, top-$p$ 0.95, $n$=16). Only the training objective and teacher differ: GRPO on the target reward, OPD

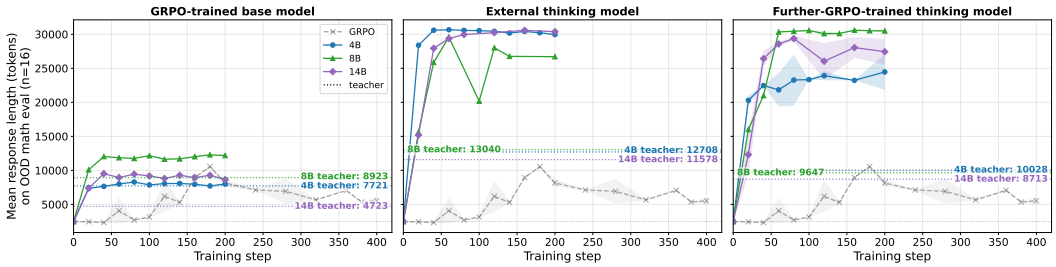



Figure 23: **Mean response length (tokens) on OOD math, by teacher type.** Qwen3-4B-Base student. Solid curves: student response length over training steps. Dotted horizontals: the teacher's own eval-set response length on the same prompts. With external thinking teachers (centre) and thinking + GRPO teachers (right), the OPD-trained student's responses are ≈2–3× longer than the teacher's own responses. Only with GRPO-trained Qwen3-4B-Base as teacher (left) does the student's response length remain nearly identical to the teacher's.

distilled from a self-GRPO teacher (the same 4B model, trained via GRPO on this data), and OPD distilled from an external thinking teacher Qwen3-4B. For each run we take the first 500 rollouts from the step-200 evaluation on AIME24/25 + AMC23 and measure response length with the Qwen3 tokenizer.

| metric | GRPO | OPD, self-GRPO teacher | OPD, external teacher |
|---|---|---|---|
| repetition loop (naive detection) | 16.6% | 15.2% | 63.6% |
| median response length (tokens) | 6,504 | 6,542 | 30,720 **(= cap)** |
| ↪ *if detected loops removed* | 4,573 | 4,937 | 21,031 |
| mean response length (tokens) | 11,233 | 11,605 | 30,229 |
| ↪ *if detected loops removed* | 7,763 | 8,353 | 19,627 |
| ≥ 95% of eval cap | 19.0% | 21.0% | 97.8% |
| CJK language collapse | 0.0% | 0.0% | 0.2% |
| missing \boxed{·} | 21.2% | 23.2% | 49.0% |

Table 8: Rollout pathologies at step 200 on math-only training (500 rollouts per run, Qwen3 tokenizer). Values in **bold red** are ≥ 2× the GRPO baseline on the same metric. "Repetition loop (naive detection)" flags a rollout when the 100-char window `response[-200:-100]` appears ≥ 3 times elsewhere in the response — a conservative probe that catches short-period token-level loops but misses longer-period, paraphrased, and semantic repetition. "↪ if detected loops removed" cuts each detected-loop response at the first occurrence of its probe pattern and re-measures length; because the probe is conservative, this is an *upper bound* on unique-content length, and still contains an unknown amount of undetected higher-order repetition. Even against this upper bound the external-teacher response is ∼ 4× the GRPO de-looped median.

GRPO and self-GRPO-teacher OPD are essentially indistinguishable: within ∼ 40 tokens of median response length and ∼ 2 percentage points on every metric. The self-GRPO teacher transfers its stopping behavior almost perfectly to the student. The external-4B-teacher OPD is qualitatively different: median response saturates at the eval cap, 97.8% of rollouts hit the cap, and 63.6% contain a detected short-period repetition loop — the student has lost its ability to stop. Removing detected loops brings the median only from 30,720 to 21,031 tokens, still ∼ 4× GRPO's de-looped median of 4,573.

The same picture holds on the 2-domain math + code mix. Joint GRPO and multi-teacher self-GRPO OPD ($G$=1, one GRPO specialist teacher per domain) both terminate cleanly: cap-hit stays under 21% on math and under 4% on code for both, and multi-teacher self-GRPO OPD is in fact *tighter* than GRPO on code (0.2% cap-hit and 0.2% detected repetition, versus 3.6% / 3.6% for GRPO). The external teacher (Qwen3-8B) collapse to the opposite regime on *both* domains: 100% cap-hit, ∼ 85% detected repetition loops, and median response saturated at the eval cap (30.7k math, 31.7k code). On top of this shared length-inflation failure, the Qwen3-8B external teacher demonstrates

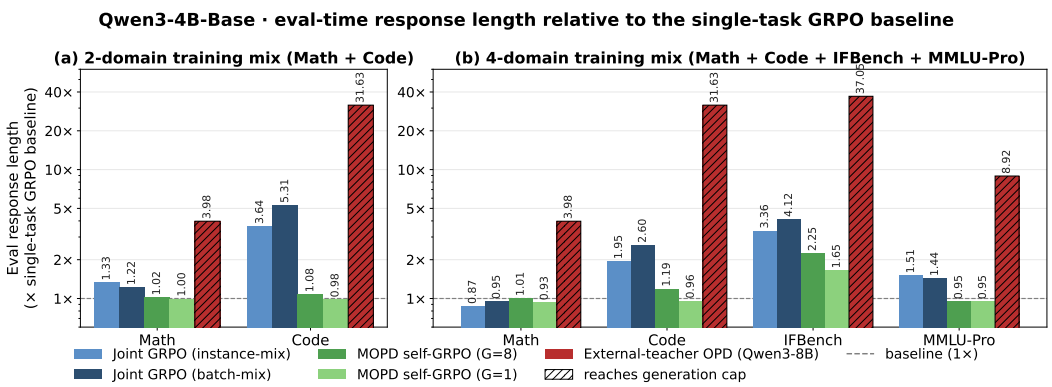



Figure 24: **Evaluation response length by regime, as a ratio to the single-task GRPO reference (Qwen3-4B-Base)**: (a) two-task mix (math, code); (b) four-task mix (adds IFBench, MMLU-Pro). The dashed line is the reference and the $y$-axis is logarithmic. Hatched bars reach the generation limit (non-terminating output).

a language-collapse mode absent elsewhere — $100\%$ of its math and code rollouts contain Chinese characters in the first 2k characters, versus $0\%$ under GRPO and multi-teacher self-GRPO.

### E.3 Length behavior in multi-domain settings

We measure mean evaluation response length (tokens, $n$=16 per prompt) for Qwen3-4B-Base, taking single-task GRPO as the per-task reference at the checkpoint each multi-teacher OPD distills from (math, IFBench, and MMLU-Pro at 200 steps, code at 80; Table 9). Figure 24 reports each regime's length as a ratio to this reference. Multi-teacher OPD, which distills from these single-task GRPO teachers, recovers the teacher length closely: within roughly $20\%$ on math, code, and MMLU-Pro ($0.93$–$1.19\times$), and somewhat longer on IFBench ($1.65$–$2.25\times$). OPD from an external teacher (Qwen3-8B) instead inflates length on every task like what happens in Appendix E.2. Joint GRPO lies between the two, and its length is setting-dependent.

Table 9: Single-task GRPO reference length per task (Qwen3-4B-Base): mean evaluation response length, in tokens, at the checkpoint used as the multi-teacher OPD teacher.

| Task | Length (tokens) |
|---|---|
| Math | 7,721 |
| Code | 1,003 |
| IFBench | 829 |
| MMLU-Pro | 804 |

Under joint GRPO, each task's response length is pulled toward the baseline lengths of the *other* tasks sharing its batch (Table 9). In the two-task math + code mix, math's long single-task baseline drags code well above its own, to $3.64\times$ the code reference, while math itself is only mildly inflated ($1.33\times$). Adding MMLU-Pro and IFBench, the two shortest single-task baselines, pulls both shared tasks back down: math from $1.33\times$ to $0.87\times$ the reference and code from $3.64\times$ to $1.95\times$. This is not a training-duration artefact — at step 160, math is $1.44\times$ in the two-task mix but only $0.32\times$ in the four-task mix. Response length under joint GRPO is thus governed by the mix of tasks in the training batch.

## F Evidence for an extrapolation component in OPD learning

Although external thinking teachers are compute-inefficient for OPD on the training domain (§4), our experiments surface an interesting side benefit: OPD with these teachers produces surprisingly strong out-of-distribution generalization that GRPO does not match. We collect two observations below suggesting that OPD's per-token objective carries information beyond the training distribution.

**Observation 1: OPD wins OOD in the difficulty cells where the teacher itself is weak.** On the difficulty-stratified DAPO-Math-17K grid (§4.1), GRPO wins in-domain in every one of the eight cells (Figure 2). On the OOD macro (AIME24 + AIME25 + AMC23), however, OPD-4B and OPD-8B finish *above* GRPO in three of the eight cells, every cell in the teacher-solves-nothing column plus the (student $= 0$, teacher $\in (0, 0.3]$) cell (Figure 8). In these cells the teacher signal is weak, so it is surprising that OPD outperforms GRPO on OOD. In parallel, OPD's OOD macro sits at a nearly constant $23.3$–$25.6$ across all eight regimes while GRPO's varies with the data it was trained on ; OPD's OOD score looks more like a property of the objective than of what it was trained on.

**Observation 2: code-trained OPD with external teachers improves math OOD by $\approx$7–13 points over code-trained GRPO.** On two code training sets (LCBv1–v5 and CodeForces), both external teachers reach math OOD macro well above the code-trained GRPO despite never seeing math training data (Figure 19 right column). OPD(Qwen3-32B) peaks at $33.3\%$ on LCBv1–v5 and $33.9\%$ on CodeForces; OPD(Qwen3-8B) peaks at $29.0\%$ and $32.1\%$ respectively; code-trained GRPO peaks at $20.7\%$ and $24.8\%$.

## G Additional results of running GRPO on teacher

The fast-saturation pattern of the self-GRPO teacher (§5.1) is not specific to the 4B student. Figure 25 repeats the comparison with a Qwen3-8B-Base student against three teachers: external Qwen3-8B, external Qwen3-32B, and the student's own GRPO-trained Qwen3-8B-Base checkpoint. The self-GRPO teacher saturates the student within $\approx 5\,\mathrm{EFLOP}$ at roughly its own accuracy, and does so faster than either external teacher; the external 8B and 32B teachers climb more slowly and remain below through the compute budget shown.

On code, the same pattern holds: the GRPO-trained 4B-base teacher reaches GRPO's LCBv6 accuracy at $2.4\times$ fewer FLOPs on LiveCodeBench-v1–v5 and ties GRPO's quality on strict CodeForces, while the external 32B teacher collapses on both (Figure 19 left column). GRPO-trained student itself is therefore the only strong-to-weak OPD setup that consistently works across both math and code domains.

Figure 26 gives the 4B + 8B math view together on a single compute axis (companion to Figure 25): at both student sizes the self-GRPO teacher reaches its own accuracy within $\sim 0.1\,\mathrm{EFLOP}$, while every external thinking teacher (4B, 8B, 32B) climbs slowly and stays below. The single-rollout ($G{=}1$) variant of the self-GRPO run replicates the same fast recovery at both scales, on this single-domain math setting.

**Non-base cold-start replication (Olmo-3-7B-Instruct-SFT).** Figure 27 runs the same self-GRPO recipe on a post-trained checkpoint from a different model family. Both curves share the step-0 Olmo-3-7B-Instruct-SFT student, the DAPO-Math-17K training set, and an identical evaluation configuration. *GRPO (self)* runs direct GRPO on the SFT student ($n{=}8$, 200 steps) and reaches $40\%$ OOD macro at $14.4\,\mathrm{EFLOP}$. *OPD $\leftarrow$ self-GRPO teacher* distills the step-200 GRPO checkpoint back into the same step-0 SFT student ($n{=}1$, 80 steps) and reaches the teacher's end-of-training accuracy within $\sim 0.36\,\mathrm{EFLOP}$ (step 20), a $\sim 40\times$ compute reduction that lines up with the same-base results in Section 5.1. Unlike the same-base setting, however, the OPD student then continues past the teacher and peaks at $51\%$ at $1.25\,\mathrm{EFLOP}$, an $\approx$11-point overshoot over the self-GRPO teacher's own plateau.

## H Measure model weight similarity by matrix spectra

Prior work (Zhu et al., 2026a) has investigated the change of model weight during post-training through the lens of spectral inheritance: RLVR can reuse the base model's weight spectra while acquiring new behavior through changes in the associated input and output singular frames. In this section, we will conduct similar analysis through the measure proposed by Zhu et al. (2026a).

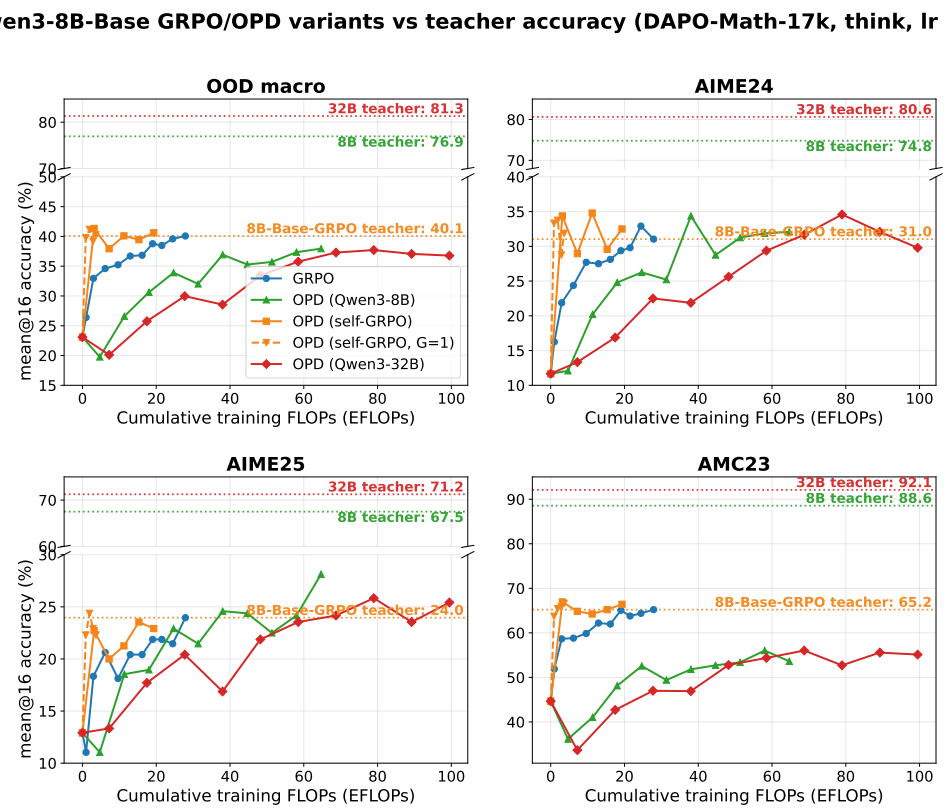



Figure 25: **Self-GRPO teacher replicates at the 8B student (companion to Figure 3).** Qwen3-8B-Base student on DAPO-Math-17K, thinking enabled, lr $2 \times 10^{-6}$; `mean@16` vs cumulative training FLOPs. Dotted horizontals mark each teacher's own accuracy. The self-GRPO run (GRPO-trained 8B-base teacher) outperform its teacher within 0.90 EFLOPs (≈40% OOD macro), surpasing GRPO checkpoint's performance (which requires 27.9 EFLOP to reach); external teachers take substantially more compute and still underperforms overall.

### H.1 Setup

For each pretrained matrix $W_B \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ of the base student, take its SVD $W_B = U_B \operatorname{diag}(\sigma_B) V_B^\top$ with $q = \min(d_{\text{out}}, d_{\text{in}})$, then decompose any candidate displacement $\Delta W = W_1 - W_B$ in the base frames. Here $\sigma(W)$ denotes the $q$-vector of singular values of $W$ – $\sigma(W) = (\sigma_1, \ldots, \sigma_q)$ in non-increasing order.

Three spectrum-based scalars (Zhu et al., 2026a) are used to characterize changes in each matrix. First, $\kappa_{\text{spec}}$:

$$\kappa_{\text{spec}}(\Delta W) = \frac{d_{\text{out}} d_{\text{in}}}{q} \cdot \frac{\|\operatorname{diag}(U_B^\top \Delta W V_B)\|^2}{\|\Delta W\|^2} \tag{2}$$

Larger $\kappa_{\text{spec}}$ indicates more eigenvalue-shift bias. Zhu et al. (2026a) shows that RLVR updates leave a student's singular-value spectrum nearly unchanged ($\kappa_{\text{spec}} \approx 1$–$3$), while SFT updates shift it by orders of magnitude ($\kappa_{\text{spec}} \gtrsim 10^2$–$10^3$).

$\rho_\Sigma$ is the share of the update magnitude that shows up as a change in singular values rather than a rotation of the base frames: $\rho_\Sigma = 0$ marks a pure Stiefel move (frames rotate, spectrum unchanged), $\rho_\Sigma = 1$ marks a purely spectral update (same singular directions, only the values change).

$$\rho_\Sigma(\Delta W) = \frac{\|\sigma(W_1) - \sigma(W_B)\|}{\|\Delta W\|} \tag{3}$$

$\delta_\Sigma$ is the absolute size of that spectral shift relative to the base weight scale: $\delta_\Sigma \approx 0$ means the spectrum barely moves compared to $W_B$, while $\delta_\Sigma \approx 1$ means the singular values change by an

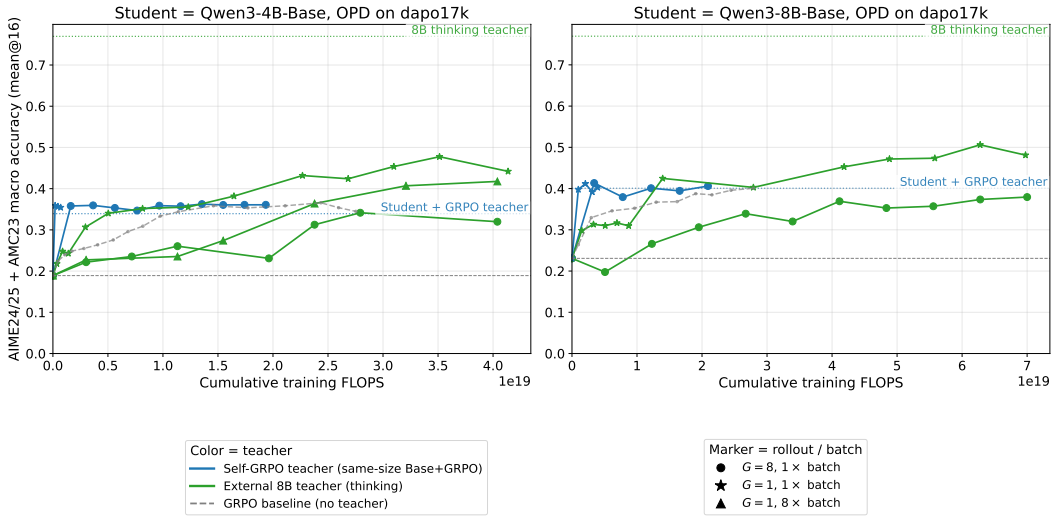



Figure 26: $G{=}1$ **OPD with external thinking teachers on math, 4B and 8B students.** OOD macro (AIME24 + AIME25 + AMC23) `mean@16` versus cumulative training FLOPs. Reducing OPD to $G{=}1$ with the external 8B-thinking teacher improves compute-efficiency at both student scales. The self-GRPO teacher ($G{=}1$) saturates at each student's own GRPO-teacher accuracy (dotted horizontal) within $\sim$0.1 EFLOP.

amount comparable to the base norm itself. On the same runs Zhu et al. (2026a) reports RLVR updates keep $\rho_\Sigma$ small (a few percent) and $\delta_\Sigma$ vanishingly small ($\sim 10^{-5}$–$10^{-4}$), while SFT-scale updates push $\rho_\Sigma \gtrsim 0.15$ and $\delta_\Sigma$ by orders of magnitude.

$$\delta_\Sigma(\Delta W) = \frac{\|\sigma(W_1) - \sigma(W_B)\|}{\|W_B\|} \tag{4}$$

Following the practice in Zhu et al. (2026a), we calculate each value for each matrix and aggregate by averaging.

### H.2 Per-pair spectrum-based scalars

Table 10 reports the three spectrum-based scalars for four (student, teacher) pairs at 4B and 8B scale. At both scales, the external thinking teacher sits far from the student in weight space ($\kappa_{\text{spec}} \approx 2000$–$3600$, $\rho_\Sigma \approx 15\%$); the student's own GRPO-trained checkpoint sits nearly isospectral ($\kappa_{\text{spec}} \approx 1$–$3$, $\rho_\Sigma \approx 6\%$–$12\%$), consistent with RLVR-vs-SFT dichotomy in Zhu et al. (2026a). Three orders of magnitude in $\kappa_{\text{spec}}$ separate the two teacher types, a weight-space signature of the overlap phase transition of §5.

**Interpretation.** Reverse-KL OPD, like GRPO, is a policy-gradient update on the student's own rollouts, and its student trajectories fall in the same RLVR $\kappa_{\text{spec}}$ band regardless of teacher. If OPD can only shift the student's spectrum by a small amount per unit of training compute, then a teacher whose spectrum sits three orders of magnitude away from the student might be out of reach at any realistic compute budget: matching the teacher would require the student to make an SFT-scale spectral move, which OPD's gradient direction does not produce. A self-GRPO teacher, by contrast, sits in the same RLVR band as anything OPD can produce, so a small number of RLVR-magnitude steps suffices. The three-orders-of-magnitude $\kappa_{\text{spec}}$ gap in Table 10 is therefore not just a marker of the overlap phase transition; it is a plausible mechanism for why external-teacher OPD fails compute-efficiently on the training domain (§4).

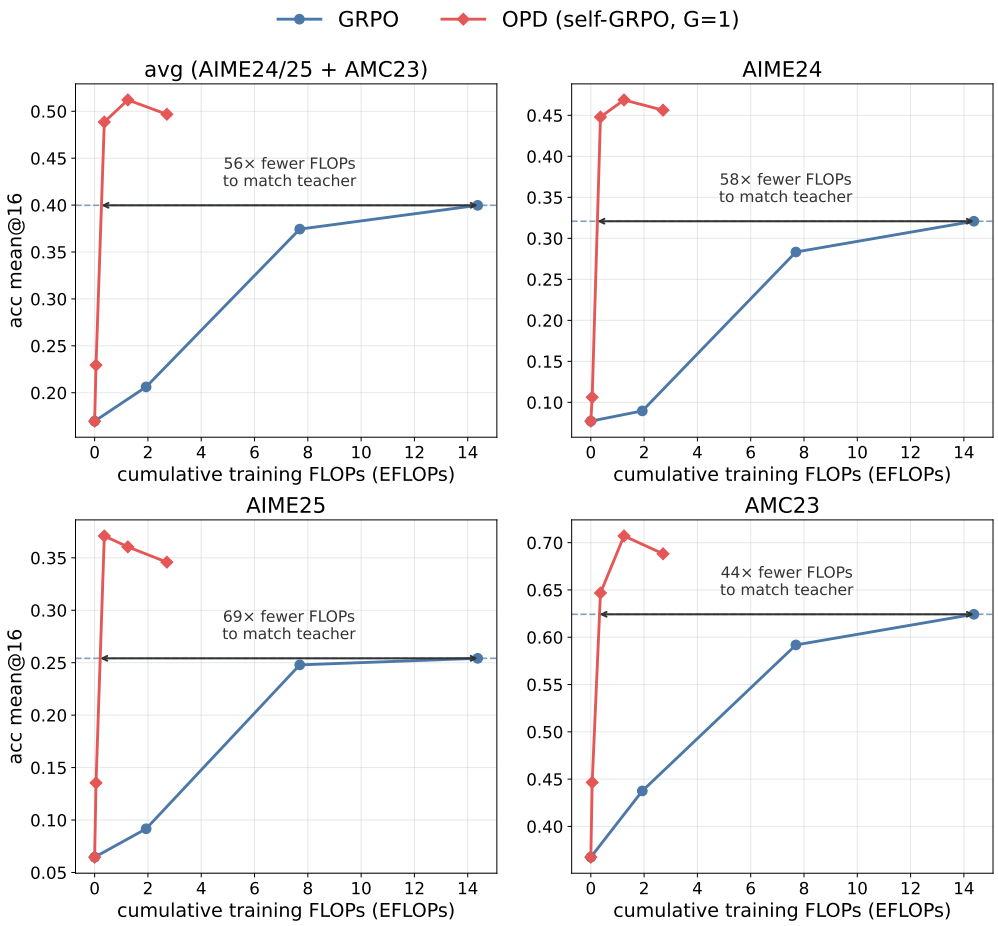



Figure 27: **Self-GRPO OPD on a non-base cold-start (Olmo-3-7B-Instruct-SFT).** OOD math `mean@16` on AIME24 + AIME25 + AMC23 versus cumulative training FLOPs. Both curves share the same step-0 SFT student, the same DAPO-Math-17K training data, and an identical eval configuration. *GRPO (self)*: direct GRPO on the SFT student ($n$=8, 200 steps). *OPD with self-GRPO teacher*: reverse-KL OPD of the step-200 GRPO checkpoint back into the same SFT student ($n$=1, 80 steps). OPD matches the GRPO teacher's step-200 accuracy at $\sim$56$\times$ fewer FLOPs and then substantially outperforms it, an overshoot not seen at this magnitude in the same-base settings of Section 5.1 or the multi-teacher merge of Section 6.

Table 10: spectrum-based scalars for four (student, teacher) pairs. $\kappa_{\text{spec}}$ is computed in the student's own SVD frames. External thinking teachers sit $\sim$3 orders of magnitude farther from the student than the student's own GRPO-trained checkpoint.

| Student | Teacher | $\kappa_{\text{spec}}$ | $\rho_\Sigma$ | $\delta_\Sigma$ | $\|\Delta W\|$ |
|---|---|---|---|---|---|
| Qwen3-4B-Base | Qwen3-4B (thinking) | 3571 | 0.155 | $1.6 \times 10^{-2}$ | 162 |
| Qwen3-4B-Base | GRPO-trained Qwen3-4B-Base | 2.81 | 0.063 | $5.3 \times 10^{-5}$ | 1.33 |
| Qwen3-8B-Base | Qwen3-8B (thinking) | 2077 | 0.155 | $1.3 \times 10^{-2}$ | 202 |
| Qwen3-8B-Base | GRPO-trained Qwen3-8B-Base | 1.19 | 0.125 | $7.7 \times 10^{-5}$ | 1.48 |

# I External OPD under single-rollout regimes

The main negative experiments on external teachers (§4) uses the default rollout group size $G$ for each experiment. A natural objection is that reducing to $G$=1 (a single rollout per prompt) changes the compute profile: OPD pays for fewer rollouts per step, sees more unique prompts at matched

total compute, and avoids the group-relative baseline that GRPO relies on. We test this on two domains.

**Math: $G{=}1$ improves compute-efficiency.** On DAPO-Math-17K with a Qwen3-4B-Base student and an external 8B-thinking teacher, reducing to $G{=}1$ at the default batch size with $8\times$ more optimizer steps (matched total compute) improves OPD compute-efficiency relative to the $G{=}8$ comparison (Figure 26) at the cost of wall-clock time due to more training steps. However, we do note that the length inflation issue (Appendix E.2) still exists.

**Code: $G{=}1$ does not help.** On LiveCodeBench-v1–v5 with the same student and teacher, we test two compute-matched $G{=}1$ variants: one at $8\times$ batch size, and one at the default batch size trained for $8\times$ more optimizer steps. Neither closes the gap: OPD remains substantially below GRPO on LCBv6 `mean@16` (Figure 28). The self-GRPO teacher run on the same recipe reaches $\approx 47\%$ LCBv6 accuracy within $\sim 0.5$ EFLOP; the external 8B-thinking teacher at either $G{=}8$ or either $G{=}1$ matched-compute variant plateaus far below and pays an order of magnitude more compute for less accuracy.

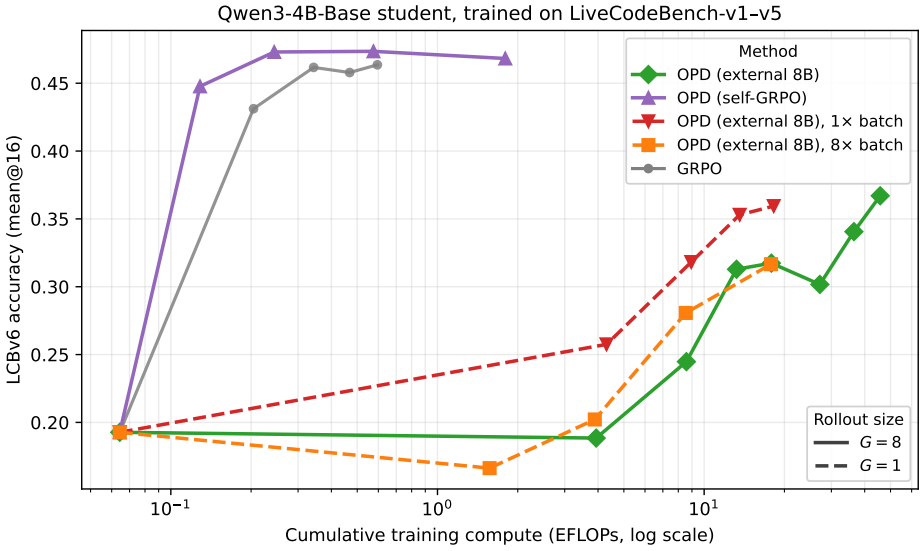



Figure 28: $G{=}1$ **ablation on LCBv6.** Qwen3-4B-Base student trained on LiveCodeBench-v1–v5, thinking, lr $2\times10^{-6}$; LCBv6 `mean@16` versus cumulative training FLOPs. Four OPD runs: external 8B-thinking teacher at $G{=}8$ (green diamonds); self-GRPO teacher at $G{=}8$ (purple triangles); external 8B at $G{=}1$ with the default batch size trained for $8\times$ more optimizer steps (red down-triangles, dashed); external 8B at $G{=}1$ with $8\times$ batch size (orange squares, dashed). The self-GRPO teacher wins by a large margin at a fraction of the compute; both matched-compute $G{=}1$ variants of the external teacher fail to close the gap. The GRPO baseline trained on the same recipe (grey dashed) is shown for reference.

**$G{=}1$ on the multi-domain setting.** The single-domain math benefit of $G{=}1$ with an external 8B-thinking teacher (Figure 26) does not transfer to the multi-domain setting: on the 4-domain mix (Figure 29, purple), the same external 8B-thinking teacher at $G{=}1$ takes the student's math accuracy from a base of 0.19 `mean@16` down to 0.07 within 80 optimizer steps, and code accuracy from 0.19 down to 0.02 over the same window.

**Reading.** Although there might be a hyperparameter configuration that improves external-teacher OPD in a narrow setting like math alone, we did not observe any such configuration being robust across other settings.

## J MULTI-TEACHER OPD: EXPERIMENT DETAIL

**Training data.** The 4-domain mix combines DAPO-Math-17K (math), LiveCodeBench-v1–v5 (code), MMLU-Pro 5-shot prompts (knowledge), and IFBench (instruction following), with equal per-domain exposure. The 2-domain (math + code) experiments use the same math and code splits without the knowledge and instruction-following components.

**Individual GRPO specialists.** From Qwen3-4B-Base we train four separate GRPO runs, one per domain. All four specialists share the hyperparameters in Tables 2 and 3; the per-domain differences (training data, response cap, batch size, and number of steps) are listed in Table 11. Each specialist is evaluated on all four domains to measure cross-domain transfer.

Table 11: Per-domain deviations for the four individual GRPO specialists. All other hyperparameters follow Tables 2 and 3.

| Domain | Training data | Prompt cap | Response cap | Batch size | Steps |
|---|---|---|---|---|---|
| Math | DAPO-Math-17K | 1024 | 7168 | 64 | 200 |
| Code | LiveCodeBench v1–v5 | 1024 | 16 384 | 32 | 80 |
| IFBench | IFBench-RLVR | 1024 | 10 240 | 64 | 200 |
| MMLU-Pro | MMLU-Pro 5-shot | 4096 | 3072 | 64 | 200 |

**Joint GRPO baseline.** A single GRPO run trains on the multi-domain mix with a balanced sampling strategy, using the GRPO-specific hyperparameters in Table 3 shared across domains. For instance-mix variant, we set the response length to 16K tokens and train for 640 optimizer steps on the 4-domain mix and 240 steps on the 2-domain mix. For instance-mix variant, we set to 680 on the 4-domain mix and 280 steps on the 2-domain mix.

**Multi-teacher OPD (MOPD).** A fresh Qwen3-4B-Base student is trained with multi-teacher OPD, routing each training-batch token to the GRPO specialist that owns its domain. We test three rollout group sizes: $G{=}8$, $G{=}4$, and $G{=}1$. No additional RL objective is applied during the merge.

**External teacher baseline.** A single Qwen3-8B (thinking) teacher is applied to the same 4-domain mix via standard OPD. This run collapses on every eval domain, confirming that no single external teacher is simultaneously proximal to all sub-distributions.

**Per-domain results.** Figures 29 and 30 show per-domain breakdowns for the 4-domain and 2-domain experiments respectively.

**Rollout group size for multi-teacher OPD.** Figures 31 and 32 isolate the effect of the rollout group size $G$ on multi-teacher OPD, holding the student, teachers, and data mix fixed. On both the 4-domain and the 2-domain mix, $G{=}1$ reaches the specialist ceiling at substantially lower compute than $G{=}8$, and the per-FLOP trajectory matches what we observe with self-GRPO OPD on a single domain (Figure 3, left panel). We therefore report the $G{=}1$ setting in Section 6.

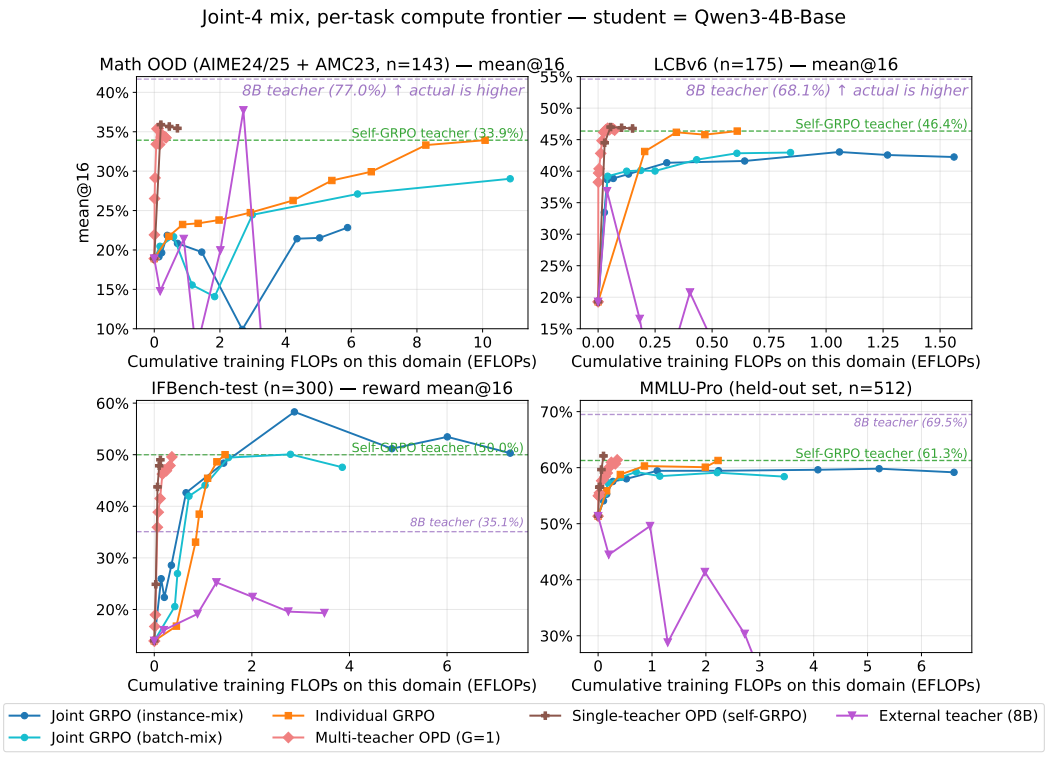



Figure 29: **Per-domain breakdown: 4-domain mix.** Qwen3-4B-Base student, `mean@16` versus training FLOPs per domain (math = AIME24/25 + AMC23, code = LCBv6, plus IFBench and MMLU-Pro). Curves shown: *Joint GRPO (instance-mix)*, *Individual GRPO* specialists, *Self-GRPO OPD (multi-teachers)*, *Self-GRPO OPD (single-teacher)* using each domain's own specialist as teacher, and *external OPD* with a single Qwen3-8B thinking teacher at both the multi-teacher-matched and $8\times$ batch sizes. Dashed lines mark teacher accuracies; all OPD runs use $G{=}1$.

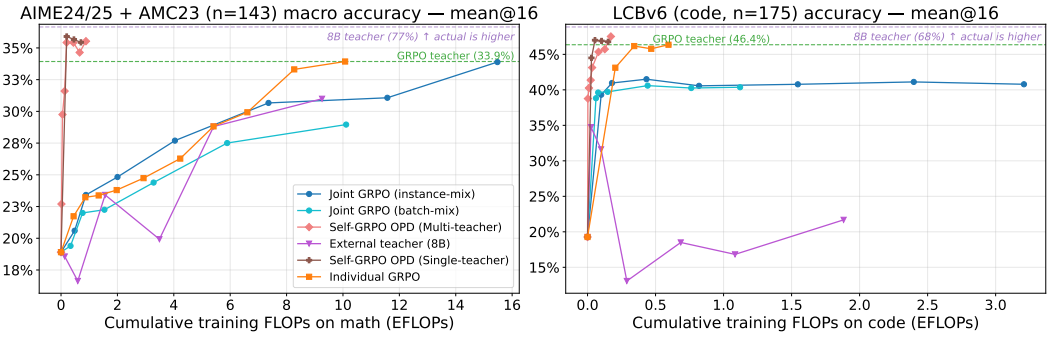



Figure 30: **Per-domain breakdown: 2-domain math + code mix.** Qwen3-4B-Base student, `mean@16` versus training FLOPs per domain (math = AIME24/25 + AMC23, code = LCBv6). Curves shown: *Joint GRPO (instance-mix)*, *Joint GRPO (batch-mix)*, *Individual GRPO* specialists, *Self-GRPO OPD (multi-teachers)*, and *external OPD* with a single Qwen3-8B thinking teacher. Dashed lines mark teacher accuracies; all OPD runs use $G{=}1$.

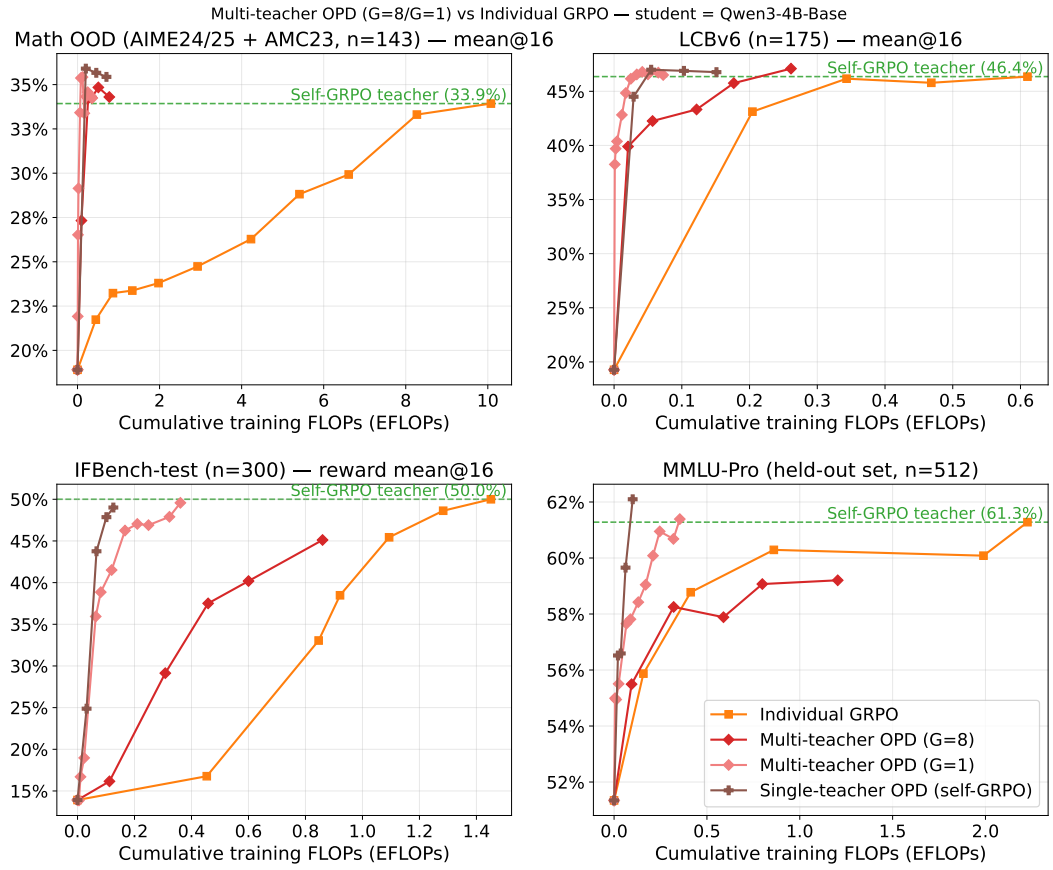



Figure 31: **Rollout group-size ablation for multi-teacher self-GRPO OPD, 4-domain mix.** Qwen3-4B-Base student on math, code, IFBench, and MMLU-Pro. Multi-teacher OPD at $G{=}\{8, 4, 1\}$ against the aggregated individual GRPO ceiling. $G{=}1$ reaches the ceiling at a fraction of the compute of $G{=}8$.

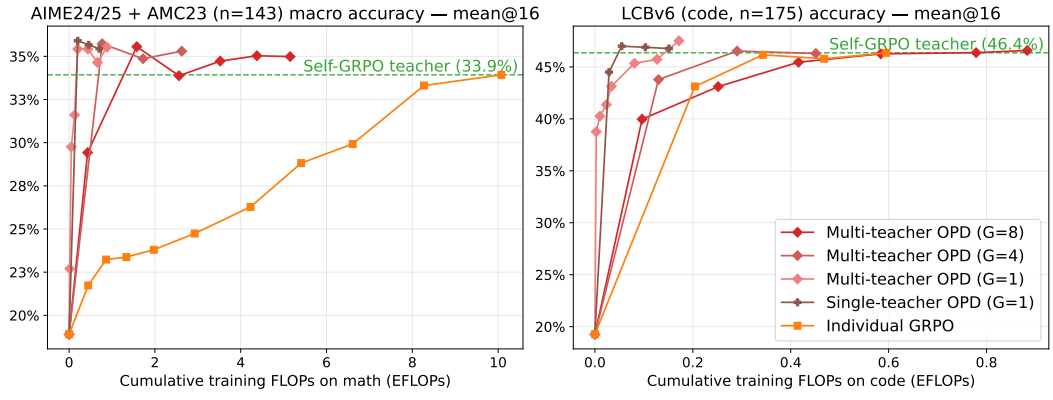



Figure 32: **Rollout group-size ablation for multi-teacher self-GRPO OPD, 2-domain math + code mix.** Same setup as Figure 31 but on the 2-domain subset. $G{=}1$ again reaches the specialist ceiling at substantially lower compute than $G{=}8$.