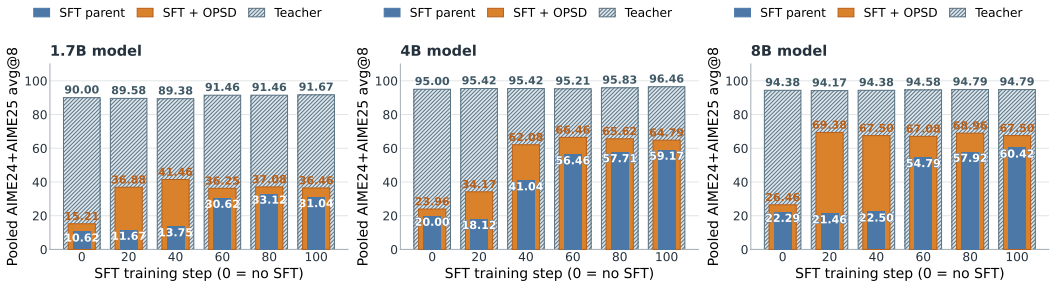
STRONGER, YET NOT BETTER: DEMYSTIFYING SELF-DISTILLATION INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy Self-Distillation (OPSD) efficiently trains LLMs by providing dense, token-level teacher supervision on the student's own rollouts. However, this supervision degrades when student errors push the student's policy too far from the teacher's, making OPSD highly sensitive to initialization: the student checkpoint and teacher configuration that keep supervision reliable and learnable. Prior OPSD methods often assume that stronger students or privileged teachers lead to better distillation. We challenge this assumption with hundreds of experiments spanning model scales, SFT checkpoints, and privileged teacher configurations. At the problem level, higher teacher accuracy on student-failed problems predicts larger post-initialization OPSD gains. At the token level, lower teacher–student divergence predicts larger gains when the student is undertrained, while this relationship weakens or reverses with overfitting. These results show that effective initialization depends not only on model strength, but also on teacher complementarity and learnable teacher–student divergence. Guided by these findings, we introduce a divergence-aware SFT warm-up that upweights low-divergence tokens during OPSD initialization, prioritizing easier-to-learn targets. It outperforms standard SFT+OPSD, yielding larger gains over initialization on AIME24 (+11.67 vs. +5.00) and AIME25 (+11.25 vs. +6.25). Overall, our results suggest that future methods should optimize OPSD initialization for both model capability and teacher–student learnability, rather than simply choosing stronger students or teachers.



(a) Qwen3-1.7B **(b)** Qwen3-4B **(c)** Qwen3-8B

Figure 1: **Stronger SFT parents or privileged teachers do not necessarily improve OPSD performance.** We report SFT performance, the best OPSD checkpoint (steps 25–100), and the corresponding privileged-teacher score.[1]

1 INTRODUCTION

On-policy distillation (OPD) transfers reasoning from a teacher to a student by supervising student-sampled trajectories, providing token-level feedback at states the student actually visits. Recent self-distillation methods—OPSD Zhao et al. (2026), SDFT Shenfeld et al. (2026), and SDPO Hübotter et al. (2026)—use the same model as teacher and student, with the teacher distinguished only by privileged contexts. This on-policy design, however, makes the effectiveness of supervision depend on *where distillation begins*. The teacher is queried on student-generated prefixes, so its supervision depends not only on teacher capability, but also on the student's state distribution and whether the

[1]Teacher scores use training data; SFT/OPSD scores use held-out benchmarks, so they are not comparable.

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.

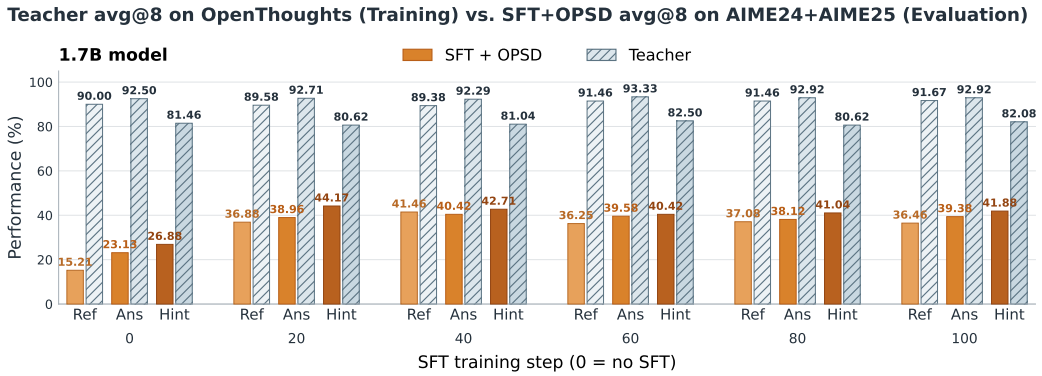


Figure 2: Teacher and student performance for Qwen3-1.7B across privileged-information types. For each SFT checkpoint, we report teacher avg@8 on OpenThoughts and SFT+OPSD student avg@8 on AIME24+AIME25 under reference (*Ref*), answer (*Ans*), and hint (*Hint*) conditioning. **Higher teacher accuracy does not necessarily yield a stronger distilled student.**

resulting teacher signal is accessible to the current student Fu et al. (2026); Jiang et al. (2026); Hübotter et al. (2026).

Recent work addresses several consequences of this dependence. For example, teacher-compatible SFT warm-up can reduce student–teacher mismatch before OPD Liu et al. (2026); Li et al. (2026b), while other methods repair problematic student trajectories or modify the distillation process itself Jiang et al. (2026); Fu et al. (2026). These approaches primarily ask *how to make distillation effective from a chosen starting policy*. They do not directly answer a different question that arises before distillation begins: ***Among possible student checkpoints and privileged teachers, what makes one initialization better suited for subsequent self-distillation?***

Prior works usually start OPSD from stronger models, assuming a better SFT checkpoint or teacher should yield a better result Li et al. (2026b); Liu et al. (2026); Gu et al. (2026). **In practice, neither reliably does.** Across Qwen3-1.7B, 4B, and 8B, stronger SFT checkpoints can leave final performance unchanged or reduce distillation gains (Figure 1), while more accurate teachers do not always produce better students (Figure 2). **We instead characterize OPSD initialization by the student–teacher relationship.**

At the problem level, downstream OPSD gains are linked to teachers whose strengths complement student weaknesses. At the token level, **we find that lower teacher–student divergence predicts larger OPSD gains only at certain training stages.** After a short SFT warm-up, lower divergence correlates with larger gains across all three model scales, but this relationship weakens or reverses with extended SFT (Section 4.3). Thus, token-level alignment is not a reliable criterion for OPSD initialization across stages, suggesting a tradeoff between ease and novelty: close teacher predictions aid undertrained students, while after further SFT, greater agreement may offer little new to learn.

This finding suggests designing SFT warm-up to maximize subsequent OPSD gains: in the short-warm-up regime, lower teacher–student divergence predicts larger improvements. We therefore introduce **Low-Divergence SFT (LD-SFT)**, a divergence-aware initialization that upweights supervised tokens with high teacher–student agreement while keeping the data and targets unchanged, aiming to provide a better initialization for OPSD.

Our experiments show that LD-SFT produces higher OPSD scores even when starting from a weaker SFT checkpoint. With the same warm-up and OPSD training budgets, LD-SFT starts below vanilla SFT on AIME24 but gains 11.7 points during OPSD, versus 5.0 points for vanilla SFT, reaching 72.9 compared with 69.2. The best LD-SFT configurations also improve AIME25 from 60.4 to 63.8, with gains over initialization reaching 13.3 points on HMMT25. Overall, optimizing SFT for downstream distillation yields better final performance than simply starting from a stronger SFT checkpoint.

Our study makes three contributions:

- We systematically examine OPSD initialization across model scales, SFT checkpoints, and multiple forms of privileged teacher context, showing that stronger students and more accurate teachers do not reliably yield better downstream self-distillation.

- We identify teacher–student complementarity and token-level distributional compatibility as useful diagnostics of downstream OPSD gain, and show that the role of divergence changes with the student's training stage.
- We introduce a new initialization procedure, Low-Divergence SFT (LD-SFT), which reweights supervised token losses using teacher–student divergence to improve subsequent OPSD performance across multiple reasoning benchmarks.

2 Background and Related Work

Vulnerability of OPD/OPSD Let a student policy p_S generate $y \sim p_S(\cdot \mid x)$ and let a privileged teacher p_T, conditioned on additional information z, provide token-level supervision along the generated trajectory. Abstractly, OPSD Zhao et al. (2026) minimizes

$$\mathcal{L}_{\text{OPSD}} = \mathbb{E}_{x,y\sim p_S}\left[\sum_t D(p_T(\cdot \mid x, z, y_{<t}), p_S(\cdot \mid x, y_{<t}))\right]. \tag{1}$$

Because supervision is evaluated on student-generated prefixes, its usefulness depends on the student's state distribution (Fu et al., 2026; Shen et al., 2026). Early student errors can produce *prefix failure*, where local teacher supervision becomes ineffective (Jiang et al., 2026), and prefixes that drift outside regions of reliable teacher guidance lead to compounding errors and noisy supervision (Fu et al., 2026; Xu et al., 2026; Li et al., 2026a). Distributional mismatch, unsuitable teacher behavior, and optimization pathologies add further instability (Zhu et al., 2026; Wang et al., 2026), and in OPSD, privileged information can yield non-transferable teacher signals that degrade student reasoning (Shen et al., 2026). Being on-policy is therefore insufficient: effective OPD/OPSD requires compatibility among student trajectories, teacher behavior, and optimization dynamics (Zhu et al., 2026).

Warm-up before on-policy (self-)distillation. A common way to mitigate these vulnerabilities is to warm up the policy before distillation to make the student more *compatible* with the teacher, and this is typically done with standard SFT. We call any stage immediately preceding OPSD the *initialization*, and the resulting policy the *pre-OPSD parent*. Existing warm-ups target different failure sources. Zhu et al. (2026) reduce student–teacher mismatch with SFT on the student and RLVR on the teacher, and Simple-OPD shows that a small LoRA SFT on teacher-compatible chains of thought suffices for effective OPD (Liu et al., 2026). Jiang et al. (2026) address prefix failure by repairing flawed rollouts before distillation, while Kim & Lee (2026) apply OPSD after SFT and RLVR to compress already-acquired reasoning. These results suggest that warm-up helps by making the pre-OPSD parent compatible with teacher guidance. *Yet it remains unclear what makes a good initialization for self-distillation, or how to build one.*

Checkpoint selection for staged post-training. Checkpoint selection, or what we called *initialization*, for later optimization has mainly been studied for SFT→RL. Kang et al. (2026) show that stronger SFT checkpoints can yield worse post-RL performance, with held-out generalization loss and large-k pass@k as better proxies. Zhang et al. (2026) likewise find that stronger offline checkpoints can underperform weaker ones after identical RL, and propose PEAR to weight offline learning by downstream-policy relevance. Intermediate models should thus be judged by how much they benefit from subsequent training, not by their own performance. OPD work, however, focuses on effective distillation and student–teacher mismatch, leaving initialization unaddressed. *We ask which pre-OPSD parent best supports distillation, which teacher-specific properties make it preferable, and how to construct a better OPSD initialization.*

3 Does a Stronger Initialization Lead to Better OPSD?

We study this question by varying student initialization and teacher context in controlled comparisons.

3.1 Experimental setup

Model and Training. We study OPSD initialization across different model scales (i.e., Qwen3-1.7B, Qwen3-4B, and Qwen3-8B (Yang et al., 2025)). We construct pre-OPSD parents from SFT checkpoints taken at different steps, and then apply the same downstream OPSD procedure to each

initialization. SFT and OPSD use the *Openthoughts_math_30k_opsd* training set (Zhao et al., 2026; Guha et al., 2025), with training subsets and example order held fixed across controlled comparisons.

Privileged Teacher Configuration. Following Zhao et al. (2026), the fixed teacher and trainable student start from the same pre-OPSD parent. The student generates non-thinking trajectories, while the teacher operates in thinking mode with one of three forms of privileged information: **Reference**, a worked solution; **Answer**, the target answer only; or **Hint**, a short key insight generated by the corresponding base model.

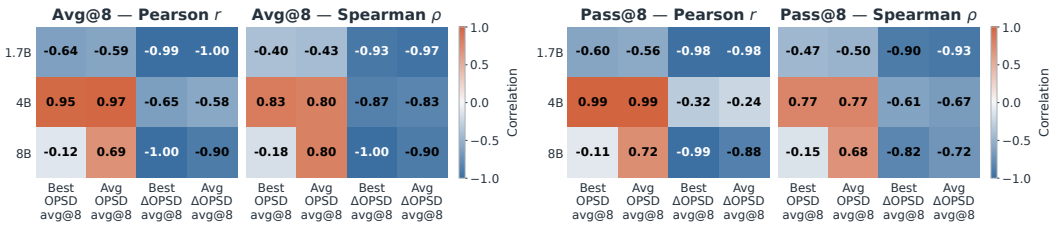
Evaluation. Following prior work Zhao et al. (2026), we evaluate on the pooled 60 problems from AIME 2024 and AIME 2025. We draw $K = 8$ *non-thinking* samples per problem and report $\mathrm{avg@}8$, the mean fraction of correct generations, and $\mathrm{pass@}8$, the fraction of problems solved by at least one generation. For diagnostic analyses, we follow Kang et al. (2026) and report Pearson r and Spearman ρ between pre-OPSD diagnostics and either absolute post-OPSD performance or the gain relative to the corresponding pre-OPSD parent.

Full implementation details (e.g., training hyperparameters, data preprocessing, and evaluation metrics) are provided in Appendix A.

3.2 Results

Here, we examine how SFT initialization affects OPSD training by comparing checkpoints with varying SFT levels in terms of final OPSD performance, OPSD gains, and SFT accuracy. Our findings are summarized below.

A short SFT warm-up stage can help OPSD even when it does not improve SFT accuracy. Results are presented in Figure 1. Across all three model sizes, just 20 SFT steps substantially improve subsequent OPSD performance while producing little SFT-only improvement, or even a slight decline, in the student's own accuracy (for 4B and 8B models). This contrast is most pronounced for Qwen3-8B, where the best post-OPSD accuracy more than doubles, suggesting that a brief warm-up can benefit subsequent distillation well before its value is apparent in standalone performance.



(a) SFT $\mathrm{avg@}8$ **(b)** SFT $\mathrm{pass@}8$

Figure 3: Correlations between SFT performance ((a) $\mathrm{avg@}8$, (b) $\mathrm{pass@}8$) and downstream OPSD scores and gains across model sizes. Each panel shows Pearson r (left) and Spearman ρ (right).

More SFT can improve SFT accuracy while saturating or hurting downstream OPSD. Figure 1 also shows that beyond a certain point, additional SFT raises pre-OPSD accuracy without improving, and sometimes degrading, OPSD performance. Thus, the checkpoint with the highest SFT accuracy is not necessarily the best starting point for OPSD.

SFT accuracy is not a reliable predictor of the benefit from OPSD. We further quantify this relationship using the correlation analysis in Section 3.1. Figure 3 shows that the correlation between SFT accuracy and final OPSD performance varies across model sizes. For SFT avg@8, the Pearson correlation with the best post-OPSD score is -0.64, 0.95, and -0.12 for Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively, showing no consistent relationship. In contrast, higher SFT accuracy is consistently associated with smaller OPSD gains, with correlations of -0.99, -0.65, and -1.00. Appendix B.6 provides evidence that this association is not driven by shared-term coupling. The same pattern holds for pass@8. Overall, while more SFT training provides a stronger starting point, it also leaves less room for improvement with OPSD, making SFT accuracy alone insufficient for selecting the best initialization.

3.3 Do stronger teachers lead to higher OPSD performance?

In this subsection, we examine whether stronger privileged teachers lead to better downstream OPSD performance. We compare teacher accuracy across SFT checkpoints and privileged-information types, and analyze its relationship with the performance of the resulting students. We discuss our key findings below.

Teacher accuracy is stably high across SFT checkpoints, but OPSD performance differs. Figure 1 compares teacher accuracy (with reference as privilege information) and downstream OPSD performance at each SFT checkpoint. For all three model sizes, SFT impacts the teacher accuracy by less than 2 points but shifts OPSD performance by about 20–40 points. From 0 to 100 SFT steps, teacher accuracy moves from 90.0 to 91.7 (Qwen3-1.7B), 95.0 to 96.5 (Qwen3-4B), and 94.4 to 94.8 (Qwen3-8B), while OPSD increases from 15.2 to 36.5, 24.0 to 64.8, and 26.5 to 67.5, respectively.

The strongest privileged teacher is not always the best teacher for distillation. We next vary the privileged information while fixing student initialization. Figure 2 shows Qwen3-1.7B results; Qwen3-4B and Qwen3-8B appear in Appendix Figures 10 and 11. Across settings, the most accurate teacher often does not produce the best OPSD student, as illustrated by hint conditioning: at the 20-step Qwen3-1.7B checkpoint, the hint-conditioned teacher has lower avg@8 (80.62) than reference (89.58) and answer (92.71), yet yields the highest OPSD score, 44.17 versus 36.88 and 38.96, respectively. Thus, task accuracy alone does not make a good distillation target. This pattern aligns with findings from Shen et al. (2026): richer privileged information, such as full reference reasoning, can shift the teacher further from the student's distribution, whereas hints leave more reasoning to the student and preserve closer token-level targets. Section 4.3 tests this via teacher–student divergence.

Teacher accuracy is not consistently correlated with downstream OPSD performance.

Figure 4 makes this pattern more explicit. For the best downstream OPSD score, the Pearson correlation with teacher avg@8 is -0.66 for Qwen3-1.7B, 0.47 for Qwen3-4B, and -0.21 for Qwen3-8B. The Spearman correlations show the same mixed pattern: $\rho = -0.63$, 0.09, and -0.18, respectively. These results suggest that teacher accuracy alone is not enough to judge teacher quality for OPSD. What matters is how the teacher differs from the student and whether the teacher's additional behavior is useful for that student. We study these teacher–student relationships in the next section.

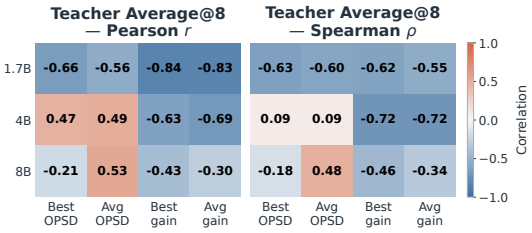


Figure 4: Correlations between privileged-teacher avg@8 and downstream OPSD scores and gains across model sizes. Pearson r (left) and Spearman ρ (right).

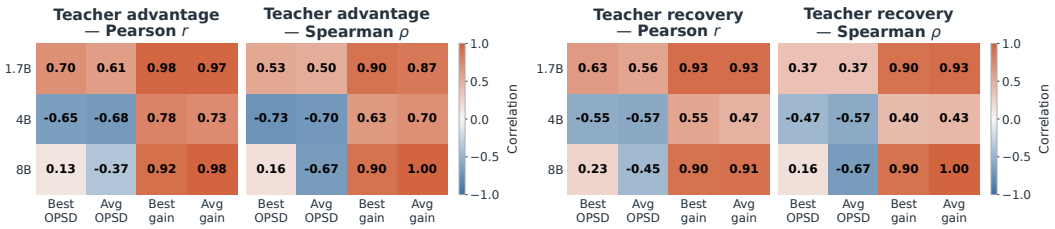
4 Teacher-Student Relationship Predicts OPSD Gain After Initialization

4.1 Experimental Setup

We use the experiment setups from Section 3.1 and additionally characterize each pre-OPSD student–teacher pair with problem- and token-level diagnostics, detailed in Appendix A.7.

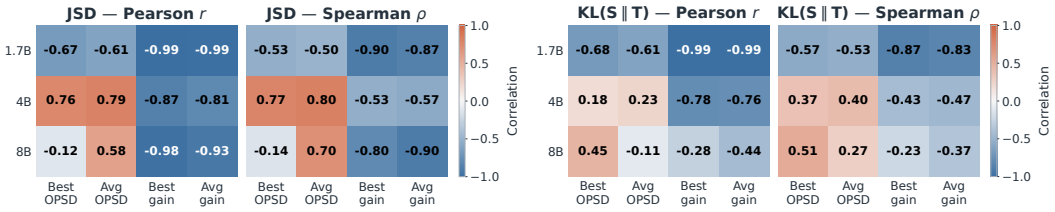
Problem-level complementarity. We evaluate each student–teacher pair on 60 OPSD training problems with eight generations per model. *Teacher advantage* measures the teacher's average gain in problem-level success, while *teacher recovery* emphasizes teacher success where the student is weak. Together, they quantify complementary teacher capability.

Token-level compatibility. We measure distributional mismatch on student-generated trajectories by evaluating both models on the same student prefix, with privileged context provided only to the teacher. We report token-averaged full-vocabulary $D_{\mathrm{KL}}(p^S \| p^T)$ and Jensen–Shannon divergence (JSD): reverse KL highlights large local mismatches, while JSD measures bounded, symmetric disagreement.



(a) Teacher advantage **(b)** Teacher recovery

Figure 5: Problem-level teacher–student complementarity predicts subsequent OPSD gain. We report correlations between (a) teacher advantage and (b) teacher recovery and downstream OPSD performance across model sizes. Each panel reports Pearson r (left) and Spearman ρ (right).



(a) JSD **(b)** $D_{\mathrm{KL}}(p^S \| p^T)$

Figure 6: Token-level teacher–student divergence predicts downstream OPSD gain in the short-SFT regime. We report correlations between (a) JSD and (b) reverse KL $D_{\mathrm{KL}}(p^S \| p^T)$ and downstream OPSD performance across model sizes. Lower divergence is consistently associated with larger OPSD gains, particularly for JSD.

4.2 Problem-Level Complementarity Predicts OPSD Gain

Figure 5 shows that OPSD gains correlate strongly with teacher advantage and recovery. For best OPSD gain, teacher advantage achieves $r = 0.98$, 0.78, and 0.92 for the 1.7B, 4B, and 8B models, while teacher recovery achieves $r = 0.93$, 0.55, and 0.90; similar trends hold across checkpoints. This suggests that distillation benefits depend more on *where the teacher outperforms the student* than on absolute teacher strength. We call this **teacher–student complementarity**: supervision is most valuable when the teacher succeeds where the student is weak.

4.3 The Effect of Token-Level Divergence Depends on the Student's Training Stage

Problem-level complementarity measures what useful capability the teacher has beyond the student, but not whether the student can learn it. We therefore use on-policy teacher–student divergence as a proxy for compatibility. Figure 6 shows that, early in SFT, lower divergence predicts larger OPSD gains: JSD correlates with best OPSD gain at -0.99, -0.87, and -0.98 for the 1.7B, 4B, and 8B models, with reverse KL showing the same, weaker trend (-0.99, -0.78, -0.28). Thus, for undertrained students, lower divergence may indicate more **usable** supervision: the teacher's policy is close enough to the student's current policy for its signal to be effectively exploited. This negative association persists in parent-controlled analyses (Appendix B.6), suggesting it is not solely explained by OPSD gain's shared dependence on pre-OPSD parent performance.

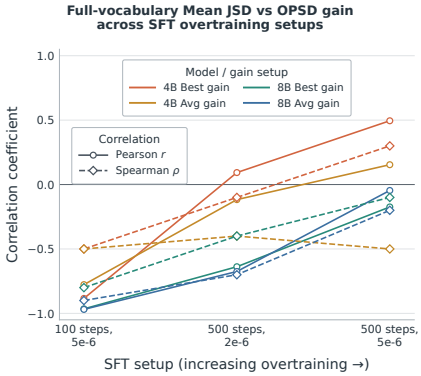


Figure 7: Lower token-level JSD predicts larger OPSD gains early in SFT, but the relationship weakens or reverses later.

Low divergence, however, is not always preferable. Figure 7 shows that this negative relationship weakens with additional SFT and can reverse. For 4B, the Pearson correlation between JSD and best OPSD gain changes from -0.884 at 100 SFT steps to

$+0.093$ and $+0.495$ under the two 500-step settings. For 8B, it weakens from -0.965 to -0.638 and then -0.174.

One possible explanation is that the role of divergence changes as SFT progresses. At an early training stage, lower divergence indicates greater teacher–student compatibility, making the supervision easier to exploit and leading to larger OPSD gains. At a later training stage, the student and teacher become more similar, so very low divergence may provide little new training signal, while a larger gap can offer more novel behavior to learn from. This may explain why the correlation between divergence and OPSD gain shifts from negative toward zero or positive as SFT training proceeds.

These results suggest a tradeoff between **ease** and **novelty**: at an early training stage, lower-divergence supervision is easier to exploit, while at a later training stage, a larger gap may provide more novel learning signal. Effective OPSD therefore requires the teacher to be *sufficiently different to be informative, yet sufficiently close to be learnable*, with the proper balance depending on the student's training stage.

5 How to Shape a Good OPSD Initialization?

Section 4.3 shows that, after short SFT, OPSD yields larger gains when the privileged teacher remains close to the student at the token level. An effective warm-up should therefore do more than maximize SFT accuracy: it should preserve teacher signals that remain accessible to the student. We incorporate this principle directly into SFT. Rather than minimizing teacher–student divergence, which could suppress the very signal OPSD exploits, we use divergence to reweight each supervised token's contribution to the SFT objective. The resulting method, **LD-SFT**, assigns greater weight to tokens for which the teacher is already well aligned with the student.

5.1 Low-Divergence Supervised Fine-Tuning (LD-SFT)

We introduce **LD-SFT**, a variant of standard SFT that reweights token-level cross-entropy by the agreement between the current policy and a privileged teacher, while leaving prompts, reasoning traces, and targets unchanged. This distinguishes it from existing OPD warm-ups, which modify *what* the student is trained on—for example, teacher-compatible chains of thought (Liu et al., 2026), joint student SFT and teacher RLVR (Zhu et al., 2026), or repaired rollouts (Jiang et al., 2026). LD-SFT instead changes only *how* a fixed supervised signal is distributed across tokens. All weighting quantities are detached, so LD-SFT is **not** a KL/JSD distillation objective like Elcock et al. (2026) and does not alter the targets.

For each target token t, we compute

$$D_t = \mathrm{JSD}(p_T(\cdot \mid s_t, z),\, p_\theta(\cdot \mid s_t))\,,$$

where p_θ is the student and p_T is the same model run without gradients and conditioned on the reference solution z. D_t thus measures how far the student is from a solution-informed prediction.

Because divergences are heavy-tailed, we replace D_t with its percentile rank $\hat{D}_t$ among the batch's supervised tokens, define a reachability score $R_t = \exp(-\hat{D}_t/\tau)$, and form centered weights

$$w_t = \max\bigl(1 + \alpha(R_t - \bar{R}),\, w_{\min}\bigr)\,,$$

where α sets the reweighting strength and $w_{\min}$ bounds the suppression of high-divergence tokens. Rank normalization keeps the weights insensitive to the scale of D_t as training proceeds, and $\alpha{=}0$ exactly recovers vanilla SFT. High-divergence tokens also tend to have higher cross-entropy, so naive reweighting shrinks the gradient. We therefore renormalize to preserve the loss scale:

$$\mathcal{L}_{\text{LD-SFT}} = \frac{1}{N}\sum_t \tilde{w}_t\,\mathrm{CE}_t, \qquad \tilde{w}_t = w_t\,\frac{\sum_j \mathrm{CE}_j}{\sum_j w_j\,\mathrm{CE}_j},$$

with the normalizer's CE terms detached. This keeps the effective step size close to vanilla SFT (Figure 9), so downstream differences can be attributed to token allocation rather than to a changed learning rate.

5.2 Empirical Evaluation

5.2.1 Experiment Setup

We follow the experiment setup of Section 3.1. All SFT variants are trained for 100 steps, matching the short-warm-up regime where lower teacher–student divergence predicts larger downstream OPSD gains. We evaluate on AIME24, AIME25, and HMMT25 using *nonthinking* mode of Qwen3 models.

We compare LD-SFT ($\alpha \in 0.25, 1$) against two groups of baselines. The first group modifies the SFT stage: **Vanilla SFT** (standard SFT), **Random-SFT** (divergence-independent reweighting), **Easy-SFT** (weighting by student-side difficulty), and **High-Divergence SFT (HD-SFT)** (high-divergence weighting). The second group keeps the Vanilla SFT initialization and modifies distillation instead, using **EMA OPSD** Zhao et al. (2026) and **Top-k OPSD** Zhao et al. (2026). This tests whether changing OPSD alone can match the gains from a better SFT initialization. Full baseline definitions are in Appendix C.1.

5.2.2 Main Results

Table 1 and Figure 8 show that low-divergence SFT initialization consistently improves subsequent OPSD. The best LD-SFT+OPSD configuration achieves avg@8 scores of **72.92**, **63.75**, and **43.33** on AIME24, AIME25, and HMMT25, versus 69.17, 60.42, and 39.58 for Vanilla SFT+OPSD. While the optimal α varies by benchmark (0.25 on AIME24; 1 on AIME25 and HMMT25), both settings remain competitive with or outperform standard initialization throughout OPSD.

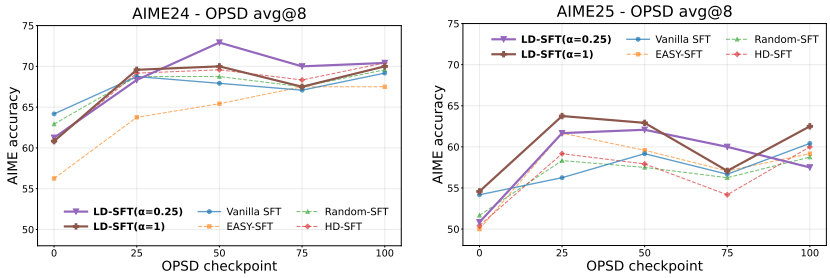


Figure 8: Comparison of downstream OPSD avg@8 performance across SFT variants on AIME24 and AIME25. We compare Vanilla SFT, EASY-SFT, Random-SFT, high-divergence SFT (HD-SFT), and low-divergence SFT (LD-SFT) with $\alpha = 0.25$ and $\alpha = 1$ across OPSD checkpoints.

More importantly, LD-SFT makes the parent benefit more from subsequent OPSD, rather than merely improving parent accuracy. On AIME24, the $\alpha = 0.25$ parent underperforms Vanilla SFT (61.25 vs. 64.17 avg@8), yet after identical OPSD training gains **+11.67** points versus **+5.00**, reversing the gap (72.92 vs. 69.17). Similar patterns hold on AIME25 (**+11.25** vs. **+6.25**) and HMMT25, where $\alpha = 1$ achieves the largest OPSD gain (**+13.33**). Thus, a strong OPSD initialization is characterized less by its initial accuracy than by how much useful learning subsequent distillation can extract from it. Results across three training seeds for Vanilla SFT and both LD-SFT settings are reported in Appendix C.3 (Table 28).

Comparisons with other SFT variants suggest that the key is how warm-up allocates the learning signal. Easy-SFT achieves large OPSD gains (+11.25 on AIME24, +11.67 on AIME25) but remains weaker overall because of its lower SFT score. LD-SFT instead preserves stronger initial capability while retaining room for OPSD improvement. Thus, the goal is not simply to maximize OPSD gain, but to start from a strong parent that can still improve substantially during OPSD.

Finally, we compare two OPSD variants that improve **teacher–student alignment**: Top-k OPSD, which distills only high-probability teacher tokens, and EMA OPSD, which uses a slowly moving teacher. From Vanilla SFT, Top-k OPSD reaches 71.25/62.08/41.67 on AIME24/AIME25/HMMT25, while LD-SFT + standard OPSD reaches 72.92/63.75/43.33. Thus, improving teacher–student alignment during OPSD helps, but shaping the initialization can be even more effective.

Table 1: Best OPSD results with reference conditioning; all SFT variants use 100 SFT warm-up steps, and the best OPSD checkpoint is selected from steps 25–100. Δ: avg@8 gain over the pre-OPSD checkpoint; gray brackets show 95% paired-bootstrap CIs. **Bold**: best; underlined: second best.

	AIME 2024		AIME 2025		HMMT 2025	
Checkpoint	avg@8	Δ	avg@8	Δ	avg@8	Δ
No SFT (Base)	22.92	–	17.08	–	11.67	–
+ OPSD	27.50	+4.58 [−0.42, +10.00]	21.25	+4.17 [+0.00, +7.50]	15.00	+3.33 [+0.83, +6.67]
Vanilla SFT	64.17	–	54.17	–	29.58	–
+ OPSD	69.17	+5.00 [−0.42, +10.83]	60.42	+6.25 [−0.42, +12.92]	39.58	+10.00 [+2.92, +17.92]
+ EMA OPSD	70.00	+5.83 [+0.00, +12.08]	58.75	+4.58 [−2.92, +12.08]	39.17	+9.58 [+4.17, +15.83]
+ Top-k OPSD	71.25	+7.08 [+2.08, +12.92]	62.08	+7.92 [+2.08, +14.58]	41.67	+12.08 [+5.42, +19.58]
Random-SFT	62.92	–	51.67	–	31.67	–
+ OPSD	69.58	+6.66 [+0.83, +12.50]	58.75	+7.08 [+1.67, +13.33]	41.25	+9.58 [+4.58, +15.42]
Easy-SFT	56.25	–	50.00	–	34.58	–
+ OPSD	67.50	+11.25 [+5.42, +17.50]	61.67	**+11.67** [+2.92, +20.83]	40.42	+5.83 [+1.67, +10.83]
HD-SFT	60.83	–	50.42	–	35.00	–
+ OPSD	70.42	+9.59 [+2.92, +17.08]	60.00	+9.58 [+3.33, +16.67]	38.75	+3.75 [−1.67, +10.00]
LD-SFT (α=0.25)	61.25	–	50.83	–	30.42	–
+ OPSD	**72.92**	**+11.67** [+4.58, +19.17]	62.08	+11.25 [+5.42, +17.92]	39.58	+9.17 [+1.67, +16.67]
LD-SFT (α=1)	60.83	–	54.58	–	30.00	–
+ OPSD	70.00	+9.17 [+0.42, +19.17]	**63.75**	+9.17 [+3.33, +15.00]	**43.33**	**+13.33** [+6.25, +21.25]

5.2.3 Training Dynamics and Hyperparameter Analysis

Figure 9 shows that both LD-SFT variants closely follow Vanilla SFT in training loss, as expected from our loss-preserving renormalization. Interestingly, $\alpha = 1$ stays closer to Vanilla SFT than $\alpha = 0.25$ and also has a smaller gradient norm. One possible explanation is that stronger reweighting downweights high-divergence tokens more aggressively, leading to smaller updates. Thus, increasing α appears to make optimization more conservative while largely preserving the overall loss scale.

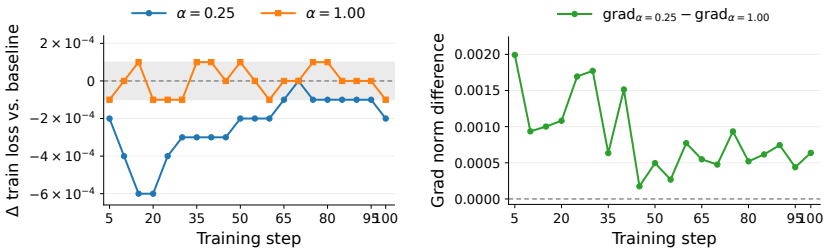


Figure 9: Effect of low-divergence reweighting strength on training. Left: loss differences from baseline remain small. Right: gradient norms are consistently larger for $\alpha = 0.25$ than $\alpha = 1.00$.

Table 1 shows that more conservative weighting is not always better: $\alpha = 0.25$ performs best on AIME24, while $\alpha = 1$ is best on the harder AIME25 and HMMT25 benchmarks. This suggests that the optimal reweighting strength depends on task difficulty: easier tasks may benefit from preserving more of the broad SFT signal, whereas harder tasks may require stronger emphasis on teacher behavior reachable by the student.

Thus, α balances preserving the broad SFT signal against emphasizing teacher-compatible tokens, with stronger bias becoming more useful on harder tasks. More broadly, our results show that a good OPSD initialization is not defined by SFT loss alone, but by how well it prepares the student to benefit from subsequent self-distillation.

6 Conclusion

A good OPSD initialization is not simply the strongest student or most accurate teacher. Gains are larger when the teacher succeeds where the student fails and provides supervision close enough for the student to use. This reveals a training-dependent tradeoff between informativeness and usability. Building on this insight, we introduce LD-SFT, a divergence-aware warm-up that outperforms standard SFT for OPSD. Overall, initialization should be chosen by expected OPSD improvement, not standalone model strength.

REPRODUCIBILITY STATEMENT

We provide detailed information needed to reproduce our experiments in Appendix A. This includes the model checkpoints and SFT initialization procedure (Appendix A.1), training data and data splits (Appendix A.2), privileged-teacher construction and OPSD training configuration (Appendix A.3), and the evaluation protocol (Appendix A.4). We specify optimization hyperparameters, sampling parameters, checkpoint-selection procedures, and the definitions of the teacher–student diagnostics used in our analyses. Definitions of the LD-SFT control objectives and alternative OPSD variants are provided in Appendix C.1. **We released the code: https://anonymous.4open.science/r/iclr-opsd-init-2C1B/**

ETHICS STATEMENT

Our experiments use publicly available pretrained language models and mathematical reasoning datasets and benchmarks. The study does not involve human subjects or the collection of private or sensitive personal data. Our experiments focus on mathematical reasoning and do not evaluate deployment in high-stakes application domains. As with model distillation methods more generally, improvements in model training may have downstream uses beyond those evaluated in this work.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing manuscript text and LaTeX presentation. All AI-assisted text and experimental artifacts were reviewed by the authors. The authors take responsibility for the final content of the paper, including text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

James Elcock, William F. Shen, Xinchi Qiu, and Nicholas D. Lane. How llm task-adaptation reshapes alignment: A multi-dimensional study of behavioral and representational drift, 2026. URL https://arxiv.org/abs/2607.22676.

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. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=uZbZb058bA.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models, 2026. URL https://arxiv.org/abs/2306.08543.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL https://arxiv.org/abs/2506.04178.

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

Li Jiang, Haoran Xu, Yichuan Ding, and Amy Zhang. Trajectory-refined distillation, 2026. URL https://arxiv.org/abs/2606.08432.

Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in SFT-RL post-training: When high SFT scores mislead and what to use instead. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uLM3BfKo19.

Jaehoon Kim and Dongha Lee. Opsd compresses what rlvr teaches: A post-rl compaction stage for reasoning models, 2026. URL https://arxiv.org/abs/2605.06188.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, and Jinqiao Wang. On-policy distillation with curriculum turn-level guidance for multi-turn agents, 2026a. URL https://arxiv.org/abs/2606.15912.

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

Tao Liu, Taiqiang Wu, Mao Zheng, Xuan Luo, Runming Yang, Xuewei Yang, Junjie Wang, and Yujiu Yang. Simple-opd: Demystifying warm-up for on-policy distillation, 2026. URL https://arxiv.org/abs/2608.06802.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, Gang Chen, and Jieping Ye. Purified opsd: On-policy self-distillation without losing how to think, 2026. URL https://arxiv.org/abs/2607.02234.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=qA6FgH0nnZ.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations, 2026. URL https://arxiv.org/abs/2607.13399.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Dylan Zhang, Yufeng Xu, Haojin Wang, Qingzhi Chen, and Hao Peng. Good SFT optimizes for SFT, better SFT prepares for reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=gSGnknLZED.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

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

APPENDIX CONTENTS

A Experimental Setup and Diagnostic Definitions **14**

A.1 Models and SFT Initialization 14

A.2 Training Data 14

A.3 Privileged Teachers and OPSD Training 14

A.4 Evaluation Protocol 15

A.5 Evaluation Mode and Thinking-Student Evaluation 15

A.6 Checkpoint Selection and Downstream Quantities 16

A.7 Teacher–Student Diagnostic Details 16

B Detailed Selection Results **17**

B.1 Correlation Analysis 17

B.2 Underfitting Settings 18

B.3 Average Results Across Privileged-Information Settings 20

B.4 Overfitting Settings 20

B.5 Comparison Across Different Privileged OPSD Settings 22

B.6 Robustness to Parent-Performance Coupling 23

C LD-SFT Implementation and Control Experiments **26**

C.1 Control Objectives and OPSD Variants 28

C.2 Control Experiment Setup and Checkpoint-Averaged Results 29

C.3 Multi-Seed SFT and OPSD Evaluation 30

A Experimental Setup and Diagnostic Definitions

This appendix provides the full experimental setup used throughout our initialization and self-distillation analyses. Unless otherwise specified, all controlled comparisons use the same training data, example ordering, optimization setup, and evaluation protocol.

A.1 Models and SFT Initialization

We use Qwen/Qwen3-1.7B, Qwen/Qwen3-4B, and Qwen/Qwen3-8B (Yang et al., 2025). We generally follow the experimental setup of Zhao et al. (2026). SFT warm-up runs are trained for $\{100, 500\}$ steps, with checkpoints saved at every 1/5 of the total training steps.

Training uses LoRA with rank 64 and scale 128 on the attention and MLP projections, learning rate $\{2 \times 10^{-6}, 5 \times 10^{-6}\}$, effective batch size 32, bfloat16, and maximum sequence length 16,000.

These intermediate SFT checkpoints serve as pre-OPSD parents for subsequent self-distillation experiments, allowing us to compare initializations obtained at different stages of supervised warm-up while holding the downstream OPSD procedure fixed.

A.2 Training Data

Warm-up and OPSD use the train split of *siyanzhao/Openthoughts_math_30k_opsd*, the dataset provided by Zhao et al. (2026), which contains 29,434 mathematical reasoning examples.

For SFT, we construct a seeded 99/1 train/evaluation split. For controlled comparisons across different SFT initializations and self-distillation settings with different forms of privileged information, we use the same training subset and example order for SFT and OPSD. Thus, differences across conditions arise from the initialization or privileged-teacher configuration rather than from changes in the underlying training examples.

A.3 Privileged Teachers and OPSD Training

Following the central construction of Zhao et al. (2026), the fixed teacher and trainable student begin from the same pre-OPSD parent. The student samples nonthinking trajectories, while the teacher operates in thinking mode with access to additional privileged context.

We consider three forms of teacher-side privileged information: **Reference**, in which the teacher is given a worked solution; **Answer**, in which the teacher is given only the target answer; and **Hint**, in which the teacher is given a short directional insight generated by the corresponding Qwen3 base model. The exact teacher-side context construction is described in Appendix A.3, and the hint-generation and filtering procedure is described in Appendix A.3.

Across these conditions, the underlying teacher–student construction remains fixed. Only the form of privileged context provided to the teacher and the corresponding prompt framing are changed.

The teacher is evaluated on the exact assistant-token prefixes sampled by the student. In particular, teacher distributions are conditioned on the realized student-generated prefix rather than on independently sampled teacher trajectories.

The downstream OPSD training setup is held fixed across initializations and privileged-teacher conditions. We train for 100 updates and save checkpoints every 25 steps. Training uses an effective batch size of 32, learning rate 5×10^{-6}, LoRA rank 64 and scale 128, and a 1,024-token completion cap. Student trajectories are sampled with temperature 1.1, top-$p = 0.95$, and top-$k = 20$.

This setup inherits the on-policy, fixed-teacher, full-vocabulary distillation design and the main training hyperparameters of Zhao et al. (2026). Unless otherwise specified, all comparisons use the same downstream OPSD training configuration.

Privileged teacher context implementation. Across all OPSD variants, the student sees only the problem and generates in nonthinking mode, while the teacher scores the same student-generated trajectory in thinking mode with additional privileged context. The student and teacher otherwise use the same problem framing and final-answer instruction.

For **Reference**, the teacher receives the full worked solution. For **Answer**, it receives only the gold final answer and is instructed to reconstruct the reasoning independently. For **Hint**, it receives a short directional hint and is instructed to use it to guide its own derivation. These variants share the same training and data-collation pipeline and differ only in the privileged text provided to the teacher.

Hint generation and filtering. Hints are generated offline by the corresponding Qwen3 base model rather than by an external teacher model. Given the problem and its reference solution, the model is prompted to produce a one–two sentence hint identifying the key insight, theorem, substitution, or intermediate step, without providing a complete derivation or the final answer.

Before training, we check that every hint is nonempty and correctly aligned with the training subset. We also remove leading formatting artifacts such as `Hint:` when doing so leaves the hint content intact, and discard generations that terminate inside a Qwen3 thinking block. Finally, we audit hints for accidental answer leakage using explicit `\boxed` markers and boundary-matched occurrences of the gold answer. Because this heuristic can produce false positives, leakage is recorded for analysis rather than used as an automatic exclusion criterion. The processed hints are then stored and reused unchanged throughout training.

A.4 Evaluation Protocol

We evaluate on the pooled AIME 2024 and AIME 2025 benchmarks. The evaluation set contains all 30 problems from the AIME 2024 train split (`HuggingFaceH4/aime_2024`) and all 30 problems from the AIME 2025 train split (`yentinglin/aime_2025`), for a total of $N = 60$ problems.

For each problem, we draw $K = 8$ nonthinking samples at temperature 1.0 and top-$p = 0.8$, with a maximum generation length of 30,000 tokens.

For each generated response, we extract the last well-formed, potentially nested boxed answer and grade mathematical equivalence using *math_verify*. When parsing fails, we fall back to normalized exact matching.

Let $i \in \{1, \dots, N\}$ index problems and $k \in \{1, \dots, K\}$ index sampled generations. Let $b_{ik}^S \in \{0, 1\}$ indicate whether student generation k for problem i is correct. We define the problem-level student success rate as

$$a_i^S = \frac{1}{K} \sum_{k=1}^{K} b_{ik}^S. \tag{2}$$

We report

$$\text{avg@}\,K(S) = \frac{1}{N} \sum_{i=1}^{N} a_i^S, \qquad \text{pass@}\,K(S) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[a_i^S > 0\right]. \tag{3}$$

Throughout the main experiments, $K = 8$, so these quantities are reported as $\text{avg@}\,8$ and $\text{pass@}\,8$.

Our evaluation protocol differs from the thinking-mode, 12-sample evaluation reported by Zhao et al. (2026). In particular, we use nonthinking generation and eight samples per problem. Consequently, absolute results from the two studies are not directly comparable under an identical evaluation protocol.

A.5 Evaluation Mode and Thinking-Student Evaluation

We use Qwen3 in *non-thinking* mode throughout the main SFT, OPSD, and evaluation pipeline. This provides a more sensitive setting for measuring warm-up effects, since thinking-mode models are already substantially stronger and leave less headroom for short SFT warm-ups. It also keeps the student's inference regime consistent across training and evaluation.

Our OPSD configuration follows the finding from the original OPSD paper (Zhao et al., 2026) that using a non-thinking student with a thinking teacher performs better than enabling thinking for the student. We adopt this student–teacher setting, but evaluate the resulting checkpoints in non-thinking mode for consistency with SFT and OPSD training, rather than switching to thinking-mode evaluation

Table 2: Thinking-student OPSD evaluated in thinking mode at Avg@8.

Stage	AIME24 Avg@8	AIME25 Avg@8	Pooled Avg@8
Qwen3-4B base	71.25	62.92	67.08
SFT parent ck100	62.50	55.42	58.96
OPSD 25	53.75	48.75	51.25
OPSD 50	55.00	52.50	53.75
OPSD 75	55.42	52.08	53.75
OPSD 100	57.50	50.42	53.96

as in the original OPSD setup from (Zhao et al., 2026). The privileged teacher therefore remains in thinking mode, as described in Section 3.1.

We additionally test an alternative in which the OPSD student trains and is evaluated in thinking mode. Results are shown in Table 2.

All thinking-mode OPSD checkpoints underperform the SFT parent on both AIME24 and AIME25, further supporting our use of a non-thinking student in the main experiments.

A.6 Checkpoint Selection and Downstream Quantities

For each pre-OPSD initialization, we evaluate the downstream OPSD checkpoints saved at steps 25, 50, 75, and 100.

We consider both the best-performing OPSD checkpoint and performance averaged across the saved OPSD checkpoints. For an evaluation metric m, the improvement of checkpoint c relative to its corresponding pre-OPSD parent S_0 is

$$\Delta m_c = m(S_c) - m(S_0), \tag{4}$$

where S_c denotes the student after OPSD checkpoint c.

Accordingly, we report both absolute post-OPSD performance and the gain relative to the corresponding pre-OPSD initialization. In the main text, these quantities are primarily reported using $\mathrm{avg@}\,8$.

A.7 Teacher–Student Diagnostic Details

This appendix provides the full definitions and measurement procedures for the teacher–student diagnostics introduced in Section 4.1. These diagnostics characterize each pre-OPSD student and its corresponding privileged teacher at both the problem level and the token level.

A.7.1 Problem-Level Complementarity

We measure two complementary aspects of the teacher's problem-level behavior on 60 problems from the OPSD training corpus, using eight student and eight teacher generations per problem. The student uses non-thinking sampling with top-$p = 0.8$, whereas the privileged thinking teacher uses top-$p = 0.95$; both use temperature 1.0 and a 16,384-token generation cap.

For teacher rung r, let $a_i^{T_r}$ denote its success rate on problem i over the eight teacher generations, and let a_i^S denote the corresponding student success rate. We define

$$A_r = \frac{1}{N}\sum_{i=1}^{N}\left(a_i^{T_r} - a_i^S\right), \qquad \text{(teacher advantage)} \tag{5}$$

$$R_r = \sum_{i=1}^{N} w_i a_i^{T_r}, \qquad w_i = \frac{1 - a_i^S}{\sum_{j=1}^{N}(1 - a_j^S)}, \qquad \text{(teacher recovery).} \tag{6}$$

Teacher advantage measures the teacher's average performance improvement over the paired student. Teacher recovery instead emphasizes problems on which the student is weak: problems with lower

student success rates receive larger weight. Thus, the two metrics capture complementary notions of teacher–student compatibility—whether the teacher is stronger than the student on average, and whether its strength is concentrated where the student has the most room to improve.

These quantities are computed on the OPSD training corpus and are therefore training-corpus diagnostics rather than held-out benchmark scores. We report a metric as undefined when its denominator is zero.

A.7.2 Token-Level Compatibility

Problem-level success does not reveal whether the teacher's token-level supervision is reachable from the student's current policy. We therefore measure teacher–student distributional divergence directly on student-generated trajectories.

For student checkpoint c and teacher configuration r, we sample

$$y \sim S_c(\cdot \mid x)$$

and evaluate both models at every assistant-token position using the same realized student prefix $y_{<t}$. Only the teacher receives privileged context z_r. Specifically,

$$p_t^T = p_{T_{c,r}}(\cdot \mid x, z_r, y_{<t}), \qquad p_t^S = p_{S_c}(\cdot \mid x, y_{<t}).$$

At metric temperature 1.0, we compute full-vocabulary reverse KL divergence and Jensen–Shannon divergence. The pooled reverse KL is

$$\begin{aligned} \overline{\mathrm{KL}}(S\|T) &= \frac{1}{M} \sum_{(x,y,t)} D_{\mathrm{KL}}\big(p_t^S \| p_t^T\big) \\ &= \frac{1}{M} \sum_{(x,y,t)} \sum_{v \in \mathcal{V}} p_t^S(v) \log \frac{p_t^S(v)}{p_t^T(v)}, \end{aligned} \tag{7}$$

where $\mathcal{V}$ denotes the full vocabulary.

We additionally compute the pooled Jensen–Shannon divergence

$$\begin{aligned} \overline{\mathrm{JSD}} &= \frac{1}{M} \sum_{(x,y,t)} J_t, \\ J_t &= \frac{1}{2} D_{\mathrm{KL}}\big(p_t^T \| m_t\big) + \frac{1}{2} D_{\mathrm{KL}}\big(p_t^S \| m_t\big), \\ m_t &= \frac{1}{2}\left(p_t^T + p_t^S\right). \end{aligned} \tag{8}$$

Here M is the total number of evaluated assistant tokens. The reported statistics therefore pool tokens across sequences rather than first averaging within each sequence. Unless otherwise specified, all distributions are evaluated over the full vocabulary $\mathcal{V}$ using natural logarithms.

The reverse KL $D_{\mathrm{KL}}(p_t^S \| p_t^T)$ is particularly sensitive to positions where the student assigns substantial probability to tokens that the teacher considers unlikely. Because reverse KL is unbounded, however, its pooled mean can be dominated by a small number of high-disagreement positions.

We therefore report Jensen–Shannon divergence alongside reverse KL. JSD is bounded and symmetric and thus provides a complementary measure of overall distributional mismatch. Comparing the two helps distinguish broad teacher–student disagreement from sparse, high-magnitude mismatches.

B Detailed Selection Results

B.1 Correlation Analysis

To characterize which properties of the pre-OPSD student–teacher pair are associated with downstream OPSD performance, we compute correlation statistics separately for the Reference, Answer, and Hint privileged-teacher conditions.

For each model size and diagnostic metric, we report both the Pearson correlation coefficient r, which measures linear association, and the Spearman rank correlation coefficient ρ, which measures monotonic association, with two types of downstream quantities: the absolute post-OPSD performance (*OPSD avg@8*) and the improvement relative to the corresponding pre-OPSD parent (*OPSD gain*, or Δ *OPSD avg@8*). We consider both the best-performing OPSD checkpoint and performance averaged across the saved OPSD checkpoints.

To summarize trends across the three privileged-teacher conditions, for each matched model-size, diagnostic-metric, and downstream-target combination, we additionally average the condition-specific correlation coefficients:

$$\bar{r} = \frac{1}{k}\sum_{j=1}^{k} r_j, \qquad \bar{\rho} = \frac{1}{k}\sum_{j=1}^{k} \rho_j, \tag{9}$$

where $k = 3$ when the statistic is available for Reference, Answer, and Hint.

B.2 Underfitting Settings

We first consider a deliberately short SFT warm-up consisting of 100 optimization steps with learning rate 5×10^{-6}. This regime is intended to capture an *underfitting* setting in which the student has only partially adapted to the supervised data. We report correlations between candidate selection statistics and both the resulting OPSD performance and the gain over the corresponding SFT initialization. For each quantity, we report Pearson correlation r and Spearman rank correlation ρ.

Table 3: $\text{OPSD}_{reference}$ Correlation Metrics for SFT Performance

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	SFT Average@8	$-0.580/-0.100$	$-0.429/-0.100$	$-0.989/-0.900$	$-0.997/-1.000$
	SFT Pass@8	$-0.474/-0.200$	$-0.314/-0.200$	$-0.958/-0.800$	$-0.976/-0.900$
Qwen3-4B	SFT Average@8	$+0.946/+0.600$	$+0.975/+0.600$	$-0.705/-0.900$	$-0.581/-0.900$
	SFT Pass@8	$+0.998/+0.975$	$+0.984/+0.975$	$-0.383/-0.564$	$-0.236/-0.564$
Qwen3-8B	SFT Average@8	$-0.311/-0.359$	$+0.776/+0.700$	$-0.999/-1.000$	$-0.925/-0.900$
	SFT Pass@8	$-0.397/-0.763$	$+0.803/+0.667$	$-0.995/-0.821$	$-0.897/-0.718$

Table 4: $\text{OPSD}_{reference}$ Correlation Metrics for Teacher Performance

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	Teacher Average@8	$-0.697/-0.667$	$-0.582/-0.667$	$-0.994/-0.821$	$-0.992/-0.718$
	Teacher advantage	$+0.668/+0.300$	$+0.506/+0.300$	$+0.992/+1.000$	$+0.989/+0.900$
	Teacher recovery	$+0.524/+0.100$	$+0.323/+0.100$	$+0.931/+0.900$	$+0.937/+1.000$
Qwen3-4B	Teacher Average@8	$+0.279/-0.154$	$+0.317/-0.154$	$-0.642/-0.667$	$-0.719/-0.667$
	Teacher advantage	$-0.687/-1.000$	$-0.754/-1.000$	$+0.743/+0.500$	$+0.604/+0.500$
	Teacher recovery	$-0.639/-1.000$	$-0.683/-1.000$	$+0.485/+0.500$	$+0.319/+0.500$
Qwen3-8B	Teacher Average@8	$-0.323/-0.289$	$+0.868/+0.821$	$-0.937/-0.975$	$-0.780/-0.872$
	Teacher advantage	$+0.016/+0.103$	$-0.456/-0.600$	$+0.898/+0.900$	$+0.996/+1.000$
	Teacher recovery	$+0.258/+0.103$	$-0.509/-0.600$	$+0.843/+0.900$	$+0.869/+1.000$

Table 5: $\text{OPSD}_{reference}$ Correlation Metrics for Student–Teacher Divergence and Agreement

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	JSD divergence	$-0.706/-0.300$	$-0.573/-0.300$	$-0.999/-1.000$	$-0.995/-0.900$
	KL(S‖T)	$-0.710/-0.400$	$-0.579/-0.400$	$-0.999/-0.900$	$-0.995/-0.800$
Qwen3-4B	JSD divergence	$+0.741/+1.000$	$+0.813/+1.000$	$-0.884/-0.500$	$-0.777/-0.500$
	KL(S‖T)	$+0.477/+0.900$	$+0.572/+0.900$	$-0.885/-0.600$	$-0.799/-0.600$
Qwen3-8B	JSD divergence	$-0.105/+0.051$	$+0.648/+0.800$	$-0.965/-0.800$	$-0.969/-0.900$
	KL(S‖T)	$+0.511/+0.462$	$-0.011/+0.500$	$-0.553/-0.600$	$-0.827/-0.800$

Across privileged-information settings, the most consistent relationship is between the *gain* from OPSD and measures of student–teacher mismatch. In particular, mean JSD is strongly negatively correlated with OPSD gain in several of the 100-step conditions. This pattern is especially clear for Qwen3-1.7B and Qwen3-8B. Intuitively, when the warm-start student remains far from the privileged teacher, subsequent distillation is less effective; conversely, warm-start checkpoints that already place the student closer to the teacher tend to realize larger gains from OPSD.

Table 6: OPSD$_{answer}$ Correlation Metrics for SFT Performance

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	SFT Average@8	$-0.444/-0.400$	$-0.631/-0.700$	$-0.998/-1.000$	$-0.999/-1.000$
	SFT Pass@8	$-0.380/-0.300$	$-0.569/-0.500$	$-0.981/-0.900$	$-0.984/-0.900$
Qwen3-4B	SFT Average@8	$+0.955/+1.000$	$+0.968/+0.900$	$-0.540/-0.900$	$-0.448/-0.800$
	SFT Pass@8	$+0.995/+0.667$	$+0.988/+0.667$	$-0.183/-0.564$	$-0.081/-0.718$
Qwen3-8B	SFT Average@8	$-0.849/-0.872$	$+0.626/+0.800$	$-1.000/-1.000$	$-0.904/-0.900$
	SFT Pass@8	$-0.783/-0.658$	$+0.655/+0.410$	$-0.989/-0.821$	$-0.879/-0.718$

Table 7: OPSD$_{answer}$ Correlation Metrics for Teacher Performance

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	Teacher Average@8	$-0.697/-0.667$	$-0.582/-0.667$	$-0.994/-0.821$	$-0.992/-0.718$
	Teacher advantage	$+0.657/+0.700$	$+0.748/+0.900$	$+0.980/+0.900$	$+0.972/+0.900$
	Teacher recovery	$+0.677/+0.400$	$+0.838/+0.700$	$+0.947/+1.000$	$+0.942/+1.000$
Qwen3-4B	Teacher Average@8	$+0.279/-0.154$	$+0.317/-0.154$	$-0.642/-0.667$	$-0.719/-0.667$
	Teacher advantage	$-0.594/-0.600$	$-0.616/-0.500$	$+0.728/+0.700$	$+0.700/+0.900$
	Teacher recovery	$-0.234/+0.200$	$-0.259/-0.100$	$+0.311/-0.100$	$+0.247/+0.000$
Qwen3-8B	Teacher Average@8	$-0.323/-0.289$	$+0.868/+0.821$	$-0.937/-0.975$	$-0.780/-0.872$
	Teacher advantage	$+0.937/+0.975$	$-0.315/-0.600$	$+0.941/+0.900$	$+0.995/+1.000$
	Teacher recovery	$+0.939/+0.975$	$-0.448/-0.600$	$+0.966/+0.900$	$+0.955/+1.000$

Table 8: OPSD$_{answer}$ Correlation Metrics for Student–Teacher Divergence

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	JSD divergence	$-0.474/-0.700$	$-0.631/-0.900$	$-0.993/-0.900$	$-0.992/-0.900$
	KL(S‖T)	$-0.487/-0.700$	$-0.632/-0.900$	$-0.991/-0.900$	$-0.989/-0.900$
Qwen3-4B	JSD divergence	$+0.761/+0.600$	$+0.792/+0.700$	$-0.797/-0.500$	$-0.721/-0.600$
	KL(S‖T)	$-0.510/-0.400$	$-0.474/-0.300$	$-0.484/+0.000$	$-0.535/-0.100$
Qwen3-8B	JSD divergence	$-0.942/-0.975$	$+0.511/+0.700$	$-0.974/-0.800$	$-0.929/-0.900$
	KL(S‖T)	$+0.228/+0.564$	$-0.840/-0.300$	$+0.692/+0.700$	$+0.422/+0.600$

Table 9: OPSD$_{hint}$ Correlation Metrics for SFT Performance

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	SFT Average@8	$-0.884/-0.700$	$-0.721/-0.500$	$-0.998/-0.900$	$-0.996/-0.900$
	SFT Pass@8	$-0.939/-0.900$	$-0.810/-0.800$	$-0.994/-1.000$	$-0.996/-1.000$
Qwen3-4B	SFT Average@8	$+0.963/+0.900$	$+0.965/+0.900$	$-0.712/-0.800$	$-0.717/-0.800$
	SFT Pass@8	$+0.992/+0.667$	$+0.990/+0.667$	$-0.397/-0.718$	$-0.405/-0.718$
Qwen3-8B	SFT Average@8	$+0.788/+0.700$	$+0.664/+0.900$	$-1.000/-1.000$	$-0.878/-0.900$
	SFT Pass@8	$+0.847/+0.975$	$+0.693/+0.975$	$-0.988/-0.821$	$-0.848/-0.718$

Table 10: OPSD$_{hint}$ Correlation Metrics for Teacher Performance

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	Teacher Average@8	$-0.592/-0.564$	$-0.501/-0.462$	$-0.523/-0.205$	$-0.522/-0.205$
	Teacher advantage	$+0.768/+0.600$	$+0.572/+0.300$	$+0.961/+0.800$	$+0.953/+0.800$
	Teacher recovery	$+0.678/+0.600$	$+0.514/+0.300$	$+0.914/+0.800$	$+0.910/+0.800$
Qwen3-4B	Teacher Average@8	$+0.852/+0.564$	$+0.850/+0.564$	$-0.611/-0.821$	$-0.625/-0.821$
	Teacher advantage	$-0.664/-0.600$	$-0.669/-0.600$	$+0.880/+0.700$	$+0.884/+0.700$
	Teacher recovery	$-0.773/-0.600$	$-0.776/-0.600$	$+0.853/+0.800$	$+0.860/+0.800$
Qwen3-8B	Teacher Average@8	$+0.008/+0.051$	$-0.135/-0.205$	$+0.593/+0.564$	$+0.658/+0.718$
	Teacher advantage	$-0.565/-0.600$	$-0.331/-0.800$	$+0.908/+0.900$	$+0.963/+1.000$
	Teacher recovery	$-0.491/-0.600$	$-0.400/-0.800$	$+0.900/+0.900$	$+0.905/+1.000$

Table 11: OPSD$_{hint}$ Correlation Metrics for Student–Teacher Divergence and Agreement

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	JSD divergence	$-0.842/-0.600$	$-0.638/-0.300$	$-0.986/-0.800$	$-0.978/-0.800$
	KL(S‖T)	$-0.842/-0.600$	$-0.633/-0.300$	$-0.985/-0.800$	$-0.976/-0.800$
Qwen3-4B	JSD divergence	$+0.765/+0.700$	$+0.771/+0.700$	$-0.919/-0.600$	$-0.920/-0.600$
	KL(S‖T)	$+0.586/+0.600$	$+0.594/+0.600$	$-0.958/-0.700$	$-0.958/-0.700$
Qwen3-8B	JSD divergence	$+0.683/+0.500$	$+0.594/+0.600$	$-0.990/-0.800$	$-0.905/-0.900$
	KL(S‖T)	$+0.624/+0.500$	$+0.526/+0.600$	$-0.972/-0.800$	$-0.922/-0.900$

The relationship with absolute OPSD performance is less uniform across model sizes. For Qwen3-4B, stronger SFT performance is positively associated with stronger final OPSD performance, whereas this relationship is substantially weaker or changes sign for some Qwen3-1.7B and Qwen3-8B conditions. This distinction motivates separating checkpoint quality as measured by *absolute performance* from checkpoint suitability for subsequent distillation as measured by *OPSD gain*.

B.3 Average Results Across Privileged-Information Settings

To summarize the 100-step results, we additionally average each correlation across the reference, answer, and hint privileged-information settings.

Table 12: Average Correlation Metrics for SFT Performance Across Different Privileged-Teacher Information Settings

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	SFT Average@8	$-0.636/-0.400$	$-0.594/-0.433$	$-0.995/-0.933$	$-0.997/-0.967$
	SFT Pass@8	$-0.598/-0.467$	$-0.564/-0.500$	$-0.978/-0.900$	$-0.985/-0.933$
Qwen3-4B	SFT Average@8	$+0.955/+0.833$	$+0.969/+0.800$	$-0.652/-0.867$	$-0.582/-0.833$
	SFT Pass@8	$+0.995/+0.770$	$+0.987/+0.770$	$-0.321/-0.615$	$-0.241/-0.667$
Qwen3-8B	SFT Average@8	$-0.124/-0.177$	$+0.689/+0.800$	$-1.000/-1.000$	$-0.902/-0.900$
	SFT Pass@8	$-0.111/-0.149$	$+0.717/+0.684$	$-0.991/-0.821$	$-0.875/-0.718$

Table 13: Average Correlation Metrics for Privileged-Teacher Performance Across Different Information Settings

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	Teacher Average@8	$-0.662/-0.633$	$-0.555/-0.599$	$-0.837/-0.616$	$-0.835/-0.547$
	Teacher advantage	$+0.698/+0.533$	$+0.609/+0.500$	$+0.978/+0.900$	$+0.971/+0.867$
	Teacher recovery	$+0.626/+0.367$	$+0.558/+0.367$	$+0.931/+0.900$	$+0.930/+0.933$
Qwen3-4B	Teacher Average@8	$+0.470/+0.085$	$+0.495/+0.085$	$-0.632/-0.718$	$-0.688/-0.718$
	Teacher advantage	$-0.648/-0.733$	$-0.680/-0.700$	$+0.784/+0.633$	$+0.729/+0.700$
	Teacher recovery	$-0.549/-0.467$	$-0.573/-0.567$	$+0.550/+0.400$	$+0.475/+0.433$
Qwen3-8B	Teacher Average@8	$-0.213/-0.176$	$+0.534/+0.479$	$-0.427/-0.462$	$-0.301/-0.342$
	Teacher advantage	$+0.129/+0.159$	$-0.367/-0.667$	$+0.916/+0.900$	$+0.985/+1.000$
	Teacher recovery	$+0.235/+0.159$	$-0.452/-0.667$	$+0.903/+0.900$	$+0.910/+1.000$

Table 14: Average Correlation Metrics for Student–Privileged-Teacher Divergence and Agreement Across Different Information Settings

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-1.7B	JSD divergence	$-0.674/-0.533$	$-0.614/-0.500$	$-0.993/-0.900$	$-0.988/-0.867$
	KL(S‖T)	$-0.680/-0.567$	$-0.615/-0.533$	$-0.992/-0.867$	$-0.987/-0.833$
Qwen3-4B	JSD divergence	$+0.756/+0.767$	$+0.792/+0.800$	$-0.867/-0.533$	$-0.806/-0.567$
	KL(S‖T)	$+0.184/+0.367$	$+0.231/+0.400$	$-0.776/-0.433$	$-0.764/-0.467$
Qwen3-8B	JSD divergence	$-0.121/-0.141$	$+0.584/+0.700$	$-0.976/-0.800$	$-0.934/-0.900$
	KL(S‖T)	$+0.454/+0.509$	$-0.108/+0.267$	$-0.278/-0.233$	$-0.442/-0.367$

The averaged results reinforce the distinction between predicting *final OPSD performance* and predicting *OPSD gain*. For Qwen3-4B, SFT performance is a strong predictor of final OPSD performance, but it is negatively correlated with OPSD gain. By contrast, teacher-advantage and teacher-recovery statistics are generally positively associated with gain. Student–teacher JSD exhibits an especially strong negative association with gain for Qwen3-1.7B and Qwen3-8B.

B.4 Overfitting Settings

We next test whether the same selection signals remain informative after substantially longer supervised warm-up. We train for 500 SFT steps using learning rates 2×10^{-6} and 5×10^{-6}. Relative to the 100-step setting, these configurations move the student further toward the SFT objective and therefore serve as potential *overfitting* regimes for the purposes of warm-start selection.

The longer-warm-up results show that the strong negative association between student–teacher divergence and OPSD gain is primarily a property of the underfit regime rather than a universal

Table 15: OPSD$_{reference}$ (500 steps, lr=2×10^{-6}) Correlation Metrics for SFT Performance

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-4B	SFT Average@8	$+0.663/+0.658$	$+0.702/+0.658$	$-0.860/-0.821$	$-0.919/-0.975$
	SFT Pass@8	$+0.727/+0.649$	$+0.226/+0.081$	$+0.118/+0.369$	$-0.251/-0.369$
Qwen3-8B	SFT Average@8	$-0.746/-0.700$	$-0.916/-0.100$	$-0.998/-0.900$	$-0.999/-1.000$
	SFT Pass@8	$-0.519/-0.783$	$+0.188/+0.224$	$-0.169/-0.783$	$-0.101/-0.447$

Table 16: OPSD$_{reference}$ (500 steps, lr=2×10^{-6}) Correlation Metrics for Teacher Performance

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-4B	Teacher Average@8	$-0.037/-0.103$	$+0.347/+0.205$	$+0.057/0.000$	$+0.275/+0.300$
	Teacher advantage	$-0.214/-0.564$	$+0.303/+0.051$	$+0.001/-0.200$	$+0.314/+0.500$
	Teacher recovery	$-0.210/-0.103$	$+0.072/+0.205$	$+0.214/0.000$	$+0.396/+0.300$
Qwen3-8B	Teacher Average@8	$-0.110/-0.100$	$+0.466/+0.500$	$+0.377/0.000$	$+0.427/-0.100$
	Teacher advantage	$+0.564/+0.300$	$+0.874/+0.600$	$+0.907/+0.400$	$+0.924/+0.300$
	Teacher recovery	$+0.432/+0.300$	$+0.787/+0.600$	$+0.821/+0.400$	$+0.844/+0.300$

Table 17: OPSD$_{reference}$ (500 steps, lr=2×10^{-6}) Correlation Metrics for Student–Teacher Divergence and Agreement

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-4B	JSD divergence	$+0.637/+0.616$	$+0.404/+0.462$	$+0.093/-0.100$	$-0.117/-0.400$
	KL(S‖T)	$+0.657/+0.616$	$+0.381/+0.462$	$+0.113/-0.100$	$-0.123/-0.400$
Qwen3-8B	JSD divergence	$-0.293/-0.300$	$-0.787/-0.600$	$-0.638/-0.400$	$-0.675/-0.700$
	KL(S‖T)	$+0.502/+0.600$	$+0.185/0.000$	$+0.426/+0.500$	$+0.391/+0.100$

Table 18: OPSD$_{reference}$ (500 steps, lr=5×10^{-6}) Correlation Metrics for SFT Performance

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-4B	SFT Average@8	$+0.447/+0.700$	$+0.490/+0.308$	$-0.673/-0.600$	$-0.869/-1.000$
	SFT Pass@8	$+0.725/+0.667$	$+0.948/+0.947$	$+0.323/+0.359$	$+0.234/-0.154$
Qwen3-8B	SFT Average@8	$+0.541/+0.395$	$+0.502/+0.051$	$-0.557/-0.564$	$-0.600/-0.872$
	SFT Pass@8	$+0.504/+0.433$	$+0.087/0.000$	$+0.664/+0.791$	$+0.243/+0.264$

Table 19: OPSD$_{reference}$ (500 steps, lr=5×10^{-6}) Correlation Metrics for Teacher Performance

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-4B	Teacher Average@8	$+0.835/+0.667$	$+0.827/+0.947$	$+0.526/+0.359$	$+0.289/-0.154$
	Teacher advantage	$-0.072/-0.564$	$+0.243/+0.105$	$+0.681/+0.564$	$+0.952/+0.872$
	Teacher recovery	$+0.529/+0.200$	$+0.511/+0.564$	$+0.849/+0.900$	$+0.742/+0.500$
Qwen3-8B	Teacher Average@8	$-0.777/-0.526$	$-0.182/+0.154$	$-0.565/-0.872$	$+0.028/-0.205$
	Teacher advantage	$-0.498/-0.553$	$-0.280/-0.103$	$-0.391/-0.359$	$-0.160/-0.205$
	Teacher recovery	$-0.539/-0.667$	$+0.153/+0.200$	$-0.559/-0.300$	$+0.115/+0.400$

Table 20: OPSD$_{reference}$ (500 steps, lr=5×10^{-6}) Correlation Metrics for Student–Teacher Divergence and Agreement

Model	Metric	Best OPSD r/ρ	Avg OPSD r/ρ	Best gain r/ρ	Avg gain r/ρ
Qwen3-4B	JSD divergence	$+0.925/+0.900$	$+0.794/+0.872$	$+0.495/+0.300$	$+0.154/-0.500$
	KL(S‖T)	$+0.918/+0.700$	$+0.774/+0.975$	$+0.532/+0.400$	$+0.190/-0.200$
Qwen3-8B	JSD divergence	$+0.735/+0.975$	$+0.900/+0.600$	$-0.174/-0.100$	$-0.046/-0.200$
	KL(S‖T)	$+0.694/+0.975$	$+0.890/+0.600$	$-0.230/-0.100$	$-0.069/-0.200$

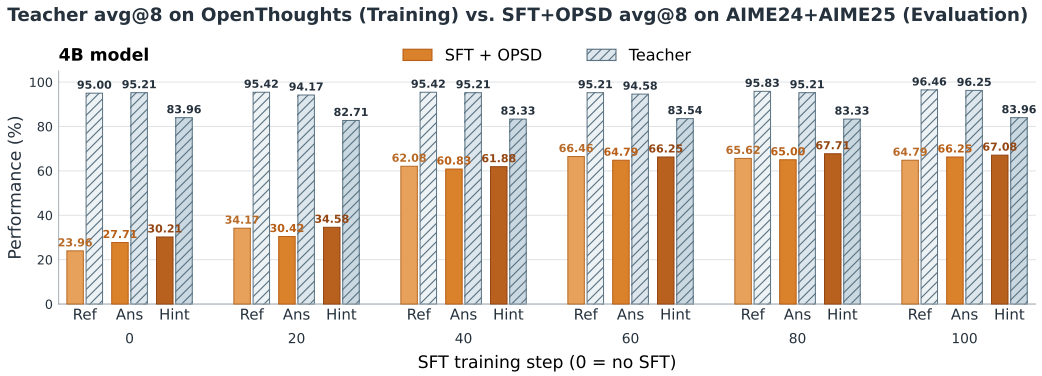


Figure 10: Cross-setting comparison for Qwen3-4B. The figure summarizes the relationship between warm-start selection statistics and downstream OPSD outcomes across the reference, answer, and hint privileged-information settings.

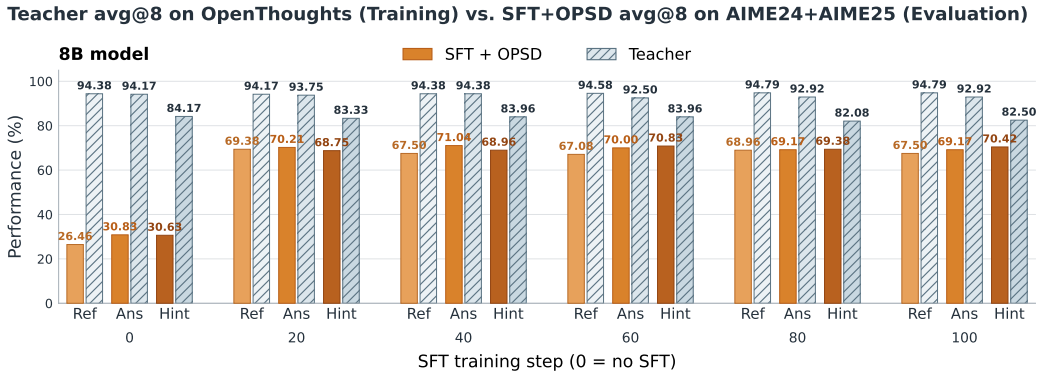


Figure 11: Cross-setting comparison for Qwen3-8B. The figure summarizes the relationship between warm-start selection statistics and downstream OPSD outcomes across the reference, answer, and hint privileged-information settings.

relationship. As the amount of SFT increases, this correlation weakens and can even reverse for Qwen3-4B. The effect is less abrupt for Qwen3-8B, but the magnitude of the negative correlation also decreases as warm-up becomes more aggressive. Thus, divergence is most useful as a selection signal while the student remains in the regime where additional supervised adaptation improves its compatibility with the privileged teacher.

B.5 Comparison Across Different Privileged OPSD Settings

We next compare the selection behavior across the reference, answer, and hint privileged-information settings. The goal is to determine whether the observed relationships are tied to one particular construction of privileged information or remain qualitatively similar across different privileged-teacher inputs.

For the 100-step warm-up, mean JSD is strongly negatively associated with gain for both Qwen3-4B and Qwen3-8B. This relationship weakens substantially after 500 SFT steps. For Qwen3-4B, the Pearson correlation even becomes positive under the more aggressive warm-up settings, while for Qwen3-8B it remains negative but approaches zero at learning rate 5×10^{-6}. These results support using student–teacher divergence as a warm-start selection criterion specifically in the short, underfit warm-up regime considered in our main experiments, rather than treating it as a general-purpose predictor across arbitrary amounts of SFT.

Table 21: Pearson r: full-vocabulary Mean JSD vs. gain

Regime	4B Best gain	4B Avg gain	8B Best gain	8B Avg gain
100 steps, 5×10^{-6}	-0.884	-0.777	-0.965	-0.969
500 steps, 2×10^{-6}	+0.093	-0.117	-0.638	-0.675
500 steps, 5×10^{-6}	+0.495	+0.154	-0.174	-0.046

Table 22: Spearman ρ: full-vocabulary Mean JSD vs. gain

Regime	4B Best gain	4B Avg gain	8B Best gain	8B Avg gain
100 steps, 5×10^{-6}	-0.500	-0.500	-0.800	-0.900
500 steps, 2×10^{-6}	-0.100	-0.400	-0.400	-0.700
500 steps, 5×10^{-6}	+0.300	-0.500	-0.100	-0.200

B.6 Robustness to Parent-Performance Coupling

Our main analysis relates properties of the pre-OPSD teacher–student pair to downstream OPSD gain. Let

$$S = \text{pre-OPSD performance}, \qquad P = \text{post-OPSD performance},$$

and define

$$G = P - S.$$

Because S enters G directly, correlations with G can be amplified when a diagnostic also covaries with parent performance. We therefore perform robustness analyses in the 100-step SFT regime.

We use two controls. First, for the relationship between parent performance S and gain G, we estimate the two quantities from disjoint evaluation samples: either disjoint generations (*generation split*) or disjoint evaluation problems (*problem split*). This removes shared evaluation noise.

Second, for a teacher–student diagnostic X, such as teacher advantage or JSD, we analyze absolute post-OPSD performance P while conditioning on S, rather than controlling the constructed outcome $G = P - S$. We report the partial correlation

$$r(X, P \mid S),$$

obtained by residualizing both X and P with respect to S.

Because each model–teacher ladder contains only five SFT checkpoints, we also report exact Freedman–Lane permutation tests for per-ladder partial correlations. Pooled regressions include model and teacher fixed effects, allow parent-performance slopes to vary across model sizes, and use CR1 cluster-robust standard errors clustered by parent checkpoint.

Overall, the negative parent-performance–gain relationship persists under independent evaluation. Teacher advantage retains a positive parent-controlled association in eight of nine settings, although the controlled relationship is weaker than the raw gain correlations. Likewise, lower teacher–student JSD remains directionally associated with stronger downstream OPSD performance in eight of nine settings, with the clearest controlled signal for the reference-conditioned teacher used to motivate LD-SFT.

Higher parent accuracy is associated with smaller OPSD gain. Because $G = P - S$, sampling noise in S appears with a negative sign in G. We therefore recompute the correlation using independent estimates of the two quantities. Table 23 shows that the negative relationship remains under both generation-level and problem-level splitting in every 100-step setting.

For Qwen3-8B with the reference-conditioned teacher, the raw correlation is

$$r(S, G) = -0.925.$$

It remains -0.916 with disjoint generations and -0.912 with disjoint problem subsets. For Qwen3-1.7B, raw correlations near -1 similarly remain approximately -0.97 to -0.98 after independent

Table 23: Independent-sample robustness of the relationship between parent performance and OPSD gain in the 100-step SFT regime. S denotes parent avg@8 and $G = P - S$ denotes OPSD gain. The generation-split column estimates S and G from disjoint generation subsets, while the problem-split column uses disjoint evaluation problems. The relationship remains negative across all nine settings.

Model	**Teacher**	**Raw** $r(S, G)$	**Generation split**	**Problem split**	**Slope** $P \sim S$
1.7B	Reference	-0.997	-0.979	-0.972	-0.04
1.7B	Answer	-0.999	-0.981	-0.975	-0.04
1.7B	Hint	-0.996	-0.978	-0.973	-0.10
4B	Reference	-0.581	-0.538	-0.519	$+0.86$
4B	Answer	-0.447	-0.419	-0.391	$+0.89$
4B	Hint	-0.717	-0.683	-0.660	$+0.78$
8B	Reference	-0.925	-0.916	-0.912	$+0.34$
8B	Answer	-0.904	-0.898	-0.893	$+0.27$
8B	Hint	-0.878	-0.871	-0.864	$+0.33$

estimation. Thus, shared evaluation noise alone does not explain the relationship between stronger parents and smaller incremental OPSD gains.

A smaller gain does not imply worse final performance. For example, the slope of P with respect to S is positive for the 4B settings (approximately 0.78–0.89), so stronger parents can reach higher final performance while receiving smaller OPSD gains.

Teacher advantage remains informative after controlling for parent performance. Raw correlations between teacher advantage and OPSD gain are positive in all nine settings, ranging from $+0.603$ to $+0.996$. Because weaker parents can leave more room for the privileged teacher to outperform them, we instead test whether teacher advantage predicts absolute post-OPSD performance after conditioning on S.

Table 24 shows that the partial correlation remains positive in eight of nine settings. For Qwen3-8B, the partial correlations are

$$+0.978, \qquad +0.985, \qquad +0.826$$

for the reference-, answer-, and hint-conditioned teachers, respectively, with exact permutation p-values of 0.01 and 0.02 for the first two.

When all 100-step settings are pooled,

$$\beta_{\text{adv}} = +1.249, \qquad \text{CR1 SE} = 0.593, \qquad p = 0.054.$$

Thus, larger teacher advantage remains positively associated with post-OPSD performance after parent control, although $p = 0.054$ is just above the conventional 0.05 threshold. The controlled results therefore suggest that the raw gain correlations overstate the independent contribution of teacher advantage, while the diagnostic still retains information beyond parent accuracy in most settings.

Lower JSD remains directionally associated with better downstream OPSD performance. The raw correlation between teacher–student JSD and OPSD gain is negative in all nine 100-step settings. Because JSD also varies strongly with parent training stage, we examine its association with P after conditioning on parent performance.

Table 25 shows that the partial correlation remains negative in eight of nine settings. For Qwen3-4B, the controlled correlations are

$$-0.720, \qquad -0.846, \qquad -0.913,$$

and for Qwen3-8B they are

$$-0.799, \qquad -0.507, \qquad -0.491.$$

The only sign reversal is for the Qwen3-1.7B hint-conditioned teacher; the Qwen3-1.7B answer-conditioned result is close to zero (-0.051).

Table 24: Teacher advantage after controlling for parent performance in the 100-step SFT regime. X denotes teacher advantage, S denotes pre-OPSD parent avg@8, P denotes mean post-OPSD performance, and $G = P - S$. The raw correlation $r(X, G)$ measures the association with gain, while $r(X, P \mid S)$ measures the association with absolute downstream performance after removing the linear effect of parent accuracy. Permutation p-values correspond to the parent-controlled partial correlation.

Model	**Teacher**	$r(X, G)$	$r(X, P \mid S)$	**Perm.** p
1.7B	Reference	$+0.989$	$+0.535$	0.44
1.7B	Answer	$+0.972$	$+0.680$	0.34
1.7B	Hint	$+0.953$	-0.834	0.19
4B	Reference	$+0.603$	$+0.278$	0.72
4B	Answer	$+0.700$	$+0.614$	0.38
4B	Hint	$+0.884$	$+0.741$	0.31
8B	Reference	$+0.996$	$+0.978$	0.01
8B	Answer	$+0.994$	$+0.985$	0.02
8B	Hint	$+0.963$	$+0.826$	0.12

Table 25: Teacher–student JSD after controlling for parent performance in the 100-step SFT regime. The raw JSD–gain relationship is negative in all settings. After conditioning on parent avg@8, the partial correlation remains negative in eight of nine settings. Because JSD also varies strongly with parent training stage, these conditional estimates should be interpreted as associations rather than causal effects.

Model	**Teacher**	$r(\mathrm{JSD}, G)$	$r(\mathrm{JSD}, P \mid S)$	**Perm.** p
1.7B	Reference	-0.995	-0.992	0.01
1.7B	Answer	-0.992	-0.051	0.96
1.7B	Hint	-0.978	$+0.860$	0.15
4B	Reference	-0.776	-0.720	0.28
4B	Answer	-0.722	-0.846	0.23
4B	Hint	-0.920	-0.913	0.20
8B	Reference	-0.968	-0.799	0.21
8B	Answer	-0.930	-0.507	0.40
8B	Hint	-0.905	-0.491	0.40

Pooling all teacher conditions gives

$$\beta_{\mathrm{JSD}} = -13.585, \qquad \mathrm{CR1\ SE} = 7.782, \qquad p = 0.103.$$

The negative coefficient is consistent with lower divergence being associated with better downstream performance, but uncertainty remains substantial. We therefore interpret this as a conditional association, not evidence that divergence is independently causal.

Reference-teacher divergence provides the clearest parent-controlled low-divergence signal. The reference-conditioned teacher is particularly relevant because it is the configuration used to motivate and evaluate LD-SFT. Its parent-controlled JSD correlations are

$$-0.992, \qquad -0.720, \qquad -0.799$$

for Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively, giving a consistent direction across model sizes.

Pooling the reference-conditioned runs while allowing parent-performance slopes to vary by model gives

$$\beta_{\mathrm{JSD}} = -22.096, \qquad \mathrm{CR1\ SE} = 9.468, \qquad p = 0.035.$$

Thus, conditional on parent performance, larger JSD is associated with lower post-OPSD performance, or equivalently, lower divergence with better downstream performance. The $p = 0.035$ result provides stronger evidence for a conditional association than the all-teacher pooled analysis.

Table 26: Pooled parent-controlled regressions for the 100-step SFT regime. The dependent variable is mean post-OPSD performance P. Models include model/teacher fixed effects where applicable and model-specific parent-performance slopes. CR1 standard errors are clustered by parent checkpoint. A positive coefficient indicates that larger values of the diagnostic are associated with higher downstream performance after parent control; a negative coefficient indicates the reverse.

Scope	Predictor	β_X	CR1 SE	p	Interpretation
All teachers	Teacher advantage	$+1.249$	0.593	0.054	Positive conditional association; suggestive pooled evidence.
All teachers	JSD	-13.585	7.782	0.103	Negative direction, but substantial uncertainty after parent control.
Reference only	JSD	-22.096	9.468	0.035	Negative conditional association with stronger statistical support.

Table 26 reports the pooled regressions. Because teacher advantage and JSD have different scales, their coefficient magnitudes are not directly comparable; the relevant directions are positive for teacher advantage and negative for JSD.

Connection to LD-SFT. These controls refine rather than eliminate the motivation for divergence-aware initialization. The raw JSD–gain correlations are partly coupled to parent performance, and the all-teacher controlled estimate remains uncertain. However, in the reference-conditioned 100-step regime used by our method, the low-divergence association is negative across all three model sizes and in the pooled parent-controlled regression ($\beta_{\mathrm{JSD}} = -22.096$, $p = 0.035$).

We therefore treat the correlation analysis as motivation for an intervention rather than a causal conclusion. LD-SFT tests this hypothesis directly by reweighting supervised tokens according to teacher–student divergence while leaving the targets unchanged. The LD-SFT+OPSD experiments in Section C provide the complementary intervention-level test of whether exploiting this low-divergence signal yields a more trainable OPSD initialization.

C LD-SFT Implementation and Control Experiments

LD-SFT implementation. We implement LD-SFT as a modification of the same prompt–completion CoT-SFT pipeline used by Vanilla SFT. Unless otherwise specified, all data processing, supervised targets, optimizer settings, LoRA configuration, batch size, sequence length, and checkpoint schedule are identical between Vanilla SFT and LD-SFT. In particular, LD-SFT changes only the token-level weighting of the supervised cross-entropy objective.

For all LD-SFT experiments in the main paper, we use

$$\tau = 1, \qquad w_{\min} = 0.8,$$

with percentile-rank normalization of the teacher–student divergence. Specifically, for all supervised target tokens in a minibatch, we rank their Jensen–Shannon divergences and map the ranks to $\hat{D}_t \in [0, 1]$. Tied divergence values receive their average rank. The resulting reachability score is

$$R_t = \exp(-\hat{D}_t/\tau).$$

We use the centered "excess" weight parameterization

$$w_t = \max\left(1 + \alpha(R_t - \bar{R}),\, w_{\min}\right),$$

where $\bar{R}$ is the mean reachability over supervised target tokens. We evaluate $\alpha \in \{0.25, 1.0\}$ in the main experiments. Setting $\alpha = 0$ produces unit weights and exactly recovers the Vanilla CoT-SFT objective.

Loss-preserving normalization. Because teacher–student divergence and token cross-entropy can be correlated, simply normalizing the weights to have mean one can change the total weighted loss and thereby introduce an unintended change in effective optimization scale. We therefore use

loss-preserving renormalization. Given detached per-token cross-entropies CE_t, the final weights are

$$\tilde{w}_t = w_t \frac{\sum_j \mathrm{CE}_j}{\sum_j w_j \mathrm{CE}_j}.$$

The cross-entropy values used to compute this normalization factor are detached from the computation graph. Likewise, the JSD values, reachability scores, and token weights are fully detached. Thus, gradients flow only through the ordinary supervised cross-entropy, and LD-SFT does not directly optimize a KL or JSD distillation loss.

Privileged teacher construction. The privileged teacher is not a separately trained or frozen model. At every LD-SFT update, the teacher and student use the same current model parameters and the same LoRA adapters. The student receives the ordinary SFT prompt, whereas the teacher receives the same problem augmented with the reference solution and the same reference-solution framing used by the subsequent reference-conditioned OPSD stage. The teacher forward pass is performed under `torch.no_grad()`, so it contributes no gradient.

No student rollouts are generated during LD-SFT. Instead, teacher and student distributions are evaluated on exactly the same supervised completion tokens. The implementation explicitly aligns the teacher and student target spans, and only completion tokens that contribute to the ordinary CoT-SFT loss receive LD-SFT weights. If the teacher-side sequence exceeds the maximum sequence length, the privileged reference solution is truncated while the supervised completion is retained unchanged.

Teacher–student divergence. For the experiments reported in the main paper, token-level JSD is computed with temperature 1.0 and the symmetric mixture coefficient $\beta = 0.5$. To reduce the memory cost of computing full-vocabulary JSD, we use a top-k approximation with $k = 128$. The support is the union of the teacher's and student's top-128 tokens; duplicate tokens are removed, and all probability mass outside this union is collapsed into a single residual bucket. We then compute JSD over this coarsened distribution. This construction retains information about probability mass outside the selected support rather than renormalizing the selected tokens to sum to one.

Divergence computation is performed in chunks of 512 target positions. The exact full-vocabulary implementation is also supported by our code, but is not used for the main LD-SFT runs reported here.

Training hyperparameters. All main LD-SFT runs use Qwen3-4B and are trained for 100 optimizer updates with learning rate 5×10^{-6}. We use LoRA with rank 64 and scaling factor 128, applied to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The per-device batch size is 2 with 16 gradient accumulation steps, giving an effective batch size of 32 for the single-process runs. The maximum sequence length is 16,000 tokens, training uses bfloat16 and gradient checkpointing, and checkpoints are saved every 20 updates.

We use the same `siyanzhao/Openthoughts_math_30k_opsd` training corpus as Vanilla SFT. Examples with an empty answer field are removed before formatting, after which the data are split into training and evaluation subsets using a seeded 99/1 split with random seed 42. The supervised target is the `COT_Reason` completion; if a completion does not already contain a boxed answer, the target answer is appended in boxed form. Student prompts are rendered with Qwen3 thinking disabled.

Compute overhead. Relative to Vanilla SFT, LD-SFT requires one additional no-gradient forward pass through the current model for the privileged teacher at each minibatch. The ordinary gradient-carrying student forward is unchanged. Our main configuration reuses the student logits from this forward when constructing the detached divergence signal and therefore does not perform an additional student forward. Consequently, the principal computational overhead of LD-SFT is the privileged-teacher forward together with the top-128 JSD computation.

The implementation additionally supports a memory-saving mode that recomputes the student under `no_grad()` before the gradient-carrying forward, preventing teacher logits and gradient-carrying student logits from coexisting in memory. We do not enable this option in the main runs, as it adds an additional student forward pass. We use the top-128 JSD approximation in the main runs to reduce the temporary memory required by divergence computation.

C.1 Control Objectives and OPSD Variants

We compare LD-SFT against both SFT-side controls and alternative OPSD constructions. Let

$$\ell_t = -\log p_\theta(y_t \mid y_{<t}, x)$$

denote the token-level cross-entropy loss, and let w_t^{LD} denote the finalized token weight assigned by LD-SFT.

- **Vanilla SFT.** Vanilla SFT assigns unit weight to every supervised token:

$$w_t = 1, \qquad \mathcal{L}_{\mathrm{SFT}} = \frac{1}{N}\sum_{t=1}^{N} \ell_t.$$

  This is the standard token-level cross-entropy objective and serves as the unweighted baseline.

- **Random-SFT ($\alpha = 0.25$).** Random-SFT controls for the effect of introducing non-uniform token weights without preserving their alignment with teacher–student divergence. We first compute exactly the same finalized per-token weights $\{w_t^{\mathrm{LD}}\}$ as LD-SFT with $\alpha = 0.25$. Within each training example, we then independently sample a random permutation π of the supervised target positions and assign

$$w_t^{\mathrm{rand}} = w_{\pi(t)}^{\mathrm{LD}}, \qquad \mathcal{L}_{\mathrm{rand}} = \frac{1}{N}\sum_{t} w_t^{\mathrm{rand}} \ell_t.$$

  Random-SFT therefore preserves the exact per-example multiset of LD-SFT weights while destroying their token-level correspondence with teacher–student divergence. It tests whether the benefit of LD-SFT can be explained simply by changing the relative contribution of supervised tokens.

- **Easy-SFT ($\alpha = 0.25$).** Easy-SFT tests whether a purely student-side notion of token difficulty is sufficient. As in Random-SFT, we begin with the finalized LD-SFT weight multiset computed using $\alpha = 0.25$, but reassign these weights according to the student's detached token-level cross-entropy.
  If

$$\ell_{i_1} \le \ell_{i_2} \le \cdots \le \ell_{i_m}$$

  are the target tokens of an example ordered from easiest to hardest, and

$$w_{(1)}^{\mathrm{LD}} \ge w_{(2)}^{\mathrm{LD}} \ge \cdots \ge w_{(m)}^{\mathrm{LD}}$$

  are its finalized LD-SFT weights sorted from largest to smallest, then

$$w_{i_j}^{\mathrm{easy}} = w_{(j)}^{\mathrm{LD}}, \qquad \mathcal{L}_{\mathrm{easy}} = \frac{1}{N}\sum_{t} w_t^{\mathrm{easy}} \ell_t.$$

  Cross-entropy is detached when constructing the ordering, and ties are resolved stably in token order. Thus, "easy" denotes tokens that the current student already predicts with low cross-entropy, testing whether generic student-side difficulty is sufficient without teacher–student distributional information.

- **HD-SFT ($\alpha = 0.25$).** HD-SFT uses the same weighting strength $\alpha = 0.25$ while reversing the divergence preference of LD-SFT: instead of assigning larger weight to tokens with smaller teacher–student divergence D_t, its weighting function is monotone increasing in D_t,

$$D_{t_1} < D_{t_2} \quad \Longrightarrow \quad w_{t_1}^{\mathrm{HD}} \le w_{t_2}^{\mathrm{HD}}, \qquad \mathcal{L}_{\mathrm{HD}} = \frac{1}{N}\sum_{t} w_t^{\mathrm{HD}} \ell_t.$$

  This control tests whether the benefit comes specifically from prioritizing reachable, low-divergence supervision rather than from divergence-aware weighting more generally.

- **EMA OPSD.** Starting from the Vanilla-SFT initialization, EMA OPSD replaces the ordinary teacher parameters with an exponential moving average of the trainable student parameters. After optimizer step s,

$$\theta_{\text{EMA}}^{(s)} = \gamma\theta_{\text{EMA}}^{(s-1)} + (1-\gamma)\theta_{\text{student}}^{(s)}, \qquad \gamma = 0.99.$$

  The EMA parameters are used only for the teacher forward pass, while the live student parameters are optimized with the same OPSD objective.

- **Top-k OPSD.** Starting from the Vanilla-SFT initialization, Top-k OPSD restricts the distillation distribution at each token to the teacher's $k = 128$ most probable vocabulary entries. Writing

$$S_t = \text{TopK}_{128}\big(p_T(\cdot \mid s_t, z)\big),$$

  the teacher and student distributions are renormalized on this shared support,

$$\tilde{p}_T(v) = \frac{p_T(v)}{\sum_{u\in S_t} p_T(u)}, \qquad \tilde{p}_S(v) = \frac{p_S(v)}{\sum_{u\in S_t} p_S(u)}, \qquad v \in S_t,$$

  and the usual OPSD divergence is replaced by

$$D_t^{\text{top-}k} = \text{JSD}\big(\tilde{p}_T, \tilde{p}_S\big).$$

Unless stated otherwise, Random-SFT, Easy-SFT, and HD-SFT all use $\alpha = 0.25$, matching the corresponding LD-SFT setting so that the controls differ in *which tokens receive large weights*, rather than in the overall weighting strength.

Across the complete three-seed evaluations, OPSD consistently improves over its corresponding SFT initialization. On AIME 2024, the highest mean OPSD performance is obtained from LD-SFT with $\alpha = 0.25$. On AIME 2025, both LD-SFT initializations yield higher best OPSD means than the Vanilla-SFT initialization.

C.2 Control Experiment Setup and Checkpoint-Averaged Results

Table 27: Average downstream OPSD performance across checkpoints 25, 50, 75, and 100. Δ denotes the difference between the checkpoint-averaged OPSD performance and the corresponding pre-OPSD parent.

Method	**AIME24**		**AIME25**	
	Avg. OPSD	**Δ vs. Parent**	**Avg. OPSD**	**Δ vs. Parent**
No SFT (Base) + OPSD	26.35	+3.85	20.21	-1.04
Vanilla SFT + OPSD	68.23	+4.06	58.13	+3.96
+ EMA OPSD	67.50	+3.33	58.02	+3.85
+ Top-k OPSD	67.40	+3.23	60.62	+6.46
Random-SFT + OPSD	68.65	+5.73	57.71	+6.04
Easy-SFT + OPSD	66.04	+9.79	59.38	+9.38
HD-SFT + OPSD	69.38	+8.54	57.81	+7.40
LD-SFT ($\alpha = 0.25$) + OPSD	**70.42**	**+9.17**	60.31	**+9.48**
LD-SFT ($\alpha = 1$) + OPSD	69.27	+8.44	**61.56**	+6.98

We use the same model, SFT data, privileged-reference teacher, OPSD training procedure, and evaluation protocol as in Section 3.1. All SFT variants are trained for 100 steps so that the comparison remains in the short-warm-up regime for which low divergence is predictive of downstream OPSD gain. In addition to AIME24 and AIME25, we evaluate on HMMT25 to test whether the effect transfers beyond the benchmarks used to motivate the method.

We compare LD-SFT against both SFT-side and OPSD-side alternatives.

- **Vanilla SFT.** Standard token-level cross-entropy training without reweighting.
- **Random-SFT.** We first compute exactly the same finalized token weights as LD-SFT and then, independently within each training example, randomly permute those weights across supervised target tokens. This preserves the per-example multiset of weights while destroying their alignment with token divergence, controlling for the possibility that redistributing the relative contribution of tokens is itself beneficial.
- **Easy-SFT.** We use the same finalized weight multiset as LD-SFT but reassign it according to the student's detached per-token cross-entropy. Within each example, target tokens are sorted from lowest to highest cross-entropy and assigned the available weights from largest to smallest, with ties resolved stably in token order. Thus, "easy" denotes tokens that the current student already predicts with low cross-entropy, testing whether generic student-side difficulty is sufficient without teacher–student distributional information.
- **HD-SFT.** We reverse the divergence preference of LD-SFT so that larger teacher–student JSD receives larger weight. This tests whether the benefit comes specifically from prioritizing reachable, low-divergence supervision rather than from divergence-aware weighting more generally.
- **EMA OPSD.** Starting from the Vanilla-SFT initialization, the teacher uses an exponential moving average of the trainable student parameters, updated after each optimizer step as $\theta_{\text{EMA}} \leftarrow 0.99\theta_{\text{EMA}} + 0.01\theta_{\text{student}}$. The live student parameters are used for optimization.
- **Top-k OPSD.** Starting from the Vanilla-SFT initialization, at each distilled token we restrict both teacher and student distributions to the teacher's top-128 vocabulary entries, renormalize both distributions over this support, and compute the same JSD objective on the resulting truncated distributions.

We evaluate LD-SFT itself with $\alpha \in \{0.25, 1\}$. The SFT-side controls test whether the effect is attributable specifically to low teacher–student divergence rather than to generic token reweighting or token difficulty, while the OPSD-side controls test whether comparable improvements can instead be obtained by modifying the subsequent distillation stage while retaining the standard SFT initialization.

C.3 Multi-Seed SFT and OPSD Evaluation

Table 28: SFT checkpoint-100 performance and the best OPSD performance obtained from each SFT initialization. Results are mean ± standard deviation of avg@8 (%) across three seeds.

Method	AIME 2024	AIME 2025
Vanilla SFT ($\alpha = 0$)	63.89 ± 1.27	54.03 ± 0.64
+ OPSD	$\mathbf{69.44 \pm 0.48}$	$\mathbf{60.14 \pm 1.05}$
LD-SFT ($\alpha = 1$)	60.83 ± 2.50	51.81 ± 4.11
+ OPSD	$\mathbf{69.58 \pm 0.42}$	$\mathbf{62.36 \pm 0.24}$
LD-SFT ($\alpha = 0.25$)	61.39 ± 2.71	54.79±0.88
+ OPSD	$\mathbf{70.56 \pm 2.29}$	$\mathbf{62.92 \pm 1.82}$

We evaluate the robustness of OPSD across multiple SFT initializations using three training seeds. Specifically, we report results for the original seed-0 run and two new replicates with master seeds 123 and 456. For the new replicates, a single master seed is propagated consistently through both stages of training. During SFT, the master seed determines the training-data order, LoRA initialization, and dropout randomness. During OPSD, the same master seed determines the optimization/data order, PyTorch randomness, and rollout-generation randomness. Each OPSD run is therefore paired with, and initialized from, the SFT checkpoint produced by the same method and the same seed; in all cases, OPSD starts from the corresponding SFT checkpoint at step 100.

The seed-0 run follows the earlier seeding regime used in our original experiments. In this run, SFT uses Trainer seed 42, the LoRA initialization does not depend on the training seed, and OPSD uses a vLLM seed of 0. Apart from these seeding differences, the training recipe and hyperparameters are held fixed across seeds. For evaluation, we fix the vLLM seed to 0 for every run, independently of

the training seed. Thus, the variation reported across seeds primarily reflects randomness introduced during SFT and OPSD training rather than changes in the evaluation-time sampling seed.

Table 28 reports SFT performance at checkpoint 100 and the best OPSD performance obtained from that SFT initialization. OPSD is evaluated at steps 25, 50, 75, and 100, and for each method and benchmark we report the checkpoint with the highest mean across the three seeds. We report only results for which all three seeds have complete evaluations. Results are reported as mean $\pm$ sample standard deviation of avg@8.