DIAGNOSING ON-POLICY SELF-DISTILLATION FOR REASONING LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) has attracted growing interest as a promising approach to improve the reasoning ability of language models. Without external rewards nor a separate stronger teacher, the self-teacher with privileged information could provide dense signals on student's trajectories. However, its behavior in language reasoning remains unclear, with reported outcomes ranging from modest gains to behavioral collapse. In this work, we diagnose OPSD for mathematical reasoning across models spanning 0.6B–8B parameters. We conduct controlled experiments and token-level analyses to fully delve into OPSD. We point out that teacher's signal is shaped by reasoning-mode alignment and the complete teacher prefix, rather than by privileged semantics alone. OPSD improves reasoning only in narrow compatibility regimes. Otherwise, it produces ineffective length growth, stable degradation, or behavioral collapse. Token-level analysis shows that teacher's signal is not stable and does not predict downstream performance. Based on these results, we argue that OPSD is a sensitive algorithm rather than a generally reliable reasoning-improvement post-training method.

1 INTRODUCTION

Post-training is important for improving the reasoning ability of large language models (LLMs). Reinforcement learning with verifiable rewards (RLVR), commonly optimized with algorithms such as Group Relative Policy Optimization (GRPO, Shao et al., 2024), is widely used when outputs can be verified automatically. However, these rewards are typically sparse and depend heavily on the final answer, making it difficult to assign credit across the entire reasoning trajectory.

On-policy distillation (OPD, Agarwal et al., 2024; Song & Zheng, 2026) uses a stronger model as teacher to provide dense signals on the student's trajectories token by token. On-policy self-distillation (OPSD, Zhao et al., 2026) replaces the teacher with a privileged self-teacher, which is a fixed copy of the student model. OPSD therefore removes the need for a stronger external model while maintaining the need for privileged training information.

Unlike the ability gap between OPD's teacher and student, the ability gap between OPSD's teacher and student is much smaller. It's difficult to transfer the ability from the teacher to the student. OPSD may therefore transfer context-conditioned behavior rather than improving unprivileged reasoning ability. Therefore, we ask three questions: **First, how does the privileged information shape the teacher's signal? Second, under what joint conditions does OPSD improve reasoning ability? Third, what can training dynamics reveal about training quality and downstream performance?** We study these questions through controlled experiments and token-level analyses mainly on Qwen3 (Qwen Team, 2025) and OLMo-3 (OLMo Team et al., 2026) models spanning 0.6B–8B parameters. Our main findings are:

- **The OPSD signal is caused by prefix distribution mismatch rather than semantic message alone.** Reasoning-mode mismatch induces the behavioral shift, and changes are little while removing the privileged information. Appropriate distirbution rather than semantic information alone can improve the performance.
- **OPSD gains are modest and confined to model- and configuration-specific regimes.** Its effect depends jointly on model ability, rollout length, training duration, teacher context, and initializa-

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.

tion stage. Improvements between OPSD settings do not necessarily become gains over the base model. Moreover, different KL divergences have different effects across models.

- **Initial signal statistics and training traces identify collapse more readily than downstream performance** Initial forward KL does not predict the signed effect, and numerically stable optimization can still degrade performance. Instability manifests as collapse, ineffective length growth, and broken formatting. Loss and entropy-based token-selection redesigns provide no consistent remedy.

2 Related Work

RLVR. RLVR obtains rewards from verifiable answers. GRPO samples G responses per question Q and replaces PPO's(Schulman et al., 2017) critic with group-normalized advantages:

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|y^{S,g}|}\sum_{i}\Big(\min\Big(\rho_{g,i}\hat{A}_g, \mathrm{clip}(\rho_{g,i}, 1-\varepsilon, 1+\varepsilon)\hat{A}_g\Big) - \beta D_{\mathrm{KL}}(\pi_\theta\|\pi_{\mathrm{ref}})\Big)\right], \tag{1}$$
$$\hat{A}_g = \frac{r_g - \bar{r}}{s_r + \epsilon}, \qquad \rho_{g,i} = \frac{\pi_\theta(y_i^{S,g} \mid Q, y_{<i}^{S,g})}{\pi_{\mathrm{old}}(y_i^{S,g} \mid Q, y_{<i}^{S,g})}.$$

DAPO (Yu et al., 2025) adds asymmetric clipping, dynamic sampling, token-level normalization, and overlength handling, whereas GSPO uses sequence-level importance ratios and clipping for greater stability (Zheng et al., 2025). Although RLVR has been widely used in reasoning models, its sparse reward and verification process takes much computation and time to train.

Following work on OPSD. Recent works modify which OPSD targets or positions should be trusted. Position-Weighted OPSD emphasizes positions with more reliable teacher supervision (Liu et al., 2026). Purified OPSD applies a PMI-based correction to the privileged target (Shen et al., 2026). And β-OPSD interpolates teacher and reference distributions with return-to-go credit assignment (Xu et al., 2026). These methods improve OPSD sometimes, but they don't analyze the underlying reasons why OPSD works and when it fails.

Rethinking OPSD. Recent analyses question the reliability and source of OPSD's gains. Privileged supervision can degrade reasoning models on long traces (Kaur et al., 2026). Another work finds that the correct reference is not always beneficial and that apparent improvements may instead reflect recovery of reasoning behavior present in the base model (Shrestha & Tessier, 2026). These works deeply analyze a part of OPSD but lack broader explorations and global diagnosis.

3 Preliminaries

Let $Q \sim \mathcal{D}$ denote a question and $y^S = (y_1^S, \ldots, y_N^S)$ a student-generated rollout, with $y_{<i}^S$ denoting the student prefix before position i. S and T identify student and teacher quantities, respectively.

3.1 On-Policy Distillation (OPD)

For a student rollout $y^S \sim \pi_\theta(\cdot \mid Q)$, let $\pi_i^S(v) = \pi_\theta(v \mid Q, y_{<i}^S)$ and $\pi_i^T(v) = \pi_\phi(v \mid Q, y_{<i}^S)$ denote the student and fixed teacher distributions on the visited student prefix $y_{<i}^S$. OPD minimizes

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{Q\sim\mathcal{D},\, y^S\sim\pi_\theta(\cdot|Q)}\left[\frac{1}{N}\sum_{i=1}^{N} D\big(\pi_i^T,\, \pi_i^S\big)\right]. \tag{2}$$

Here D is a divergence between next-token distributions. Each rollout is sampled from the current student. At every response position, the teacher conditions on the student-generated prefix and computes its next-token distribution. This provides dense supervision for OPD along the rollout trajectory. Gradients are not propagated through the rollout sampling process.

3.2 On-Policy Self-Distillation (OPSD)

OPSD (Zhao et al., 2026) specializes OPD by using a frozen copy π_{θ_0} of the initial student as the teacher and providing it with privileged context C. For each pair (Q, C), the student conditions

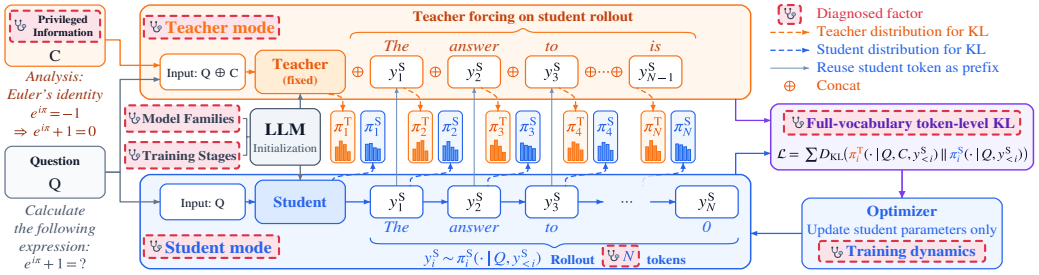


Figure 1: Overview of our diagnosis. We design a comprehensive diagnostic pipeline to study how OPSD behaves in different settings and what can token-level dynamics reveal during training.

only on Q, whereas the teacher conditions on $Q \oplus C := \mathcal{T}_{\text{teacher}}(Q, C)$. The privileged information C may contain a reference solution or verified answer unavailable to the student. On a student-generated prefix $y^S_{<i}$, the teacher distribution is $\pi^T_i(v) = \pi_{\theta_0}(v \mid Q, C, y^S_{<i})$, while the student distribution remains as in OPD. OPSD instantiates D in Equation 2 as a clipped forward KL. It caps each vocabulary contribution $d_{i,v} = \pi^T_i(v)\big[\log \pi^T_i(v) - \log \pi^S_i(v)\big]$ before summation:

$$\mathcal{L}^{\text{clip}}_{\text{OPSD}}(\theta) = \mathbb{E}_{(Q,C)\sim\mathcal{D},\, y^S\sim\pi_\theta(\cdot|Q)}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{v\in\mathcal{V}}\min(d_{i,v}, \tau_{\text{clip}})\right]. \tag{3}$$

The clipping limits large contributions from individual vocabulary items.

4 Empirical Methodology and Setup

We write $m_S, m_T \in \{\text{NTh}, \text{Th}\}$ for the student and teacher training or inference modes. The student uses m_S for on-policy generation. The teacher uses m_T when scoring the student's rollout. In the paper, we abbreviate THINK as Th and NO-THINK as NTh. We mainly train Qwen3 and OLMo-3 models spanning 0.6B–8B parameters. We additionally use DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025), Falcon-H1R-7B (Falcon LLM Team et al., 2026), MiMo-7B-RL (LLM-Core-Team Xiaomi, 2025), Qwen3.5-4B (Qwen Team, 2026), and Qwen3-4B-Thinking-2507(Qwen3-4B-Thinking,Qwen Team (2025)) for additional tests.

The main OPSD experiments mainly use OpenThoughts Math 30k (Zhao et al., 2026). Default OPSD setup uses a frozen same-model teacher, full-solution teacher context. We set 1,024-token prompt and rollout caps, 100 training steps, learning rate 10^{-6}, forward KL ($\beta = 0$), $\tau_{\text{clip}} = 0.05$, seed 42, and AdamW(Loshchilov & Hutter, 2019) optimizer. We evaluate on AIME 2024, AIME 2025, AIME 2026, and HMMT February 2025, which are sufficiently challenging to evaluate model reasoning ability. We sample eight responses per question and report Avg@8, Pass@8, and mean output length. Reported scores are average of the four benchmarks. Complete model-specific training and evaluation defaults are given in section B. Detailed Results are given in section G.

We use several phrases to describe model's some performances. **Ineffective deliberation** denotes little decline of accuracy accompanied by substantial evaluation length growth. **Stable degradation** denotes that as training step or student rollout length increases, accuracy declines gradually while evaluation rollout length grows. **Behavioral collapse** denotes a large performance degradation accompanied by failures such as repetition, hallucination, missing of <eos> token, corrupted text, or loss of boxed-answer formatting.

5 Experimental Results

We first establish a valid comparison setting. We then examine training steps, rollout length, teacher context, loss design, model ability, pipeline stage, and OPSD versus GRPO. Figure 1 summarizes the experimental pipeline. Per-dataset results and robustness checks are in section G and subsection F.1.

5.1 Valid Comparison Setting

Reasoning-mode mismatch confounds attribution because Qwen3's Th and NTh modes differ in both reasoning ability and response style. It's necessary to confirm the training mode.

Table 1: Controls for establishing a valid comparison setting. **(a)** Reasoning-mode results averaged over Qwen3-1.7B/4B/8B. For Avg@8 and Pass@8, Δ is the absolute OPSD–Base change in percentage points. For Length, it is the relative change. Eval is the inference mode of both Base and OPSD, and S/T denote student/teacher modes. **(b)** Qwen3-4B feature rates. Self-correction includes Hmm/Actually/No/But. **(c)** No-privilege same-prefix control for mismatched NTh/Th, evaluated in Th mode. Parentheses show change from the corresponding full-solution checkpoint. Per-model results are in Table 14. Per-dataset results are in Table 49.

(a) Reasoning-mode comparison

Model	S	T	Eval	Avg@8			Pass@8			Length		
				Base	OPSD	Δ	Base	OPSD	Δ	Base	OPSD	Δ
Qwen3 mean	Th	Th	Th	53.5	50.7	−2.8	72.2	71.1	−1.1	18,013	21,250	+17.9%
	Th	NTh	Th	53.5	36.6	−16.9	72.2	59.5	−12.7	18,013	13,291	−26.0%
	NTh	Th	NTh	15.8	19.6	+3.8	32.2	39.7	+7.5	5,510	14,093	+154.8%
	NTh	NTh	NTh	15.8	15.2	−0.6	32.2	28.6	−3.6	5,510	13,634	+144.3%
	NTh	Th	Th	53.5	56.2	+2.7	72.2	73.1	+0.9	18,013	19,527	+8.4%
	Th	NTh	NTh	15.8	14.0	−1.8	32.2	27.8	−4.4	5,510	5,104	−8.7%

(b) Qwen3-4B feature rates

Stylistic feature	NTh Base	NTh/Th	NTh/NTh	Th base
Contains "wait"	31.6%	47.0%	18.2%	100.0%
Self-correction	35.7%	49.9%	21.0%	99.7%
Contains `<think>`	0.0%	0.1%	0.0%	100.0%

(c) No-privilege control

Model	Avg@8	Pass@8	Length
Qwen3-1.7B	41.0 (+0.2)	62.5 (-0.8)	20,486 (+2.4%)
Qwen3-4B	60.9 (-2.1)	75.0 (-4.2)	19,715 (+2.2%)
Qwen3-8B	64.3 (-0.5)	80.8 (+4.1)	19,699 (+2.1%)

Table 1, panel (a), shows a directional effect: a NTh teacher suppresses a Th student's accuracy and length, whereas a Th teacher increases both for a NTh student. Although there is no explicit `<think>` leakage, the latter still shows more Th-associated discourse features (panel (b)).

The no-privilege same-prefix control in panel (c) supports this interpretation. Across model sizes, the no-privilege versus full-solution means differ only slightly (55.4/72.8/20.0k vs 56.2/73.1/19.5k), so the apparent gain primarily reflects mode transfer rather than privileged semantics.

Because mismatch separates the rollout and evaluation policies and relies on a Qwen-specific mode switch, making comparisons across model families unavailable, all subsequent experiments are trained and tested on matched native modes, with matched Th/Th for hybrid reasoning models.

5.2 Training Steps and Maximum Student Rollout Length

We first ask whether more training steps can improve final performance. We train Qwen3-1.7B and OLMo-3-7B-Think for up to 300 steps and evaluate every 50 steps (Figure 2). The result shows that overall performance declines as training continues, despite small intermediate rebounds. Sustained training therefore does not reliably improve model capability.

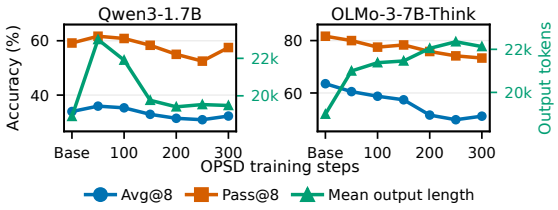


Figure 2: Stability over training steps. Per-dataset results are in Table 52.

We next compare loss-token positions with a 256-position supervision budget. First-256 uses the equivalent OPSD-256 condition, while Uni-256 and Final-256 use a 1,024-token rollout and retain uniformly sampled or final valid response positions (Appendix F.1.2). First-256 gives the highest Avg@8 and shortest outputs while Final-256 performs the worst (Table 2).

With the comparison restricted to matched modes, we then vary the maximum student training rollout length over {128, 256, 512, 1024, 2048, 4096, 6144} tokens. For instruct models, performance decreases as rollout length increases: short or intermediate rollout lengths are competitive with or better than the base, whereas the longest settings lose accuracy.

Table 2: Results of position loss. Parentheses: change from Base. Per-dataset results are in Table 51.

Positions	Avg@8	Pass@8	Length
First-256	**49.5** (+0.7)	71.3 (+0.8)	**21.4k** (+14.7%)
Uni-256	47.1 (-1.7)	**72.1** (+1.6)	22.0k (+17.9%)
Final-256	47.8 (-1.0)	69.2 (-1.3)	22.4k (+20.1%)

Under matched Th/Th, both models exhibit stable degradation. Prior work reports that privileged OPSD does not improve reasoning models. We find that small gains remain possible, but only for some models and only at short rollout lengths.

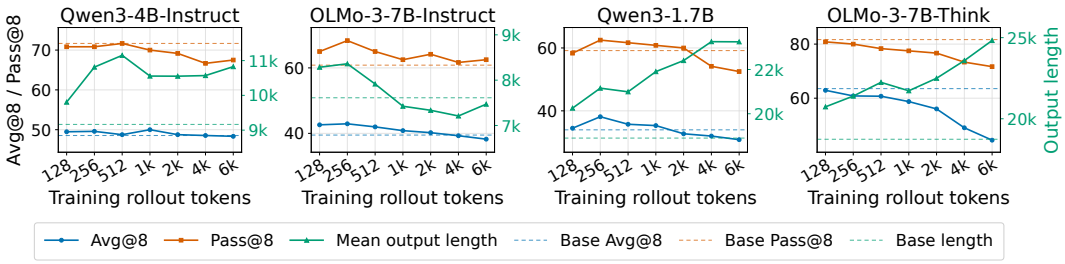


Figure 3: OPSD vs. rollout cap for four models. Solid: OPSD. Dashed: Base. Left axes: Avg@8/Pass@8. Right: mean eval output length. Per-dataset results are in Table 50 and Table 51.

5.3 Privileged Information

We next vary privileged information to test whether target-specific information alone explains training utility. The primary conditions use target problems from OpenThoughts Math 30k. Besides the full solution, ANSWER provides only the correct final answer, while UNRELATED provides a deduplicated problem–solution pair from DAPO-Math-17k (Yu et al., 2025). The unrelated pair changes the privileged information without supplying evidence relevant to the target problem. The student receives only the target problem in all conditions.

For context type $z \in \{\text{SOLUTION}, \text{ANSWER}, \text{UNRELATED}\}$, the teacher input is $Q, C^{(z)} := \mathcal{T}_{\text{teacher}}(Q, C^{(z)})$. Because the intervention changes a complete prefix, it estimates the overall context effect rather than separately identifying length or target-problem position. We vary these three contexts across five model settings at maximum rollout lengths of 256 and 1,024 tokens (Table 3).

Table 3: Macro-average effect of privileged information across five model settings. Parentheses report change from Base in percentage points for Avg@8 and Pass@8 and relative percent for Length. Per-model and per-dataset results are in Table 53.

Teacher context	Rollout length: 256			Rollout length: 1,024		
	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length
Full solution	**50.5** (+0.9)	**71.5** (+1.0)	16.2k (+12.9%)	**48.6** (−1.0)	**69.3** (−1.2)	16.4k (+12.2%)
Unrelated solution	48.9 (−0.7)	69.7 (−0.8)	15.4k (+6.0%)	47.8 (−1.8)	67.3 (−3.2)	18.0k (+25.3%)
Answer only	45.9 (−3.7)	65.7 (−4.8)	16.3k (+6.4%)	44.3 (−5.3)	66.3 (−4.2)	17.4k (+12.9%)

Context variants exhibit a non-monotonic ordering. At both rollout lengths in Table 3, Unrelated solution outperforms Answer only. Furthermore, we apply CoT+SOLUTION to Qwen3-1.7B and OLMo-3-7B-Think by concatenating the OpenThoughts CoT trace with the gold solution on the teacher side, using exactly the same target problems and row order as Full solution while increasing the teacher prompt budget from 1,024 to 8,192 tokens. For Qwen3-1.7B, Avg@8 and Pass@8 both decrease. For OLMo-3-7B-Think, neither metric improves (Table 4). Thus, providing more detailed teacher-side information does not necessarily yield better training outcomes.

Table 4: Results at a 1,024-token rollout cap. Per-dataset results are in Table 53.

	Qwen3-1.7B			OLMo-3-7B-Think		
Teacher context	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length
None (Base)	34.0	59.2	18.6k	63.5	81.7	18.7k
Answer only	35.6	59.2	25.8k	56.1	75.8	24.1k
Unrelated solution	**37.1**	60.0	22.0k	57.5	76.7	23.4k
Full solution	35.3	**60.8**	22.4k	**58.8**	**77.5**	21.7k
CoT + solution	30.8	55.8	18.6k	58.6	**77.5**	20.8k

5.4 Local Loss-Design

Following the OPSD objective (Zhao et al., 2026), at position i the interior branch $0 < \beta < 1$ uses the generalized-JSD loss

$$\ell_\beta(\pi_i^S, \pi_i^T) = \beta D_{\text{KL}}(\pi_i^T \| m_{\beta,i}) + (1-\beta) D_{\text{KL}}(\pi_i^S \| m_{\beta,i}), \quad m_{\beta,i} = (1-\beta)\pi_i^S + \beta \pi_i^T. \tag{4}$$

The endpoints are defined separately: $\ell_0 = D_{\text{KL}}(\pi_i^T \| \pi_i^S)$ (forward KL) and $\ell_1 = D_{\text{KL}}(\pi_i^S \| \pi_i^T)$ (reverse KL). We use $\beta = 0$ by default and ablate $\beta \in \{0.5, 1\}$. The result shows that only a few trained checkpoints improve over their corresponding base models, and the leading objective varies by model (Table 5, panel (a)). Changing divergence direction therefore redistributes outcomes rather than providing a reliable remedy. Vocabulary truncation, teacher-favored position selection, and student-entropy masking are likewise non-universal among models (subsubsection C.4.2 and subsubsection C.4.3).

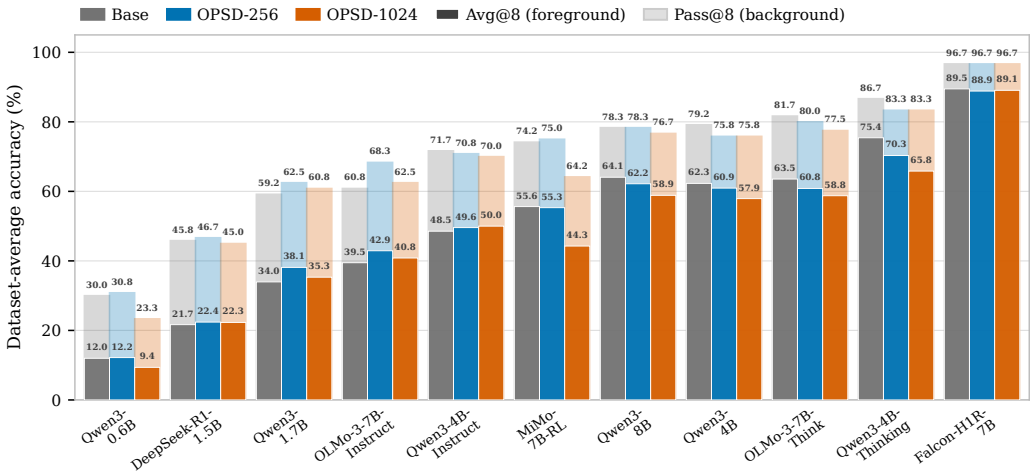


Figure 4: OPSD results across baseline performance levels. Each model cluster reports Base, OPSD-256, and OPSD-1024. Dark foreground bars show Avg@8. Light background bars show Pass@8. Models are ordered by Base Pass@8. Per-dataset results are in Table 58.

5.5 Initial Model Ability

We then ask whether OPSD's performance depends on the model itself. We compare Base, OPSD-256, and OPSD-1024 across eleven models under their matched student–teacher mode and standard full-solution privileged information.

The result shows that only some models with lower reasoning ability benefit from OPSD, whereas most higher-performing models degrade (Figure 4). Consistent with prior work (Kaur et al., 2026), instruction-tuned models are often less adversely affected than reasoning models, which might be attributed to their instruction following ability. The result suggests that models with stronger reasoning performance may be less likely to benefit from OPSD, and instruction models benefit more than reasoning models.

Table 5: Optimization sensitivity. **(a)** Avg@8 and Pass@8 changes after 100 steps of OPSD-1024 under forward KL, symmetric JSD, and reverse KL. Bold marks the best change for each model. Per-model and per-dataset results are in Table 16 and Table 55. **(b)** Post-training pipeline results. Format is the boxed-answer rate (%) and Length is mean output length. Per-checkpoint results are in Table 22. Per-dataset results are in Table 61 and Table 64.

(a) Divergence objectives

	Fwd. KL		JSD		Rev. KL	
Model	Avg	Pass	Avg	Pass	Avg	Pass
Qwen3-1.7B	+1.3	+1.6	−2.7	−4.2	−7.5	−11.7
Qwen3-4B	−4.4	−3.4	−5.7	−3.4	−30.3	−22.5
Qwen3-4B-Thinking	−9.6	−3.4	−7.4	−4.2	−1.2	−1.7
OLMo-3-7B-Think	−4.7	−4.2	−1.4	0.0	−1.1	+0.8
OLMo-3-7B-Instruct	+1.3	+1.7	−2.3	−2.5	+3.0	+0.9
Falcon-H1R-7B	−0.4	0.0	−1.1	−0.9	−1.4	0.0
MiMo-7B-RL	−11.3	−10.0	−3.0	−1.7	−5.4	−1.7

(b) Post-training stages

Model	Avg@8	Pass@8	Format	Length
Qwen3-4B-Base	7.29	20.00	87.19	3.0k
+ OPSD	4.79	17.50	74.59	10.3k
+ SFT	34.69	60.00	89.27	14.3k
+ SFT + OPSD	1.98	8.33	3.44	32.9k
+ SFT + GRPO	33.65	60.83	94.79	14.2k
+ SFT + GRPO + OPSD	3.54	14.17	5.62	37.6k
Qwen3-4B	62.3	79.2	97.1	17.8k
+ OPSD	57.9	75.8	86.4	19.9k

5.6 Sensitivity to Post-Training Stage

We next test a separate source of sensitivity: the post-training stage at which OPSD is applied. The SFT stage uses 15,000 steps on the `cot` split of OpenMathReasoning. The GRPO stage uses DAPO-Math-17k. Full configurations appear in subsection B.3 and subsection B.4.

The failure mode changes with pipeline stage. Direct OPSD on Qwen3-4B-Base produces ineffective deliberation, so we treat this low-ability starting point only as a sanity check. OPSD after either our SFT or SFT+GRPO stage instead produces behavioral collapse (Table 5, panel (b)). GRPO itself approximately preserves the SFT checkpoint, but the subsequent OPSD collapses. Matched Th/Th OPSD on the released Qwen3-4B instead produces stable degradation.

5.7 TRAINING-COMPUTE TRADE-OFF: OPSD VERSUS GRPO

Finally, we compare OPSD and GRPO on Qwen3-1.7B / 4B. Both methods train for 50 steps on exactly the same OpenMathReasoning problems in the same order. OPSD conditions its teacher on the full solution, whereas GRPO uses only the final answer for verification. Table 6 reports accuracy and training GPU-hours. Full configurations are reported in subsection C.7.

Table 6: Cells report Avg@8 / Pass@8 / GPU hours. Bold marks the best Avg@8 or Pass@8. Per-dataset results are in Table 62 and Table 63.

Model	Base	OPSD_{50}	GRPO_{50}
Qwen3-1.7B	34.0/59.2/–	**38.0/63.3**/0.66	37.1/**63.3**/114.27
Qwen3-4B	62.3/79.2/–	60.1/**80.0**/1.13	**62.8**/78.3/171.81

OPSD performs better than GRPO on 1.7B model with less training computation. However, the advantage decreases on 4B model. This may imply that OPSD's accuracy advantage does not consistently scale with model size, although its computational advantage remains substantial.

6 ANALYSIS: FROM TOKEN-LEVEL SIGNALS TO OPTIMIZATION DYNAMICS

We analyze the frozen token-level signal across reasoning mode, checkpoint type, teacher context, and response position, then ask what training trajectories can diagnose. Each model uses 2,048 prompts with two on-policy rollouts ($n = 4{,}096$). Completions are capped at 1,024 tokens except in the 6,144-token response-position analysis. Settings and per-model results are in subsection D.1.

6.1 MISMATCHED MODE EFFECT

Table 7 shows that mismatched pairs have substantially higher initial forward KL and high-divergence-position rates than matched pairs, despite high top-1 agreement. Mode mismatch amplifies the initial signal: both forward KL and the share of high-divergence positions increase relative to matched pairs. The effect is asymmetric, with a larger shift for Th/NTh than for NTh/Th.

Table 7: Qwen3-1.7B grid. S/T are student/teacher modes, and $D > .05$, Agree., Enc., and Disc. are percentages.

S	T	Fwd. KL	$D > .05$	Agree.	Enc.	Disc.
NTh	NTh	0.100	17.8	94.3	32.3	37.9
NTh	Th	0.193	24.4	93.3	14.0	70.9
Th	NTh	0.468	42.9	87.5	15.3	63.0
Th	Th	0.144	20.6	92.9	34.5	36.0

Top-1 agreement changes little, indicating redistribution below the argmax. For $a_i = \log \pi_i^T(y_i^S) - \log \pi_i^S(y_i^S)$, mismatch shifts sampled positions from Enc. toward Disc. relative to matched baselines. Since $y_i^S \sim \pi_i^S$, this relative shift is more informative than the absolute sign balance. Together with the prior results, the pattern supports mode transfer over privileged-semantic transfer.

6.2 PRIVILEGED INFORMATION EFFECT

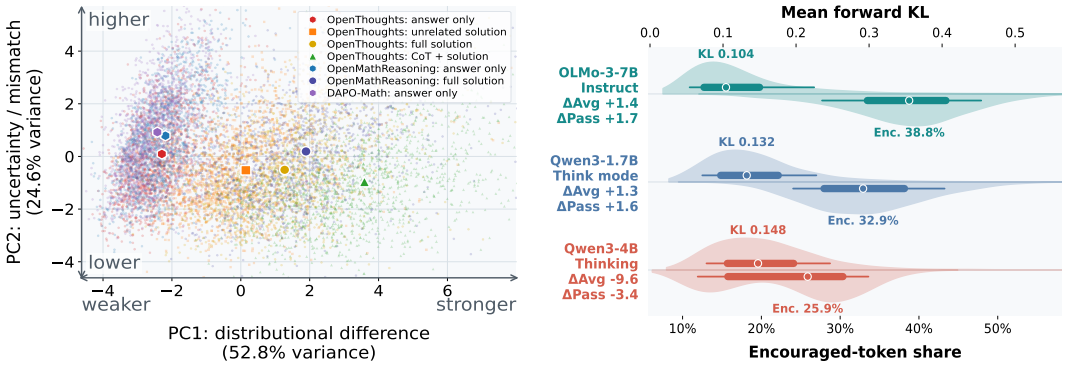
To isolate privileged information, we rescore fixed student rollouts under four conditions as training, Answer only, Full solution, CoT+Solution, and Unrelated solution. To explore whether dataset influences the effect, we add OpenMathReasoning for answer only, full solution and DAPO-Math-17k for answer only. We change privileged information and teacher prompt length while keeping student's input same in all conditions . Panel (a) of Figure 5 applies PCA after within-model standardization. Panel (a) of Table 9 reports raw ten-model macro averages.

The signal shifts progressively from Answer only to CoT+Solution. Even for different datasets, the signals are still clustered in the nearby area. This indicates that the type of privileged information has a greater impact on the signal. Unrelated solution also produces a clear shift despite containing no task-relevant content. These show that teacher's signal therefore reflects the full teacher prefix, different type of privileged information would yield different signal.

Table 8: Model-type signal analysis. HE20: highest-entropy 20%.

Metric	Inst.	Reason.
Mean Fwd. KL	0.151	0.139
Discouraged (%)	44.1	**61.0**
HE20 KL (%)	**57.2**	47.9
HE20 |adv.| (%)	**70.5**	64.6

The known answer can act as an instruction to reconsider the sampled reasoning, making the Instruct–Reasoning contrast another prefix-context effect. Across the paired Qwen3-4B and OLMo-3-7B checkpoints, mean forward KL is similar, but Reasoning has more discouraged tokens while Instruct concentrates more signal at high-entropy positions (Table 8). All four entries in this comparison are computed from the same entropy-analysis rollout pool for each checkpoint. This is consistent with instruction following turning privileged information into targeted corrections that



(a) Context-condition PCA **(b)** Full-solution signal contrast

Figure 5: Privileged-context effects on the initial teacher–student signal. **(a)** PCA of teacher's signal on student rollout across teacher-context conditions. **(b)** Full-solution of OpenThoughts signal distributions across models with different downstream OPSD performances.

can make OPSD more effective. This agrees with Kaur et al. (2026)'s fork-suppression analysis. Our controls further show that prefix mismatch can shift the signal even with task-irrelevant context.

A stronger signal is not necessarily more useful. Panel (b) of Figure 5 contrasts two settings that improve after full-solution OPSD with one that degrades. Their rollout-level signal distributions overlap, so neither raw mean forward KL nor the encouraged-token share predicts the downstream performance. On matched rows, CoT+Solution increases initial divergence but reduces Avg@8 relative to the full solution (Table 9, panel (b)). Training-time forward KL has the same limitation: its change varies across models and rollout caps without tracking downstream performance (Table 10).

Privileged information should therefore be treated as a context-induced distributional intervention. Adding context can strengthen the teacher–student difference without making the supervision more useful. Context construction should control the nature and magnitude of this difference, not maximize context length, information, or initial KL.

Table 9: Privileged-context effects. **(a)** Ten-model OpenThoughts signal averages. $D > .05$, Abs. adv., and Agree. are the high-KL-position rate, mean absolute sampled-token log advantage, and top-1 agreement. **(b)** First-batch forward KL and Avg@8 under a 1,024-token rollout cap. Only teacher context and its prefix cap change. Full per-model signal results are split across Table 26 and Table 27; downstream per-dataset results are in Table 53.

(a) Context signal

Context	Fwd. KL	$D > .05$	Abs. adv.	Agree.
Answer only	0.016	2.9	0.039	98.0
Unrelated solution	0.054	14.9	0.100	95.0
Full solution	0.108	20.3	0.151	93.3
CoT + solution	0.179	30.8	0.224	90.6

(b) Signal and utility

Model	Context	Prefix cap	Fwd. KL	Avg@8
Qwen3-1.7B	Full solution	1,024	0.141	35.3
	CoT + solution	8,192	0.218	30.8
OLMo-3-7B-Think	Full solution	1,024	0.115	58.8
	CoT + solution	8,192	0.205	58.6

6.3 Forward KL at Different Positions

Because OPSD supervises tokens throughout the sampled response, we examine how its signal changes with position.

The result show that both forward KL and SNR decrease substantially toward later positions (Table 11), indicating that the signal is concentrated early in the response. Consistently, First-256 outperforms Final-256 under the same loss-token budget (Table 2). Together with Zhou et al. (2026), this result suggests a common limitation of OPD and OPSD: once a rollout takes a wrong branch, the weaker late-token signal on the student's trajectory may be insufficient or even harmful. OPSD need not compute the loss over an entire trajectory. Applying the loss only to the early tokens of each trajectory can achieve better performance, and it takes less training time and computation.

6.4 OPSD Training Dynamics

We examine whether training-time statistics can diagnose OPSD behavior before evaluation. Specifically, we track student entropy, forward KL to privileged teacher, the clipped distillation loss, and

Table 10: Forward KL and downstream change under full-solution OPSD. KL First10 and Final10 average KL loss over the first and final ten steps. Δ Avg and Δ Pass are Avg@8 and Pass@8 changes from Base. DS-Qwen-1.5B refers to DeepSeek-R1-Distill-Qwen-1.5B.

	256-token rollout cap				1,024-token rollout cap			
Model	KL First10	KL Final10	Δ Avg	Δ Pass	KL First10	KL Final10	Δ Avg	Δ Pass
DS-Qwen-1.5B	0.062	0.073 ↑	+0.7	+0.9	0.032	0.040 ↑	+0.6	−0.8
Falcon-H1R-7B	0.112	0.087 ↓	−0.6	0.0	0.077	0.054 ↓	−0.4	0.0
Qwen3-0.6B	0.125	0.334 ↑	+0.2	+0.8	0.073	0.212 ↑	−2.6	−6.7
MiMo-7B-RL	0.175	0.340 ↑	−0.3	+0.8	0.098	0.243 ↑	−11.3	−10.0
OLMo-3-7B-Think	0.193	0.377 ↑	−2.7	−1.7	0.108	0.251 ↑	−4.7	−4.2
OLMo-3-7B-Instruct	0.206	0.204 ↓	+3.4	+7.5	0.120	0.151 ↑	+1.3	+1.7
Qwen3-1.7B	0.261	0.458 ↑	+4.1	+3.3	0.142	0.336 ↑	+1.3	+1.6
Qwen3-4B-Thinking	0.246	0.341 ↑	−5.1	−3.4	0.156	0.325 ↑	−9.6	−3.4
Qwen3-4B-Instruct	0.260	0.232 ↓	+1.1	−0.9	0.164	0.161 ↓	+1.5	−1.7

Table 11: Forward KL by response-position window. Each entry reports mean/SNR, where $\mathrm{SNR} = \mu/\sigma$ and σ is the token-level standard deviation. Per-model results are in Table 29.

Response position	[0, 128)	[128, 256)	[256, 512)	[512, 1k)	[1k, 2k)	[2k, 4k)	[4k, 6k)
Mean/SNR	0.287 / 0.248	0.190 / 0.217	0.126 / 0.185	0.087 / 0.160	0.067 / 0.138	0.056 / 0.121	0.048 / 0.103

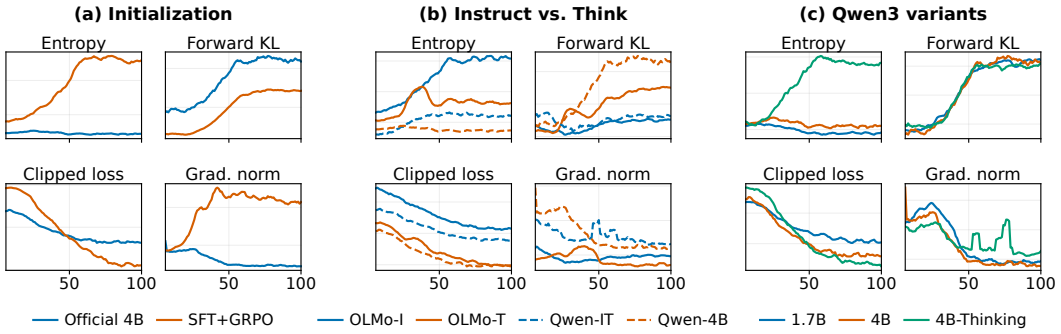


Figure 6: OPSD training dynamics across (a) initialization, (b) Instruct/Think checkpoints, and (c) Qwen3 variants. Per-window results for the initialization comparison are in Table 31.

gradient norm, which characterize output uncertainty, student–teacher discrepancy, and optimization behavior. Figure 6 compares these trajectories across different initializations, Instruct and Think checkpoints, and Qwen3 variants to explore further.

The initialization contrast exhibits the clearest collapse signature: student entropy and gradient norm rise together for the SFT+GRPO actor, whereas both remain bounded or decrease for the released Qwen3-4B. The Instruct/Think comparison shows consistent mode-dependent differences: within both OLMo and Qwen, Instruct checkpoints exhibit lower forward KL and less negative clipped loss than Think checkpoints, without the joint entropy–gradient rise associated with collapse. The Qwen3 variants further show that similar forward KL values can correspond to substantially different downstream performances. Thus, the joint entropy–gradient rise is a useful online warning of collapse, but no single trace reliably predicts final performance.

7 Conclusion

In this paper, we design a comprehensive diagnostic pipeline to study the behavior of OPSD. Controlled experiments and token-level analyses show that OPSD is a sensitive algorithm. The teacher's signal of OPSD is driven by contextual-distribution differences rather than privileged semantics alone. Both of the training configuration such as rollout cap and model itself including model type affect the performance of OPSD. Training dynamics can be used to warn of OPSD collapse rather than predicting downstream performance.

Compared with RLVR, OPSD requires no external reward or stronger privileged teacher, provides dense supervision, and uses less computation. However, OPSD offers no free lunch. Its efficiency comes at the cost of relying on unverified, training-state-dependent supervision. OPSD should therefore be viewed as a complementary post-training method rather than a complete replacement for RLVR. Its effect and generalization remain to be tested on larger models and more diverse tasks.

AI USE STATEMENT

We used GPT-5.6, Cursor Composer, and Grok primarily for routine assistance, including generating portions of auxiliary code (such as dataset-processing code following author-specified procedures), organizing and summarizing experimental outputs, and drafting and polishing the manuscript. The overall research direction, experimental design, data processing strategy, methodology, experimental design, analysis, interpretation of the findings, and conclusions were developed and completed by the authors. Formal proof-related tasks are not applicable. The authors reviewed all AI-assisted materials, independently verified the relevant code, processed data, experimental results, and final manuscript, and take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

To support the accuracy and reproducibility of our diagnostic findings, we provide the accompanying code as supplementary material and document our experimental configurations in detail in the appendix. We also prepare to release the code for reproducibility in the future.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

DeepSeek-AI. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Falcon LLM Team, Iheb Chaabane, Puneesh Khanna, Suhail Mohmad, Slim Frikha, Shi Hu, Abdalgader Abubaker, Reda Alami, Mikhail Lubinets, Mohamed El Amine Seddik, and Hakim Hacid. Falcon-h1r: Pushing the reasoning frontiers with a hybrid model for efficient test-time scaling, 2026. URL https://arxiv.org/abs/2601.02346.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models, 2026. URL https://arxiv.org/abs/2607.05184.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of LLMs?, 2026. URL https://arxiv.org/abs/2603.24472.

Team Kimi. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing, 2026a. URL https://arxiv.org/abs/2604.02288.

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.

Yu Li, Shu Hong, and Tian Lan. Localizing credit at the divergence: Path-conditioned self-distillation for LLM reasoning, 2026c. URL https://arxiv.org/abs/2606.15576.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning, 2026. URL https://arxiv.org/abs/2605.21606.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

LLM-Core-Team Xiaomi. Mimo: Unlocking the reasoning potential of language model – from pretraining to posttraining, 2025. URL https://arxiv.org/abs/2505.07608.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

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

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. AIMO-2 winning solution: Building state-of-the-art mathematical reasoning models with OpenMathReasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025. URL https://arxiv.org/abs/2504.16891.

OLMo Team, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, et al. OLMo 3, 2026. URL https://arxiv.org/abs/2512.13961.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: Reinforcement learning with contrastive on-policy self-distillation, 2026. URL https://arxiv.org/abs/2606.11709.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020. URL https://arxiv.org/abs/1910.02054.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

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.

Samyak Shrestha and Alexander Tessier. Rethinking privileged information in on-policy self-distillation, 2026. URL https://arxiv.org/abs/2608.18271.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models, 2026. URL https://arxiv.org/abs/2604.00626.

Zhiquan Tan and Yinrong Hong. Self-supervised on-policy distillation for reasoning language models, 2026. URL https://arxiv.org/abs/2605.17497.

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.

Jiawei Xu, Minghui Liu, Juzheng Zhang, Tom Goldstein, and Furong Huang. β-OPSD: Deriving with Policy Optimization, Training with Self-Distillation, 2026. URL https://arxiv.org/abs/2607.28582.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR, 2026a. URL https://arxiv.org/abs/2604.03128.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-OPD: Efficient and reliable on-policy distillation for long-horizon reasoning, 2026b. URL https://arxiv.org/abs/2605.07804.

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

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. Fast and effective on-policy distillation from reasoning prefixes, 2026. URL https://arxiv.org/abs/2602.15260.

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

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs, 2024. URL https://arxiv.org/abs/2312.07104.

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

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

APPENDIX CONTENTS

A Extended Related Work 14
B Experimental Setup 15
C Supplementary Experimental Results 20
D Supplementary Analysis Results 28
E Additional Experiments and Analyses 33
F Robustness Checks 37
G Per-Dataset Evaluation Results 40

A Extended Related Work

A.1 RLVR

RLVR optimizes reasoning policies with automatically verifiable outcome rewards. These rewards provide reliable sequence-level supervision but only coarse credit assignment within a reasoning trajectory.

GRPO replaces PPO's learned critic with group-normalized advantages from multiple responses to the same question (Shao et al., 2024). DAPO adds asymmetric clipping, dynamic sampling, token-level normalization, and overlength handling (Yu et al., 2025). GSPO instead applies importance weighting and clipping at the sequence level (Zheng et al., 2025). Recent analyses identify response-length bias in GRPO and suggest that RLVR often reinforces successful trajectories already supported by the base model (Liu et al., 2025; Yue et al., 2025). Although RLVR is widely used in reasoning models(Qwen Team, 2025; Kimi, 2026; DeepSeek-AI, 2025), it requires substantial compute and wall-clock time because each update depends on multiple on-policy rollouts and reward-based optimization. This cost motivates lower-compute dense alternatives such as OPD and OPSD.

A.2 Following Work on OPSD

Recent work improves OPSD by selecting more reliable positions, correcting privileged targets, or combining dense self-distillation with verified outcomes.

Position-Weighted OPSD emphasizes positions with more reliable teacher supervision (Liu et al., 2026). Purified OPSD applies a PMI-based correction to the privileged target (Shen et al., 2026). β-OPSD interpolates teacher and reference distributions and uses return-to-go credit assignment (Xu et al., 2026).

RLSD uses rewards to set the token-update direction and the privileged teacher–student gap to scale its magnitude (Yang et al., 2026a). SRPO routes correct responses to GRPO and failed responses to self-distillation (Li et al., 2026a). RLCSD contrasts correct- and incorrect-hint teachers to reduce privilege-induced style shifts (Pan et al., 2026). SSOPD distills the shortest verified-correct response into prefixes of a long failed response (Tan & Hong, 2026). Hindsight Self-Distillation uses a successful peer to supervise where successful and failed paths diverge (Li et al., 2026c). These methods sometimes improve OPSD by changing which targets, positions, or trajectories are trusted, or by anchoring the update with verified outcomes. However, they primarily redesign the algorithm and do not explain the underlying conditions under which vanilla OPSD works or fails across models and configurations.

A.3 Rethinking OPD and OPSD

Recent analyses question when dense on-policy supervision improves reasoning and when it merely changes behavior or causes degradation.

Privileged supervision can degrade reasoning models on long traces (Kaur et al., 2026). Correct references are not always beneficial, and apparent gains may recover behavior already present in the base model (Shrestha & Tessier, 2026). Li et al. (2026b) show that OPD requires compatible thinking patterns and a teacher with genuinely new capabilities. Zhu et al. (2026) find OPSD sensitive to teacher and loss design, and question whether students internalize privileged information.

Wang et al. (2026) view OPD as an exploration catalyst limited by signal quality, teacher–student mismatch, and length exploitation. Rich privileged context can suppress verbalized uncertainty and harm generalization (Kim et al., 2026). Supervision may concentrate near the response prefix (Zhang et al., 2026), while prefix drift reduces the relevance of later teacher feedback (Yang et al., 2026b). Length inflation, repetition, and truncation can consequently lead to training collapse (Luo et al., 2026). These studies analyze individual factors such as trace length, teacher compatibility, privileged context, or prefix drift in depth. However, they lack a global diagnosis that jointly controls reasoning mode, context content, rollout length, objective, model ability, and post-training stage, and that connects frozen token-level signals to end-to-end training outcomes.

B EXPERIMENTAL SETUP

We report the training, evaluation, and systems settings used throughout the paper. Each ablation changes only the named factor.

B.1 TRAINING FRAMEWORK AND SYSTEMS CONFIGURATION

We implement OPSD with PyTorch, Hugging Face Transformers, and TRL. Distributed jobs are launched with Accelerate and use the DeepSpeed implementation of ZeRO-3 (Rajbhandari et al., 2020) to shard the trainable student, optimizer states, and the frozen teacher across data-parallel workers. All main experiments perform full-parameter student updates. No parameter-efficient adapters are used. Both student and teacher dropout are disabled. Student forward passes use PyTorch scaled dot-product attention, with non-reentrant gradient checkpointing and the key–value cache disabled during training. Parameters and activations use bfloat16, and TF32 matrix multiplication is enabled. Training runs use 2–8 NVIDIA A800 GPUs.

The evaluated models come from the Qwen3, Qwen3.5, DeepSeek-R1, OLMo-3, Falcon-H1R, and MiMo families. A colocated inference engine generates on-policy rollouts: vLLM (Kwon et al., 2023) for Qwen3 and DeepSeek-R1-Distill, and SGLang (Zheng et al., 2024) for Qwen3.5-4B, OLMo-3, Falcon-H1R-7B, and MiMo-7B-RL. The rollout copy of the student is synchronized after every optimizer update. Each worker uses tensor parallelism of 1 for generation and data parallelism otherwise. The vLLM path uses PyTorch 2.6.0, Transformers 4.57.0, TRL 0.22.1, Accelerate 1.7.0, DeepSpeed 0.18.9, and vLLM 0.8.5. The SGLang path uses PyTorch 2.8 and SGLang 0.5.4. Both backends use the same experimental settings.

B.2 TRAINING DATASETS AND PREPROCESSING

B.2.1 OPSD DATA

Unless noted otherwise, OPSD experiments use the OpenThoughts Math 30k training split. The public revision contains 29,434 rows. Field normalization leaves 29,418 records before model-specific prompt-length filtering. Each record provides a mathematics problem, reference solution, final answer, and separate CoT trace. Student and teacher prompts are rendered independently with the model's native chat template and designated reasoning mode. Only the user-side text is reproduced below. We keep examples for which both rendered prompts contain at most 1,024 tokens and filter longer prompts without truncation. Each model–mode configuration is filtered separately because its tokenizer and template differ. Under full-solution conditioning, the resulting sets contain 21,981 examples for the standard Qwen3 models, 21,918 for Qwen3-4B-Thinking(Qwen-3-4B-Thinking-2507), 20,557 for Qwen3.5-4B with a Th teacher, 22,049 for DeepSeek-R1-Distill-Qwen-1.5B, 20,650 for OLMo-3-7B-Think, 21,911 for OLMo-3-7B-Instruct, 16,386 for Falcon-H1R-7B, and 21,642 for MiMo-7B-RL. Teacher-context controls are filtered under their modified teacher templates while retaining the student message. The data sampler uses seed 42.

For the CoT+Solution intervention, we retain the 21,981 Qwen3-1.7B or 20,650 OLMo-3-7B-Think rows and their order, concatenate the CoT and gold-solution fields in the teacher context, and raise the rendered teacher-prompt cap to 8,192 tokens. The cross-dataset replication uses the OpenMath-Reasoning `cot` split (Moshkov et al., 2025) with Qwen3-1.7B. We separate each generated solution into its internal CoT and post-thinking solution, then form subsets with post-thinking lengths of at most 1,024 tokens (609,613 rows) or 2,048–4,096 tokens (15,023 rows). Within each subset, the solution-only and CoT-plus-solution conditions use identical rows and ordering. The solution-only prompt caps are 1,024 and 4,096 tokens, respectively. Both augmented conditions use a 12,288-token cap.

B.2.2 TRAINING PROMPTS

The default student user message is:

```
Problem: {problem}
Please reason step by step, and put your final answer within \boxed{}.
```

All teacher-context experiments use this student message. The teacher-side message varies as follows. Under the full-solution teacher context, the teacher user message is:

```
Problem: {problem}
Here is a reference solution to this problem:
=== Reference Solution Begin ===
{solution}
=== Reference Solution End ===
After reading the reference solution above, make sure you truly understand the reasoning behind each step---do not copy or paraphrase it. Now, using your own words and independent reasoning, derive the same final answer to the problem above. Think step by step, explore different approaches, and don't be afraid to backtrack or reconsider if something doesn't work out:
Please reason step by step, and put your final answer within \boxed{}.
```

Under CoT+Solution, the reference block contains the dataset-provided CoT followed by the gold solution. The student message and sampled rows remain fixed. Under answer-only conditioning, the teacher receives the verified final answer without its derivation:

```
Problem: {problem}
Here is the verified answer:
{answer}
After understanding the privileged information, solve the problem using your own reasoning.
Please reason step by step and put the final answer within \boxed{}.
```

Under the unrelated-solution control, an independently sampled problem–solution pair precedes the target problem and never claims to solve it:

```
Problem: {problem_B}
Solution: {solution_B}
Problem: {problem_A}
Please reason step by step, and put your final answer within \boxed{}.
```

For each retained prompt, the current student produces one sampled completion. The frozen teacher then scores exactly the same completion tokens under its own prompt, and the loss is applied only to non-padding response positions. Thus neither reference-solution tokens nor prompt tokens are direct prediction targets.

For the no-privilege same-prefix control in Table 1, panel (c), both sides receive the default student message above: the teacher receives no solution, answer, or additional problem. The student generates in NTh mode, the teacher scores in Th mode, and evaluation uses Th mode. All other optimization, rollout, and model-specific batch settings are the default OPSD settings below.

B.2.3 Supervised-Initialization Data

The Qwen3-1.7B-Base and Qwen3-4B-Base supervised reasoning initialization uses the NVIDIA OpenMathReasoning `cot` split and excludes the tool-integrated-reasoning, GenSelect, and additional-problem splits. The source split contains 3,201,061 problem–solution rows. We use `problem` as the user turn and `generated_solution` as the assistant target, render both with the Qwen3 chat template in Th mode, and keep sequences of at most 12,288 tokens. The resulting pretokenized corpus contains 2,598,095 examples. Supervised loss applies to assistant tokens, without sequence packing.

OpenMathReasoning provides the supervised initialization, cross-dataset context replication, and the matched-corpus OPSD–GRPO comparison, whereas OpenThoughts provides the problems and privileged teacher context for the main OPSD experiments. The datasets are not mixed within a training batch. Before tokenization and length filtering, we decontaminate both corpora against all 120 evaluation items from AIME 2024–2026 and HMMT25. Matched records are removed, and a second check finds zero overlaps in every final preprocessed training set. We use no training-time validation split or validation-based checkpoint selection. The four test sets in section 4 are reserved for downstream evaluation.

The teacher is a frozen copy of the initial student and receives no gradient or parameter updates. Teacher behavior therefore changes only through context or reasoning mode. The objective contains no task reward, correctness signal, policy-gradient advantage, or auxiliary supervised loss.

B.3 Supervised-Initialization Configuration

We perform full-parameter causal-language-model SFT for both Qwen3-1.7B-Base and Qwen3-4B-Base. Both runs use the 12,288-token supervised corpus above, DeepSpeed ZeRO-2 without CPU offload, a global batch of 64 sequences, 15,000 updates (960,000 sampled sequences), and checkpoints every 500 updates. We use AdamW with peak learning rate 1×10^{-5}, zero weight decay, gradient-norm clipping at 1.0, and a cosine schedule with 10% linear warmup. Training uses bfloat16, gradient checkpointing, no sequence packing, and seed 42. Each base checkpoint is formatted with its model-matched released Qwen3 instruction checkpoint's chat template, with thinking enabled.

The model-specific distributed settings are 4 A800 GPUs, per-device batch size 2, and gradient accumulation 8 for Qwen3-1.7B-Base, and 8 A800 GPUs, per-device batch size 1, and gradient accumulation 8 for Qwen3-4B-Base. Thus both runs have the same global batch and sample budget. Only the model, model-matched chat template, GPU count, and per-device micro-batch differ.

B.4 GRPO Control Configuration

Both GRPO recipes use full-parameter actor updates, one PPO epoch per update, DAPO clip ratios $0.2/0.28$, actor weight decay 0.01, gradient-norm clipping at 1.0, and no entropy bonus. They use a frozen-reference low-variance KL loss with coefficient 0.01 (and no reward-side KL), disable the overlong-response penalty, and offload the frozen reference parameters between scoring passes. Rollouts use eight generations per prompt, temperature 0.6, top-$p = 0.95$, and top-$k = 20$. The reward extracts a boxed answer first and falls back to an `Answer:` pattern. Correctness is checked with Math-Verify.

The GRPO control in subsubsection C.6.2 initializes from the Qwen3-4B-Base SFT-15k checkpoint and trains for 100 updates on DAPO-Math-17k. The corpus is rendered with the Qwen3 Th chat template and filtered to prompts of at most 1,024 tokens. We use 8 A800 GPUs with a colocated hybrid actor/vLLM layout, train batch size 16, PPO mini-batch size 8, learning rate 5×10^{-7}, and a maximum response length of 20,480. Checkpoints are saved every 25 updates. We evaluate the step-100 actor after merging FSDP shards. Decoding at evaluation matches the Qwen3 Th protocol in Table 13. The stacked OPSD run in subsubsection C.6.3 changes only the initialization, setting both student and teacher to the GRPO step-100 actor, and keeps all other Qwen3-4B OPSD settings fixed.

The released-model GRPO runs in subsection C.7 use the integer-answer OpenMathReasoning subset rendered in Th (source solutions capped at 12,288 tokens), with prompts filtered to at most 1,024 tokens. We use a colocated hybrid actor/vLLM layout on 4 A800 GPUs for Qwen3-1.7B and 8 for Qwen3-4B, train batch size 64, PPO mini-batch size 64 (1 optimizer step per trainer step), learning rate 1×10^{-6}, and a maximum response length of 24,576. Actor-parameter and optimizer offload are enabled. We evaluate and merge the step-50 actors and follow the Qwen3 Th protocol in Table 13.

B.5 OPSD Optimization and Rollout Hyperparameters

B.5.1 Default OPSD Configuration

Unless stated otherwise, OPSD uses a frozen same-model teacher, matched native student–teacher reasoning modes, full-solution teacher context, 1,024-token prompt and rollout caps, and full-parameter student updates. Training runs for 100 training steps, with checkpoints saved every 25 updates and update 100 used for evaluation without validation-based selection. We use AdamW (Loshchilov & Hutter, 2019) with peak learning rate 1×10^{-6}, zero weight decay, and gradient-norm clipping at 0.1. The learning rate follows a cosine schedule with linear warmup over the first 10% of updates. The random and data seeds are 42.

For the loss in Equation 3, we set generalized-JSD parameter $\beta = 0$ and pointwise vocabulary-contribution cap $\tau_{\text{clip}} = 0.05$. At temperature 1.1, the implementation computes full-vocabulary

contributions, caps them coordinate-wise, sums over the vocabulary, and averages over selected valid response positions. The top-k ablation in paragraph C.4.3.1 restricts vocabulary support. The entropy-masking study in subsubsection C.4.2 retains the highest- or lowest-entropy response tokens (HE20/HE80 or LE20/LE80). First-256 reuses the equivalent OPSD-256 run; Uni-256 and Final-256 keep a 1,024-token rollout but retain at most 256 uniformly sampled or final valid positions. The AdvT ablation in paragraph C.4.3.2 retains positions where the frozen teacher ranks the sampled token among its top-4 or top-16 outcomes. Other runs average over all valid response positions. Rollout top-k and loss-support k are distinct.

Student rollouts use multinomial sampling with temperature 1.1, top-$p = 0.95$, top-$k = 20$, min-$p = 0$, presence penalty 0, and repetition penalty 1.0. In the rollout-length study, we vary only the rollout cap over $\{128, 256, 512, 1024, 2048, 4096, 6144\}$ tokens where reported and keep all other settings fixed. We rebalance the per-device micro-batch and gradient accumulation at longer rollout lengths to preserve the default model-specific global batch in Table 12.

Architecture-transfer experiments also use Qwen3-4B-Thinking, DeepSeek-R1-Distill-Qwen-1.5B, Falcon-H1R-7B, and MiMo-7B-RL. Table 12 gives the distributed configuration for every OPSD checkpoint. A parenthetical teacher denotes a cross-model experiment. Otherwise, student and teacher use the same initial model.

Table 12: Distributed configurations at the default 1,024-token rollout cap. Global batch equals micro-batch × accumulation × GPUs. The length sweep rebalances micro-batch and accumulation while preserving global batch.

Student model (teacher if different)	Engine	GPUs	Micro-batch	Accumulation	Global batch
Qwen3-0.6B	vLLM	2	16	4	128
Qwen3-1.7B	vLLM	2	8	4	64
Qwen3-4B / Qwen3-4B-Instruct	vLLM	2	4	4	32
Qwen3-4B-Instruct (Qwen3-4B)	vLLM	2	4	4	32
Qwen3-4B-Thinking	vLLM	2	4	4	32
Qwen3-8B	vLLM	4	4	2	32
Qwen3.5-4B	SGLang	2	4	4	32
DeepSeek-R1-Distill-Qwen-1.5B	vLLM	2	8	4	64
OLMo-3-7B-Think / OLMo-3-7B-Instruct	SGLang	4	4	4	64
Falcon-H1R-7B	SGLang	4	2	4	32
MiMo-7B-RL	SGLang	4	2	4	32

Over 100 updates, the configurations in Table 12 correspond to 12,800 sampled trajectories for Qwen3-0.6B, 6,400 for Qwen3-1.7B, DeepSeek-R1-Distill-Qwen-1.5B, and the OLMo-3 models, and 3,200 for every remaining configuration. The Qwen3-1.7B and Qwen3-4B rows also apply, unchanged, when OPSD starts from their Base, SFT, or SFT+GRPO checkpoints.

B.5.2 Main-Text Configuration Coverage

The reasoning-mode comparison uses the default OPSD recipe with only the stated student, teacher, and evaluation modes changed. The no-privilege control is defined above. The extended-training experiment changes the maximum step count to 300 and saves/evaluates every 50 updates. The rollout-length, loss-position, teacher-context, and divergence studies change only the named factor from the default recipe. The initial-model-ability study uses the model rows in Table 12. The post-training-stage study combines the SFT and GRPO recipes above with the default size-matched OPSD recipe, and the compute comparison uses default OPSD versus the released-model GRPO recipe. Token-analysis sampling and scoring settings are reported separately in subsection D.1.

B.6 Evaluation Protocol and Inference Settings

Each of the 30 problems in every test set is evaluated zero-shot with eight sampled completions and seed 42. Base–OPSD comparisons differ only in whether evaluation uses the base model or the corresponding trained checkpoint. All evaluation settings remain fixed. Table 13 lists the model-specific inference defaults. The evaluation user message is:

```
{problem}
Please reason step by step, and put your final answer within \boxed{}.
```

Unlike the training student prompt, evaluation does not prepend a `Problem:` label. The message is rendered with the model's native chat template and the designated Th or NTh mode. MiMo-7B-RL uses an empty system message. All runs use 1 A800 GPU with tensor parallelism 1. The vLLM path uses PyTorch 2.6.0, Transformers 4.57.0, and vLLM 0.8.5 (Kwon et al., 2023). The SGLang path uses PyTorch 2.8 with SGLang 0.5.4 for OLMo-3 and MiMo-7B-RL, and SGLang 0.5.10 for Qwen3.5-4B and Falcon-H1R-7B (Zheng et al., 2024). Answers are extracted from the final balanced `\boxed{}` expression and checked with Math-Verify.

Table 13: Inference settings. Max new and Context are token counts. † For MiMo-7B-RL, 32,768 is the nominal cap. The effective cap is the context window minus the rendered prompt length and one guard token.

Model and eval mode	Engine	Max new	Context	Temp.	Top-p	Top-k	Pres. pen.
Qwen3-0.6/1.7/4/8B, Th	vLLM	38,912	40,960	0.6	0.95	20	0.0
Qwen3-1.7/4/8B, NTh	vLLM	32,768	40,960	0.7	0.80	20	0.0
Qwen3-4B-Instruct, NTh	vLLM	32,768	40,960	0.7	0.80	20	0.0
Qwen3-4B-Thinking	vLLM	81,920	90,112	0.6	0.95	20	0.0
Qwen3.5-4B, Th or NTh	SGLang	81,920	131,072	1.0	0.95	20	1.5
DeepSeek-R1-Distill-Qwen-1.5B	vLLM	32,768	40,960	0.6	0.95	off	0.0
OLMo-3-7B-Think	SGLang	32,768	40,960	0.6	0.95	off	0.0
OLMo-3-7B-Instruct	SGLang	32,768	40,960	0.6	0.95	off	0.0
Falcon-H1R-7B	SGLang	65,536	73,728	0.6	0.95	off	0.0
MiMo-7B-RL	SGLang	32,768†	32,768	0.6	0.95	off	0.0

C SUPPLEMENTARY EXPERIMENTAL RESULTS

This section follows Section 5 and reports the corresponding extended comparisons, ablations, and model-level results.

C.1 VALID COMPARISON SETTING: EXTENDED REASONING-MODE RESULTS

Table 14 gives the model-level results underlying the Qwen3 means in Table 1. Qwen3-1.7B, Qwen3-4B, Qwen3-8B, and Qwen3.5-4B use the same six-configuration grid.

Table 14: Model-level reasoning-mode results underlying Table 1, with the complete Qwen3.5-4B grid. Rows are macro-averaged over four benchmarks. Notation follows Table 1. Per-dataset results are in Table 49.

Model	S	T	Eval	Avg@8			Pass@8			Length		
				Base	OPSD	Δ	Base	OPSD	Δ	Base	OPSD	Δ
Qwen3-1.7B	Th	Th	Th	34.0	35.3	+1.3	59.2	60.8	+1.6	18,624	22,380	+20.2%
Qwen3-1.7B	Th	NTh	Th	34.0	19.2	−14.8	59.2	39.2	−20.0	18,624	11,109	−40.4%
Qwen3-1.7B	NTh	Th	NTh	9.8	14.1	+4.3	20.8	28.3	+7.5	3,923	9,731	+148.1%
Qwen3-1.7B	NTh	NTh	NTh	9.8	9.2	−0.6	20.8	20.8	0.0	3,923	8,463	+115.7%
Qwen3-1.7B	NTh	Th	Th	34.0	40.8	+6.8	59.2	63.3	+4.1	18,624	19,996	+7.4%
Qwen3-1.7B	Th	NTh	NTh	9.8	7.5	−2.3	20.8	20.0	−0.8	3,923	3,003	−23.5%
Qwen3-4B	Th	Th	Th	62.3	57.9	−4.4	79.2	75.8	−3.4	17,769	19,949	+12.3%
Qwen3-4B	Th	NTh	Th	62.3	47.9	−14.4	79.2	69.2	−10.0	17,769	15,108	−15.0%
Qwen3-4B	NTh	Th	NTh	19.6	21.5	+1.9	38.3	45.8	+7.5	6,694	17,204	+157.0%
Qwen3-4B	NTh	NTh	NTh	19.6	17.2	−2.4	38.3	32.5	−5.8	6,694	15,907	+137.6%
Qwen3-4B	NTh	Th	Th	62.3	63.0	+0.7	79.2	79.2	0.0	17,769	19,294	+8.6%
Qwen3-4B	Th	NTh	NTh	19.6	16.0	−3.6	38.3	30.0	−8.3	6,694	5,538	−17.3%
Qwen3-8B	Th	Th	Th	64.1	58.9	−5.2	78.3	76.7	−1.6	17,647	21,422	+21.4%
Qwen3-8B	Th	NTh	Th	64.1	42.8	−21.3	78.3	70.0	−8.3	17,647	13,655	−22.6%
Qwen3-8B	NTh	Th	NTh	18.0	23.2	+5.2	37.5	45.0	+7.5	5,914	15,343	+159.4%
Qwen3-8B	NTh	NTh	NTh	18.0	19.1	+1.1	37.5	32.5	−5.0	5,914	16,533	+179.5%
Qwen3-8B	NTh	Th	Th	64.1	64.8	+0.7	78.3	76.7	−1.6	17,647	19,292	+9.3%
Qwen3-8B	Th	NTh	NTh	18.0	18.5	+0.5	37.5	33.3	−4.2	5,914	6,771	+14.5%
Qwen3.5-4B	Th	Th	Th	80.6	0.0	−80.6	91.7	0.0	−91.7	38,096	81,920	+115.0%
Qwen3.5-4B	Th	NTh	Th	80.6	8.6	−72.0	91.7	27.5	−64.2	38,096	73,214	+92.2%
Qwen3.5-4B	NTh	Th	NTh	52.7	0.0	−52.7	77.5	0.0	−77.5	10,193	81,920	+703.7%
Qwen3.5-4B	NTh	NTh	NTh	52.7	8.0	−44.7	77.5	35.0	−42.5	10,193	70,710	+593.7%
Qwen3.5-4B	NTh	Th	Th	80.6	0.0	−80.6	91.7	0.0	−91.7	38,096	81,920	+115.0%
Qwen3.5-4B	Th	NTh	NTh	52.7	36.3	−16.4	77.5	64.2	−13.3	10,193	13,649	+33.9%

The model-level results support the conclusion in subsection 5.1: reasoning-mode mismatch transfers mode-specific behavior rather than privileged semantics. Qwen3.5-4B is a collapse case under both matched modes and is analyzed in subsection E.4.

C.2 TRAINING STEPS, ROLLOUT LENGTH, AND LOSS-TOKEN POSITION

Figures 2 and 3 and Table 2 report the model-level trends. First-256 reuses OPSD-256; Uni-256 and Final-256 use 1,024-token rollouts and retain uniformly sampled or final valid response positions. The results match the findings in subsection 5.2 and subsection 5.2: longer training and rollouts do not reliably improve accuracy, and early positions perform best under the fixed 256-position budget. section G gives the benchmark-level values, and subsection F.1 gives the seed repetitions.

C.3 TEACHER-SIDE CONTEXT: PREFIX-LENGTH STRATIFICATION

At a fixed 256-token rollout length, we compare standard OPSD (Full) with training on gold-trace length deciles 0–4 (Short-prefix) or 7–9 (Long-prefix). Consistent with subsection 5.3, longer teacher prefixes provide no consistent advantage; the ordering varies across models (Table 15).

Table 15: Teacher-prefix-length stratification at a 256-token rollout length. Short-prefix and Long-prefix retain gold-trace length deciles 0–4 and 7–9. Δ is the OPSD–Base change (percentage points for accuracy, relative for Length). Bold marks the best Avg@8 and Pass@8 among Full, Short-prefix, and Long-prefix within each model (ties are all bolded). Per-dataset results are in Table 72.

Model	Subset	Avg@8			Pass@8			Length		
		Base	OPSD	Δ	Base	OPSD	Δ	Base	OPSD	Δ
Qwen3-1.7B	Full	34.0	**38.1**	+4.1	59.2	**62.5**	+3.3	18.6k	21.4k	+15.2%
	Short-prefix	34.0	37.1	+3.1	59.2	**62.5**	+3.3	18.6k	20.9k	+12.3%
	Long-prefix	34.0	36.7	+2.7	59.2	**62.5**	+3.3	18.6k	20.8k	+11.5%
Qwen3-4B	Full	62.3	60.9	−1.4	79.2	75.8	−3.4	17.8k	19.1k	+7.4%
	Short-prefix	62.3	**61.0**	−1.3	79.2	**79.2**	+0.0	17.8k	19.0k	+6.9%
	Long-prefix	62.3	60.6	−1.7	79.2	74.2	−5.0	17.8k	19.1k	+7.4%
Qwen3-4B-Instruct	Full	48.5	**49.6**	+1.1	71.7	**70.8**	−0.9	9.1k	10.8k	+18.0%
	Short-prefix	48.5	48.9	+0.4	71.7	68.3	−3.4	9.1k	11.0k	+20.5%
	Long-prefix	48.5	48.9	+0.4	71.7	67.5	−4.2	9.1k	10.6k	+15.8%
OLMo-3-7B-Instruct	Full	39.5	42.9	+3.4	60.8	**68.3**	+7.5	7.6k	8.3k	+9.8%
	Short-prefix	39.5	**44.4**	+4.9	60.8	65.8	+5.0	7.6k	8.1k	+7.4%
	Long-prefix	39.5	41.7	+2.2	60.8	64.2	+3.4	7.6k	8.2k	+8.2%

C.4 Local Loss-Design Ablations

These ablations change the distributional contributions or response positions included in the OPSD objective.

C.4.1 Divergence-Objective Ablation

The official OPSD implementation uses the forward-KL endpoint of generalized JSD ($\beta = 0$). We additionally train with symmetric JSD ($\beta = 0.5$) and reverse KL ($\beta = 1$), holding all other settings fixed.

C.4.1.1 Loss Definitions and Implementation. At a valid response position i, write $\pi_i^S(v) = \pi_\theta(v \mid Q, y^S_{<i})$ for the active student and $\pi_i^T(v) = \pi_{\theta_0}(v \mid Q, C, y^S_{<i})$ for the teacher. For any two next-token distributions a, b on the vocabulary $\mathcal{V}$,

$$D_{\mathrm{KL}}(a\|b) = \sum_{v\in\mathcal{V}} a(v) \log \frac{a(v)}{b(v)}. \tag{5}$$

Accordingly, the two KL endpoints evaluated here are

$$\ell_i^{\mathrm{fwd}} = D_{\mathrm{KL}}(\pi_i^T\|\pi_i^S), \qquad \ell_i^{\mathrm{rev}} = D_{\mathrm{KL}}(\pi_i^S\|\pi_i^T). \tag{6}$$

For $0 < \beta < 1$, let $m_{\beta,i} = (1-\beta)\pi_i^S + \beta\pi_i^T$ and use the generalized-JSD interior

$$\begin{aligned} \ell_i^{\mathrm{JSD}_\beta} &= \beta D_{\mathrm{KL}}(\pi_i^T\|m_{\beta,i}) + (1-\beta)D_{\mathrm{KL}}(\pi_i^S\|m_{\beta,i}), \\ \ell_i^{\mathrm{JSD}_{0.5}} &= \tfrac{1}{2}D_{\mathrm{KL}}\left(\pi_i^T\middle\|\tfrac{\pi_i^S+\pi_i^T}{2}\right) + \tfrac{1}{2}D_{\mathrm{KL}}\left(\pi_i^S\middle\|\tfrac{\pi_i^S+\pi_i^T}{2}\right). \end{aligned} \tag{7}$$

The implementation dispatches the two endpoints separately rather than taking a literal $\beta \to 0$ or $\beta \to 1$ limit of Equation 7: $\beta = 0$ selects ℓ_i^{fwd}, $\beta = 0.5$ selects the symmetric JSD above, and $\beta = 1$ selects ℓ_i^{rev}. The option named `jsd_token_clip` clips each pointwise vocabulary contribution to 0.05 before summing over v and averaging over valid response positions. Because negative contributions are not lower-clipped, the resulting surrogate can be negative.

Our forward-KL branch matches the official OPSD objective: both implementations use full-vocabulary distributions and pointwise-before-sum clipping. Our trainer supports separate student-sampling and teacher-scoring temperatures and arbitrary loss masks. Here both temperatures are 1.1 and all valid response tokens are retained. We freeze a full-parameter teacher copy rather than the LoRA teacher documented in the official fixed-teacher path, without changing the per-token loss.

Changing divergence direction redistributes outcomes rather than providing a reliable remedy, as in subsection 5.4. Forward KL leads in some settings, while reverse KL or symmetric JSD leads in others (Table 16). paragraph C.4.3.1 further restricts the divergence to a vocabulary top-k support.

Table 16: Macro-average divergence-ablation results over four benchmarks. Δ is the absolute change from Base for Avg@8/Pass@8 and the relative change for Length. Bold marks the best Avg@8 and Pass@8 within each model. Decoding settings are fixed across objectives. Per-dataset results are in Table 55.

Model	Objective	Avg@8		Pass@8		Length	
		Value	Δ	Value	Δ	Value	Δ
Qwen3-1.7B	Fwd. KL ($\beta = 0$)	**35.31**	+1.3	**60.83**	+1.6	22.4k	+20.3%
	JSD ($\beta = 0.5$)	31.25	−2.7	55.00	−4.2	15.2k	−18.4%
	Rev. KL ($\beta = 1$)	26.46	−7.5	47.50	−11.7	24.4k	+31.0%
Qwen3-4B	Fwd. KL ($\beta = 0$)	**57.92**	−4.4	**75.83**	−3.4	19.9k	+12.0%
	JSD ($\beta = 0.5$)	56.56	−5.7	**75.83**	−3.4	14.1k	−20.7%
	Rev. KL ($\beta = 1$)	31.98	−30.3	56.67	−22.5	27.9k	+57.0%
Qwen3-4B-Thinking	Fwd. KL ($\beta = 0$)	65.83	−9.6	83.33	−3.4	43.4k	+90.5%
	JSD ($\beta = 0.5$)	68.02	−7.4	82.50	−4.2	17.7k	−22.3%
	Rev. KL ($\beta = 1$)	**74.17**	−1.2	**85.00**	−1.7	34.1k	+49.7%
OLMo-3-7B-Think	Fwd. KL ($\beta = 0$)	58.75	−4.7	77.50	−4.2	21.7k	+15.8%
	JSD ($\beta = 0.5$)	62.08	−1.4	81.67	0.0	17.3k	−7.7%
	Rev. KL ($\beta = 1$)	**62.40**	−1.1	**82.50**	+0.8	19.9k	+6.2%
OLMo-3-7B-Instruct	Fwd. KL ($\beta = 0$)	40.83	+1.3	**62.50**	+1.7	7.4k	−2.1%
	JSD ($\beta = 0.5$)	37.19	−2.3	58.33	−2.5	7.5k	−0.8%
	Rev. KL ($\beta = 1$)	**42.50**	+3.0	61.67	+0.9	8.5k	+12.4%
Falcon-H1R-7B	Fwd. KL ($\beta = 0$)	**89.06**	−0.4	**96.67**	0.0	22.1k	+6.6%
	JSD ($\beta = 0.5$)	88.44	−1.1	95.83	−0.9	20.4k	−1.6%
	Rev. KL ($\beta = 1$)	88.13	−1.4	**96.67**	0.0	20.4k	−1.3%
MiMo-7B-RL	Fwd. KL ($\beta = 0$)	44.27	−11.3	64.17	−10.0	16.7k	−0.3%
	JSD ($\beta = 0.5$)	**52.60**	−3.0	**72.50**	−1.7	14.1k	−15.5%
	Rev. KL ($\beta = 1$)	50.21	−5.4	**72.50**	−1.7	19.6k	+17.3%

C.4.2 Entropy-Based Token Selection

Table 17: Selected macro-average results for student-entropy loss masking at a 256-token rollout length. Parentheses show changes from Base (percentage points for Avg@8/Pass@8, relative percentage for Length). HE/LE retain the highest/lowest-entropy 20%/80%. Bold marks the best accuracy per model. Complete results are in Table 56.

Model	Subset	Avg@8	Pass@8	Length
Qwen3-1.7B	Full	**38.1** (+4.1)	62.5 (+3.3)	21.4k (+15.2%)
	HE20	36.9 (+2.9)	**65.8** (+6.6)	21.7k (+16.7%)
	HE80	37.8 (+3.8)	61.7 (+2.5)	20.8k (+11.6%)
	LE20	37.6 (+3.6)	64.2 (+5.0)	18.0k (−3.5%)
	LE80	33.2 (−0.8)	59.2 (+0.0)	17.7k (−5.2%)
Qwen3-4B-Thinking	Full	70.3 (−5.1)	83.3 (−3.4)	30.5k (+34.0%)
	HE20	71.8 (−3.6)	**85.0** (−1.7)	30.1k (+32.2%)
	HE80	72.1 (−3.3)	83.3 (−3.4)	30.2k (+32.6%)
	LE20	73.9 (−1.5)	**85.0** (−1.7)	22.5k (−1.3%)
	LE80	**74.5** (−0.9)	83.3 (−3.4)	26.0k (+14.1%)
OLMo-3-7B-Instruct	Full	42.9 (+3.4)	68.3 (+7.5)	8.3k (+9.8%)
	HE20	42.2 (+2.7)	**69.2** (+8.4)	8.1k (+7.0%)
	HE80	42.3 (+2.8)	68.3 (+7.5)	8.1k (+6.4%)
	LE20	40.2 (+0.7)	66.7 (+5.9)	7.9k (+3.7%)
	LE80	**44.1** (+4.6)	**69.2** (+8.4)	8.5k (+11.3%)

We test whether OPSD should focus on positions where the student is uncertain. At the 256-token rollout length, HE20 and HE80 retain the highest-entropy 20% and 80% of response positions, LE20 and LE80 retain the corresponding lowest-entropy fractions, and Full retains all positions. All other settings are fixed. Table 17 reports selected macro averages.

Entropy-based position selection provides no consistent improvement over Full, matching subsection 5.4. Low-entropy variants more often shorten or stabilize outputs, whereas high-entropy variants more often retain standard OPSD's length growth (Table 56). subsubsection C.4.4 reports the corresponding initial token-level signals.

C.4.3 SELECTIVE-SUPERVISION ABLATIONS

We test two ways to concentrate OPSD supervision: restricting the divergence to a small vocabulary support and retaining only response positions favored by the teacher. The two panels in Table 18 use their respective rollout lengths and summarize the model-dependent trends in subsection 5.4.

Table 18: Selective OPSD supervision: Avg@8 change from Base (percentage points). Panels (a) and (b) use 1,024- and 256-token rollouts. Full metrics are in Table 19 and Table 20.

(a) Vocabulary support (1,024-token rollout)

Model	Full	Top-16	Top-4
Qwen3-1.7B	+1.3	−1.0	−8.2
Qwen3-4B-Instruct	+1.5	+3.5	−2.8
OLMo-3-7B-Think	−1.1	−3.6	−60.0

(b) Teacher-favored positions (256-token rollout)

Model	Full	AdvT4	AdvT16
Qwen3-1.7B	+4.1	+0.7	+3.2
Qwen3-4B	−1.4	−1.8	−0.1
Qwen3-4B-Thinking	−5.1	−1.0	−0.2
OLMo-3-7B-Think	−2.7	−1.2	−1.7

As in subsection 5.4, neither vocabulary truncation nor teacher-favored position selection improves consistently across models.

C.4.3.1 Vocabulary Top-k Distillation Loss. The default OPSD loss in Equation 3 is a full-vocabulary divergence. We vary only its vocabulary support: the top-k variant evaluates per-token KL on k tokens and renormalizes both student and teacher distributions on that support, while all other settings remain fixed.

Forward KL ($\beta = 0$) takes the teacher's top-k tokens and uses $D_{\mathrm{KL}}(\widetilde{\pi}_i^T \| \widetilde{\pi}_i^S)$. Reverse KL ($\beta = 1$) takes the student's top-k tokens and uses $D_{\mathrm{KL}}(\widetilde{\pi}_i^S \| \widetilde{\pi}_i^T)$.

We report three models: Qwen3-1.7B in matched Th/Th with forward KL, Qwen3-4B-Instruct in matched NTh/NTh with forward KL, and OLMo-3-7B-Think in matched Th/Th with reverse KL, the stronger OLMo endpoint in Table 16. Table 19 compares $k \in \{4, 16\}$ against the matched full-vocabulary run.

Table 19: Macro-average vocabulary top-k loss results. The Qwen3 rows use forward KL with teacher Top-k. OLMo-3-7B-Think uses reverse KL with student Top-k. Full-vocab rows are the matched controls from Table 16. Δ is the absolute change from Base for Avg@8/Pass@8 and the relative change for Length. Bold marks the best Avg@8 and Pass@8 within each model. Ties are all bolded. Per-dataset results are in Table 57.

Model	Loss support	Avg@8		Pass@8		Length	
		Value	Δ	Value	Δ	Value	Δ
Qwen3-1.7B	Full vocab	**35.31**	+1.3	60.83	+1.6	22.4k	+20.3%
	$k = 16$	33.02	−1.0	**62.50**	+3.3	22.6k	+21.4%
	$k = 4$	25.83	−8.2	47.50	−11.7	28.8k	+54.6%
Qwen3-4B-Instruct	Full vocab	50.00	+1.5	70.00	−1.7	10.6k	+15.2%
	$k = 16$	**51.98**	+3.5	**70.83**	−0.9	10.6k	+15.4%
	$k = 4$	45.73	−2.8	69.17	−2.5	8.2k	−10.9%
OLMo-3-7B-Think	Full vocab	**62.40**	−1.1	**82.50**	+0.8	19.9k	+6.2%
	$k = 16$	59.90	−3.6	81.67	0.0	20.2k	+7.9%
	$k = 4$	3.54	−60.0	23.33	−58.4	32.2k	+71.9%

Vocabulary truncation provides no reliable improvement: Top-16 remains close to the full-vocabulary objective, whereas Top-4 degrades more sharply across the tested models (Table 19).

C.4.3.2 Teacher-Favored Position Selection (AdvT4 / AdvT16). The default OPSD loss averages over all valid response tokens. We vary only the position-selection rule: AdvTk restricts the per-completion average to positions where the frozen teacher ranks the sampled student token among its top-k vocabulary outcomes (`pos_adv_teacher_topk`$= k$), with $k \in \{4, 16\}$ and all

other settings fixed. Table 20 compares Base, the matched 256-token full-loss control, and both AdvT checkpoints.

Table 20: Teacher-favored position selection (AdvT4 / AdvT16) versus the matched 256-token full-loss control. Bold marks the best Avg@8 / Pass@8 within each model. Values are unweighted averages over AIME 2024–2026 and HMMT25.

Model	Stage	Avg@8	Pass@8	Length
Qwen3-1.7B	Base	34.0	59.2	18.6k
	c256	**38.1**	**62.5**	21.4k
	AdvT4	34.7	59.2	19.8k
	AdvT16	37.2	61.7	19.5k
Qwen3-4B	Base	**62.3**	**79.2**	17.8k
	c256	60.9	75.8	19.1k
	AdvT4	60.5	76.7	18.4k
	AdvT16	62.2	78.3	18.0k
Qwen3-4B-Thinking	Base	**75.4**	**86.7**	22.8k
	c256	70.3	83.3	30.5k
	AdvT4	74.4	85.8	26.9k
	AdvT16	75.2	85.0	26.8k
OLMo-3-7B-Think	Base	**63.5**	81.7	18.7k
	c256	60.8	80.0	21.4k
	AdvT4	62.3	**82.5**	21.2k
	AdvT16	61.9	80.8	21.4k

Neither AdvT4 nor AdvT16 consistently improves over the matched 256-token full-loss control. AdvT16 remains closer to the control on the Qwen3 models, while AdvT4 produces the milder OLMo degradation (Table 20).

C.4.4 Entropy-Stratified Signal by Checkpoint

Signal concentration does not imply training utility: entropy-based position selection provides no consistent benefit, as in subsection 5.4, although initial divergence concentrates at high-entropy positions (Table 21).

Table 21: Token-level entropy-tail signals for Table 17. HE20 KL/Adv are the highest-entropy 20% shares of total forward KL and absolute advantage, with each total computed from the same entropy-analysis rollout pool.

Model	Mode	Highest 20%		Lowest 20%		HE20 share	
		KL	$D > .05$	KL	$D > .05$	KL	Adv
Qwen3-1.7B	Th/Th	0.371	66.4	0.004	0.1	51.5	69.8
Qwen3-1.7B	NTh/NTh	0.277	61.5	0.003	0.1	55.3	72.9
Qwen3-4B	Th/Th	0.349	68.4	0.002	0.1	51.2	68.4
Qwen3-4B	NTh/NTh	0.221	57.2	0.002	0.1	60.1	74.4
Qwen3-4B-Instruct	NTh/NTh	0.522	66.0	0.003	0.1	59.7	76.3
Qwen3-4B-Thinking-2507	Th/Th	0.405	71.1	0.004	0.2	50.0	69.1
OLMo-3-7B-Think	Th/Th	0.266	62.9	0.006	0.3	45.9	60.1
OLMo-3-7B-Instruct	NTh/NTh	0.347	70.5	0.004	0.3	54.8	64.7

C.5 Initial Model Ability

Table 58 reports the complete benchmark-level results behind Figure 4, including output length. The breakdown supports the conclusion in subsection 5.5: gains are model-dependent, with improvements confined to a few lower-ability or Instruct checkpoints and degradation on most stronger reasoning checkpoints.

C.6 Post-Training Stage and SFT Initialization Sweep

We initialize OPSD from Qwen3-1.7B-Base and Qwen3-4B-Base checkpoints after 0, 5,000, 10,000, or 15,000 SFT updates. Student and teacher use the same checkpoint. subsection B.3 gives the SFT configurations.

We evaluate each SFT checkpoint directly on AIME 2024–2026 and HMMT25, then run OPSD with the same checkpoint for student and teacher. Reported values are unweighted benchmark averages. Checkpoints with extended decoding use the 38,912-token evaluation.

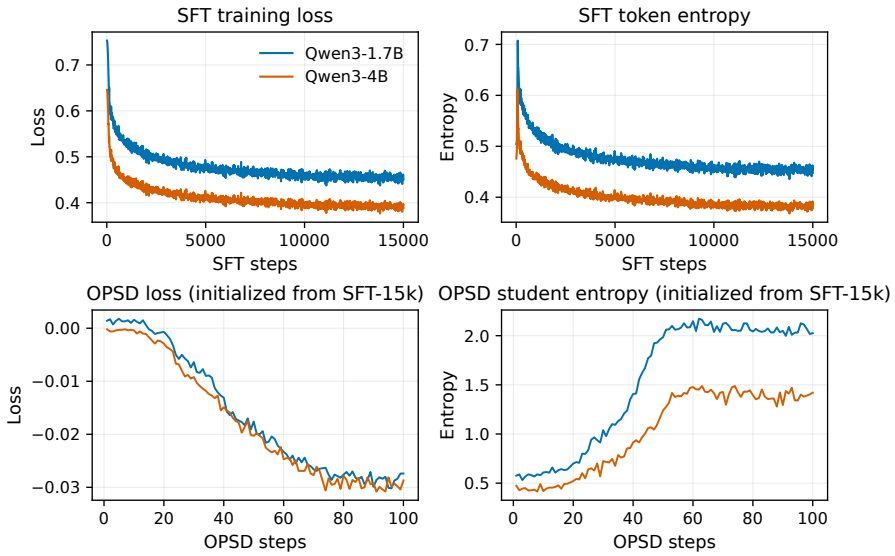


Figure 7: Training traces for the SFT-initialization sweep. Top: SFT loss and token entropy for Qwen3-1.7B-Base and Qwen3-4B-Base. Bottom: loss and student entropy during 100-step OPSD runs from the 15k SFT checkpoints.

Table 22: Macro-average direct-SFT and downstream OPSD results. Format is the boxed-answer rate (%). Downstream OPSD initializes both student and teacher from the SFT checkpoint. Bold marks the best Avg@8 and Pass@8 within each model and stage. Per-checkpoint and per-dataset results, including the additional 7.5k and 12.5k direct-SFT evaluations, are in Table 59 and Table 60.

Model	SFT steps	Direct SFT			After 100-step OPSD			
		Avg@8	Pass@8	Format	Avg@8	Pass@8	Format	Length
Qwen3-1.7B-Base	Base	2.08	8.33	80.63	1.15	7.50	63.23	8.0k
	5k	7.08	21.67	51.98	0.31	2.50	0.83	34.6k
	10k	9.59	20.83	64.58	**3.12**	**11.67**	6.56	35.2k
	15k	**11.46**	**29.17**	63.54	1.56	7.50	3.85	36.2k
Qwen3-4B-Base	Base	7.29	20.00	87.19	**4.79**	**17.50**	74.59	10.3k
	5k	25.63	50.83	76.04	0.83	4.17	2.08	33.1k
	10k	32.29	**60.84**	90.52	1.67	8.33	4.17	32.8k
	15k	**34.69**	60.00	89.27	1.98	8.33	3.44	32.9k

OPSD remains sensitive to the post-training stage, as in subsection 5.6. Direct SFT ability rises with training, but OPSD collapses from the partially SFT-trained checkpoints (Table 22).

C.6.1 Output Degeneration Examples

For the unstable SFT initializations, we inspect all eight samples for each of the 120 evaluation problems (960 generations per row). Consistent with the failure mode in subsection 5.6, Table 23 shows a shift from partially successful direct-SFT generations to repetitive, non-terminating outputs with few boxed answers after OPSD. The length increase reflects collapse, not longer successful reasoning.

Table 23: Generation-level diagnostics pooled over AIME 2024–2026 and HMMT25. Boxed is the fraction of samples with a boxed final answer. Avg@8 is the fraction of correct samples. OPSD denotes evaluation after 100-step OPSD from that SFT checkpoint.

Checkpoint	Cond.	N	Tokens	$\geq$32k	$\geq$38k	Boxed	Avg@8
1.7B SFT-5k	direct	960	23.8k	39.8%	19.5%	52.0%	7.1%
1.7B SFT-5k	OPSD	960	34.7k	48.0%	32.4%	0.8%	0.3%
1.7B SFT-10k	OPSD	960	35.2k	71.2%	55.4%	6.6%	3.1%
4B SFT-5k	OPSD	960	33.1k	80.0%	0.0%	2.1%	0.8%

On AIME 2024 problem 0 (answer 204), the direct 1.7B SFT-5k checkpoint produces three correct boxed answers among eight samples. One trajectory correctly uses the two stated total times:

```
[AIME 2024-0, Qwen3-1.7B SFT-5k, direct, excerpt]
9/s = 4 - t/60 9/(s + 2) = 2.4 - t/60
... final answer: \boxed{204}
```

After 100-step OPSD, none of the eight samples contains a boxed or correct answer, and all span 32.6k–38.9k tokens. In the 38,912-token trajectory below, the model leaves the stated four-hour total as an unknown and repeatedly searches for a nonexistent third constraint (ellipses denote omitted tokens):

```
[AIME 2024-0, same problem, Qwen3-1.7B SFT-5k after 100-step OPSD, excerpt]
Total Time T1 = (9/s) + (t/60) = ?
Alternatively perhaps there is a third condition involving different speeds
and times? Alternatively perhaps I can get another equation from another
scenario where she walks slower than s?
Hmm. Hmm. ... [repeated 15,134x: Hmm. 78x: Alternatively]
. . . [no \boxed{} answer, 38,912 tokens]
```

The 4B SFT-5k checkpoint after OPSD exhibits the same failure to enter a final-answer mode on AIME 2025 problem 0. Its 33,583-token unboxed trajectory ends as follows:

```
[AIME 2025-0, Qwen3-4B SFT-5k after 100-step OPSD, final characters]
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! ... [no \boxed{} answer, 33,583 tokens]
```

C.6.2 GRPO Control on the Collapsing 4B SFT Initialization

We continue the Qwen3-4B-Base SFT-15k checkpoint with 100 GRPO updates under a DAPO-style verifiable math reward. This checkpoint has the strongest direct SFT result in Table 22 and collapses under 100-step OPSD. subsection B.4 gives the GRPO configuration.

OPSD's failure mode remains pipeline-dependent, as in subsection 5.6: from the same initialization, GRPO is approximately flat and format-stable, whereas OPSD collapses; applying OPSD after GRPO collapses again (Table 5, panel (b)).

C.6.3 OPSD after SFT then GRPO

We next initialize both student and teacher from the GRPO step-100 actor and apply the Qwen3-4B OPSD protocol without other changes.

Table 24: Matched Th/Th OPSD from released Qwen3-4B and the SFT-then-GRPO checkpoint. Per-dataset results and training traces are in Table 64 and Table 31.

Initialization	Stage	Avg@8	Pass@8	Format	Length
Qwen3-4B (Th/Th)	Base	62.3	79.2	97.1	17.8k
	+ 100-step OPSD	57.9	75.8	86.4	19.9k
4B SFT-15k+GRPO	GRPO-100	33.65	60.83	94.79	14.2k
	+ 100-step OPSD	3.54	14.17	5.62	37.6k

The pipeline-stage conclusion in subsection 5.6 also holds here: released Qwen3-4B degrades without collapse, whereas the SFT-then-GRPO checkpoint collapses after OPSD. Their logs separate early: entropy and gradient norm rise for the SFT-then-GRPO initialization but remain bounded or decrease for released Qwen3-4B (subsection 6.4).

C.7 OPSD versus GRPO on Released Qwen3 Thinking Models

For the OPSD–GRPO comparison in Table 6, we train both methods from the released Qwen3-1.7B and Qwen3-4B models in their native Th mode and evaluate after 50 updates. Both methods consume exactly the same 3,200 examples from the OpenMathReasoning integer-answer subset in the same order: we materialize the first 3,200 rows produced by the GRPO random sampler (seed 1) as the sequential OPSD dataset. GRPO uses the final answer only for reward verification, whereas OPSD supplies the corresponding post-thinking solution as privileged teacher context.

The matched OPSD runs use a frozen same-model teacher, full-parameter updates, a 1,024-token rollout cap, global batch size 64, learning rate 1×10^{-6}, forward KL, token-contribution clipping at

0.05, and seed 42. Both model sizes train on 2 A800 GPUs; Qwen3-1.7B uses per-device batch size 8 with gradient accumulation 4, and Qwen3-4B uses per-device batch size 4 with gradient accumulation 8. The longest rendered prompt in the materialized dataset is 2,584 tokens. subsection B.4 gives the GRPO configuration, and Table 63 and Table 62 report the benchmark-level results.

The cost column in Table 6 uses training-loop time through the evaluated checkpoint and excludes evaluation and data preprocessing. The results support the trade-off in subsection 5.7: neither method dominates accuracy, while OPSD uses substantially fewer GPU-hours under these recipes. At 1.7B, OPSD has higher Avg@8 and equal Pass@8; at 4B, GRPO leads Avg@8 and OPSD leads Pass@8.

D SUPPLEMENTARY ANALYSIS RESULTS

D.1 TOKEN-LEVEL ANALYSIS DETAILS

This subsection reports the protocol and per-model results underlying section 6. Unless noted otherwise, the analysis uses 2,048 prompts with two on-policy rollouts each; the position-window study uses a 6,144-token rollout cap.

D.1.1 TOKEN-LEVEL ANALYSIS PROTOCOL

We compute the token-level statistics in section 6 offline from the same student and teacher distributions used in training. For each initial checkpoint, we sample 2,048 training prompts with two on-policy rollouts each ($n = 4{,}096$) at temperature 1.1 and a 1,024-token rollout cap. The frozen self-teacher scores the visited response tokens under its privileged context. Reported forward KL is the vocabulary-summed, unclipped $\beta = 0$ divergence. Unless explicitly described as shared, the matched-mode, teacher-context, and entropy analyses use independent rollout draws under this common protocol. The model-type summary in Table 8 computes all of its metrics within the entropy-analysis draw for each checkpoint.

The $D_i^{\mathrm{fwd}} > 0.05$ statistic thresholds this summed diagnostic and is distinct from the training-time pointwise contribution cap. Encourage and Discourage are the fractions of positions with sampled-token log advantage $a_i > 0$ and $a_i < 0$. The remainder are ties at stored precision. Because $y_i^S \sim \pi_i^S$, $\mathbb{E}_{y_i^S \sim \pi_i^S}[a_i] = -D_{\mathrm{KL}}(\pi_i^S \| \pi_i^T) \leq 0$, so student-policy sampling yields a non-positive expectation. Measurements at initialization exclude student drift and optimizer-state effects.

For the teacher-side context conditions of Table 9, Answer only, Unrelated solution, short and long Full solution, and CoT+Solution share the same student rollouts (sampled under the long-solution pool. See subsubsection D.1.5). For the position-window analysis of subsubsection D.1.6, the protocol is identical except that completions are capped at 6,144 tokens and metrics are aggregated within fixed response-position intervals.

D.1.2 ADDITIONAL MATCHED-MODE TOKEN-LEVEL SIGNALS

Table 25 complements Table 7 with eight additional checkpoints in their native matched modes.

Table 25: Additional matched-mode signals (%). Ordered by Base Pass@8. Definitions follow Table 7.

Model	S	T	Fwd. KL	$D > .05$	Agree.	Enc.	Disc.
Qwen3-0.6B	Th	Th	0.074	21.7	92.6	36.8	62.3
DeepSeek-R1-Distill-Qwen-1.5B	Th	Th	0.033	12.5	95.1	42.7	56.9
OLMo-3-7B-Instruct	NTh	NTh	0.126	24.1	92.3	38.9	51.4
Qwen3-4B-Instruct	NTh	NTh	0.175	18.9	93.5	21.8	36.8
MiMo-7B-RL	Th	Th	0.104	23.3	93.4	20.0	75.8
OLMo-3-7B-Think	Th	Th	0.115	23.7	92.6	37.1	60.9
Qwen3-4B-Thinking-2507	Th	Th	0.162	24.6	92.3	23.1	61.1
Falcon-H1R-7B	Th	Th	0.058	17.4	94.1	30.4	68.6

Consistent with subsection 6.2, matched-mode signal statistics vary across checkpoints and do not order downstream OPSD outcomes.

D.1.3 WHERE THE DIVERGENCE MASS FALLS IN THE VOCABULARY

The vocabulary analysis supports the mode-transfer conclusion in subsection 5.1: large cumulative divergence contributions often come from punctuation, delimiters, the frequent word `the`, LaTeX markers, and reasoning-transition tokens such as "Let," "So," "We," and "I."

D.1.4 TEACHER-SIDE CONTEXT SIGNAL BY CHECKPOINT

Tables 26 and 27 report seven context conditions for ten checkpoints; Figure 8 shows their per-model PCA views on a shared basis. The checkpoint-level results support the context finding in subsection 5.3: context type shapes the signal. Within OpenThoughts, CoT+Solution produces the largest signal for every checkpoint, while Full solution and Unrelated solution are closest for DeepSeek-R1-Distill-Qwen-1.5B.

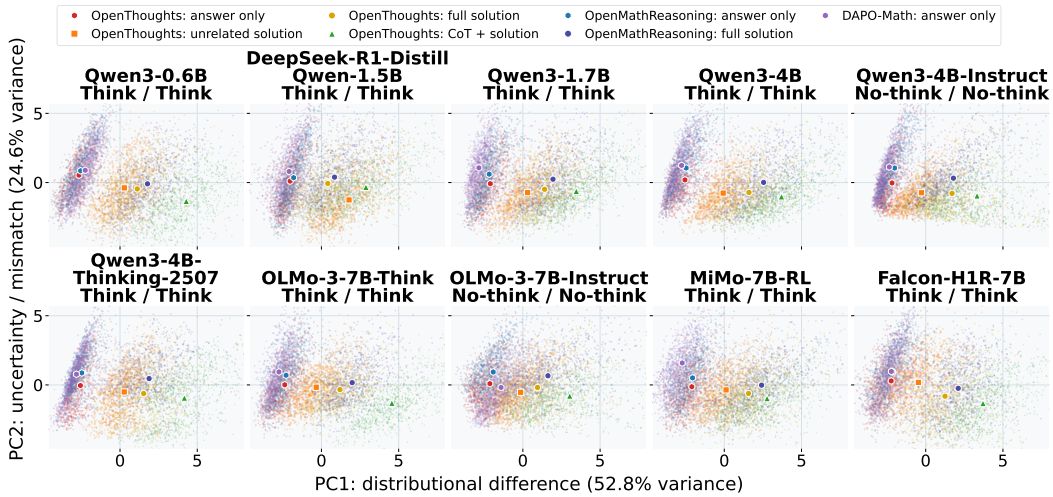


Figure 8: Per-model teacher-context PCA for the ten checkpoints underlying the aggregate view in Figure 5. Points are rollout-level signal summaries and large markers are centroids for the seven dataset–context conditions. The panels use a common PCA basis and shared axes. Features are standardized within each model–mode setting before the joint PCA fit.

Table 26: Teacher-side context signal by checkpoint (models 1–5). Unqualified conditions use OpenThoughts; OMR and DAPO denote OpenMathReasoning and DAPO-Math-17k. CoT+Sol. denotes CoT+Solution. $D > .05$ and Agree. are percentages, and Abs. adv. is the mean absolute sampled-token log advantage.

Model	Mode	Context	Fwd. KL	$D > .05$	Abs. adv.	Agree.
Qwen3-0.6B	Th/Th	Answer	0.010	1.8	0.036	97.9
		Unrelated	0.046	17.5	0.112	93.8
		Full	0.074	21.5	0.139	92.7
		CoT+Sol.	0.134	36.9	0.214	89.1
		OMR Answer	0.008	1.3	0.034	98.0
		OMR Full	0.089	24.2	0.157	91.7
		DAPO Answer	0.004	0.8	0.032	98.1
DeepSeek-R1-Distill-Qwen-1.5B	Th/Th	Answer	0.004	0.9	0.024	98.6
		Unrelated	0.037	11.5	0.083	95.4
		Full	0.032	12.3	0.086	95.3
		CoT+Sol.	0.063	25.3	0.142	92.5
		OMR Answer	0.004	0.7	0.023	98.6
		OMR Full	0.038	14.5	0.098	94.5
		DAPO Answer	0.002	0.2	0.020	98.8
Qwen3-1.7B	Th/Th	Answer	0.019	3.4	0.042	98.0
		Unrelated	0.059	15.4	0.107	95.0
		Full	0.144	20.2	0.184	92.9
		CoT+Sol.	0.225	27.7	0.280	90.1
		OMR Answer	0.016	2.6	0.037	98.1
		OMR Full	0.177	22.7	0.220	91.9
		DAPO Answer	0.007	1.9	0.032	98.4
Qwen3-4B	Th/Th	Answer	0.025	3.7	0.047	97.8
		Unrelated	0.052	15.1	0.097	95.2
		Full	0.135	22.0	0.176	92.8
		CoT+Sol.	0.203	30.3	0.242	90.3
		OMR Answer	0.023	3.0	0.045	97.9
		OMR Full	0.167	25.3	0.208	91.5
		DAPO Answer	0.010	1.8	0.035	98.2
Qwen3-4B-Instruct	NTh/NTh	Answer	0.018	3.0	0.035	98.4
		Unrelated	0.050	11.1	0.080	96.4
		Full	0.155	18.1	0.175	93.8
		CoT+Sol.	0.248	23.1	0.257	91.7
		OMR Answer	0.015	2.9	0.036	98.2
		OMR Full	0.140	21.2	0.183	92.9
		DAPO Answer	0.006	1.4	0.025	98.7

Table 27: Teacher-side context signal by checkpoint (models 6–10 and ten-model macro average). Condition and metric definitions follow Table 26.

Model	Mode	Context	Fwd. KL	$D > .05$	Abs. adv.	Agree.
Qwen3-4B-Thinking-2507	Th/Th	Answer	0.018	3.2	0.039	98.1
		Unrelated	0.105	19.2	0.128	94.1
		Full	0.159	24.1	0.181	92.4
		CoT+Sol.	0.275	34.8	0.270	89.0
		OMR Answer	0.018	2.8	0.038	98.1
		OMR Full	0.176	25.3	0.208	91.5
		DAPO Answer	0.007	1.3	0.026	98.6
OLMo-3-7B-Think	Th/Th	Answer	0.023	2.5	0.038	98.1
		Unrelated	0.055	17.5	0.107	94.6
		Full	0.111	23.1	0.153	92.8
		CoT+Sol.	0.206	38.8	0.234	89.6
		OMR Answer	0.023	1.8	0.035	98.2
		OMR Full	0.132	25.4	0.171	91.9
		DAPO Answer	0.004	0.7	0.023	98.8
OLMo-3-7B-Instruct	NTh/NTh	Answer	0.015	3.8	0.043	97.7
		Unrelated	0.048	15.5	0.099	95.0
		Full	0.112	22.7	0.162	92.8
		CoT+Sol.	0.195	30.9	0.254	89.8
		OMR Answer	0.012	3.3	0.044	97.6
		OMR Full	0.126	27.0	0.191	91.3
		DAPO Answer	0.011	3.3	0.042	97.7
MiMo-7B-RL	Th/Th	Answer	0.019	3.6	0.046	97.9
		Unrelated	0.056	16.9	0.105	95.0
		Full	0.103	23.0	0.147	93.5
		CoT+Sol.	0.154	32.5	0.190	92.0
		OMR Answer	0.018	2.9	0.044	97.9
		OMR Full	0.122	26.0	0.171	92.4
		DAPO Answer	0.009	1.6	0.036	98.2
Falcon-H1R-7B	Th/Th	Answer	0.012	2.7	0.038	97.9
		Unrelated	0.034	9.8	0.082	95.6
		Full	0.051	15.5	0.104	94.5
		CoT+Sol.	0.086	27.4	0.158	92.0
		OMR Answer	0.010	1.9	0.034	98.1
		OMR Full	0.059	18.0	0.120	93.6
		DAPO Answer	0.007	1.3	0.029	98.4
Macro average	–	Answer	0.016	2.9	0.039	98.0
		Unrelated	0.054	14.9	0.100	95.0
		Full	0.108	20.3	0.151	93.3
		CoT+Sol.	0.179	30.8	0.224	90.6
		OMR Answer	0.015	2.3	0.037	98.1
		OMR Full	0.122	23.0	0.173	92.3
		DAPO Answer	0.007	1.4	0.030	98.4

D.1.5 Long Full-Solution Teacher Prefixes

The main teacher-context analysis uses the 1,024-token prompt cap from training. The length control rescores the same student rollouts with a 12,288-token full-solution teacher prefix. Answer only and Unrelated solution retain the short budget, and token-level CoT+Solution uses a 10,240-token cap. All five prefixes share identical student completions sampled from the long-solution pool, with the short condition truncating C in $Q \oplus C$. This offline control differs from the training intervention in subsection 5.3, which adds a separate CoT field to matched OpenThoughts rows and uses an 8,192-token teacher cap.

Consistent with subsection 5.3, prefix length alone changes the initial signal only slightly; the Answer only $<$ Unrelated solution $<$ Full solution ordering remains unchanged (Table 28).

Table 28: Length-control extension of the full-solution signal in Table 9: short vs. long full-solution teacher prefixes on shared student rollouts. SOL uses a 1,024-token teacher prompt cap. SOL-LONG raises it to 12,288. $D > .05$ and Agree. are percentages. Abs. adv. is the mean absolute sampled-token log advantage.

Model	Mode	Prefix	Fwd KL	$D > .05$	Abs. adv.	Agree.
Qwen3-1.7B	Th/Th	SOL	0.144	20.2	0.184	92.9
		SOL-LONG	0.147	20.4	0.188	92.9
Qwen3-1.7B	NTh/NTh	SOL	0.090	16.7	0.125	94.7
		SOL-LONG	0.094	17.2	0.129	94.6
Qwen3-4B	Th/Th	SOL	0.135	22.0	0.176	92.8
		SOL-LONG	0.137	22.3	0.178	92.7
Qwen3-4B	NTh/NTh	SOL	0.066	15.2	0.109	95.2
		SOL-LONG	0.068	15.6	0.111	95.1
Qwen3-4B-Instruct	NTh/NTh	SOL	0.155	18.1	0.175	93.8
		SOL-LONG	0.158	18.3	0.178	93.7
OLMo-3-7B-Think	Th/Th	SOL	0.110	23.0	0.153	92.8
		SOL-LONG	0.114	23.6	0.156	92.7
OLMo-3-7B-Instruct	NTh/NTh	SOL	0.112	22.6	0.162	92.8
		SOL-LONG	0.115	23.1	0.165	92.7
Macro average	–	SOL	0.116	19.7	0.155	93.6
		SOL-LONG	0.119	20.1	0.158	93.5

D.1.6 POSITION-WINDOW SIGNAL BY CHECKPOINT

The position-window results support early-token supervision in subsection 6.3: forward KL decreases with response position in every checkpoint (Table 29).

Table 29: Forward KL by response-position window (rollout cap 6,144). Each cell reports mean/SNR, where $\mathrm{SNR} = \mu/\sigma$ and σ is the token-level standard deviation. The macro row assigns equal weight to each checkpoint, matching Table 11.

Model	Mode	$[0, 128)$	$[128, 256)$	$[256, 512)$	$[512, 1\mathrm{k})$	$[1\mathrm{k}, 2\mathrm{k})$	$[2\mathrm{k}, 4\mathrm{k})$	$[4\mathrm{k}, 6\mathrm{k})$
Qwen3-1.7B	Th/Th	0.291/0.238	0.224/0.214	0.138/0.186	0.091/0.153	0.066/0.124	0.048/0.104	0.041/0.090
Qwen3-4B	Th/Th	0.249/0.233	0.197/0.230	0.132/0.203	0.094/0.173	0.073/0.143	0.059/0.124	0.048/0.113
Qwen3-4B-Instruct	NTh/NTh	0.339/0.224	0.230/0.202	0.157/0.166	0.100/0.143	0.064/0.125	0.046/0.114	0.046/0.104
OLMo-3-7B-Think	Th/Th	0.258/0.278	0.158/0.255	0.106/0.221	0.077/0.173	0.062/0.132	0.050/0.106	0.039/0.088
OLMo-3-7B-Instruct	NTh/NTh	0.299/0.312	0.139/0.254	0.095/0.218	0.071/0.196	0.071/0.182	0.077/0.159	0.068/0.120
Macro average	–	0.287/0.248	0.190/0.217	0.126/0.185	0.087/0.160	0.067/0.138	0.056/0.121	0.048/0.103

D.1.7 TRAINING-TIME FORWARD KL VERSUS DOWNSTREAM CHANGE

Table 10 pairs training-time forward KL with Avg@8 and Pass@8 changes under full-solution OPSD for nine checkpoints. For each run, we average the logged unclipped, vocabulary-summed forward KL over optimizer steps 1–10 and 91–100. Table 30 adds Qwen3-4B using its complete 256- and 1,024-token logs. Absolute Base and OPSD results for every model, rollout cap, and benchmark appear in Table 58.

Table 30: Qwen3-4B row supplementing the training-time forward-KL analysis in Table 10. KL First10 and Final10 average the first and final ten optimizer-step records. Δ Avg and Δ Pass are changes from Base in percentage points.

	256-token rollout cap				1,024-token rollout cap			
Model	KL First10	KL Final10	Δ Avg	Δ Pass	KL First10	KL Final10	Δ Avg	Δ Pass
Qwen3-4B	0.226	$0.412\uparrow$	-1.4	-3.4	0.137	$0.329\uparrow$	-4.4	-3.4

Consistent with subsection 6.2, neither early nor late forward KL tracks downstream changes across rollout caps and metrics for Qwen3-4B (Table 30).

D.1.8 INITIALIZATION-DEPENDENT TRAINING DYNAMICS

Table 31: Mean diagnostics over early (steps 1–10) and late (steps 91–100) windows in the initialization block of Figure 6. Both runs use matched Th/Th and $\tau_{\text{clip}} = 0.05$. H_S is student entropy and $\|\nabla\|$ is the pre-clipping gradient norm.

Initialization	Steps	H_S	$D_{\text{KL}}(\pi^T \| \pi^S)$	$\|\nabla\|$	Loss
Qwen3-4B (Th/Th)	1–10	0.24	0.137	0.39	-0.009
	91–100	0.22	0.329	0.18	-0.020
SFT+GRPO then OPSD	1–10	0.41	0.047	0.35	-0.001
	91–100	1.29	0.215	0.86	-0.029

Consistent with subsection 6.4, only the collapsing SFT+GRPO initialization shows a joint rise in student entropy and gradient norm.

E Additional Experiments and Analyses

E.1 Cross-Dataset Teacher-Context Replication

We repeat the solution-only versus CoT-plus-solution comparison on two disjoint OpenMathReasoning subsets using Qwen3-1.7B, with identical examples and row order within each pair. More detailed teacher context does not necessarily improve training, as in subsection 5.3: adding CoT lowers Avg@8 from 34.4 to 27.8 on the at-most-1k subset and from 34.7 to 29.0 on the 2–4k subset despite the larger teacher prompt budget (Table 54).

E.2 Instruction-Only Teacher Prefixes

We remove privileged teacher content and vary only the English student and teacher instructions. Concise-to-detailed (C→D) asks the student for a concise solution and the teacher for a detailed one. D→C reverses the instructions.

Table 32: Instruction-only prefix controls after 100 OPSD steps. Each cell reports Avg@8 / Pass@8 / mean output length. Per-dataset results are in Table 77.

Model	Base	Full solution	No priv.: C→D	No priv.: D→C
Qwen3-1.7B	34.0 / 59.2 / 18.6k	35.3 / 60.8 / 22.4k	23.1 / 46.7 / 12.6k	36.3 / 62.5 / 25.0k
OLMo-3-7B-Think	63.5 / 81.7 / 18.7k	58.8 / 77.5 / 21.7k	60.7 / 80.8 / 16.5k	56.4 / 75.0 / 24.1k

Teacher context shapes OPSD beyond privileged semantics, extending subsection 5.3: instruction-only prefixes produce direction-dependent changes without privileged content.

E.3 Additional Training Controls

E.3.1 Pointwise Contribution-Clip Threshold Sweep

We vary only the pointwise vocabulary-contribution cap, retraining Qwen3-1.7B and OLMo-3-7B-Think with $\tau_{\text{clip}} \in \{0.01, 0.1, 0.2\}$ while keeping all other settings fixed, and compare against Base and the default $\tau_{\text{clip}} = 0.05$ run (Table 33).

Table 33: Vocabulary-contribution clip threshold sweep after 100 OPSD steps. Bold marks the best Avg@8 / Pass@8 within each model (ties both bolded). Per-dataset results are in Table 79.

Model	τ_{clip}	Avg@8	Pass@8	Length
Qwen3-1.7B	Base	34.0	59.2	18.6k
	0.01	31.4	56.7	26.2k
	0.05	**35.3**	**60.8**	22.4k
	0.1	32.9	60.0	16.6k
	0.2	30.6	53.3	15.5k
OLMo-3-7B-Think	Base	**63.5**	**81.7**	18.7k
	0.01	58.2	79.2	20.6k
	0.05	58.8	77.5	21.7k
	0.1	63.1	80.8	19.4k
	0.2	63.1	81.7	17.0k

Consistent with subsection 5.4, changing the clipping threshold provides no reliable remedy. The leading threshold varies by model and does not consistently improve over the default or Base (Table 33).

E.3.2 Short SFT then OPSD

We insert 10 supervised steps before 100-step same-prefix OPSD without privileged teacher context. The primary runs use OpenThoughts for both stages. An additional Qwen3-1.7B control uses OpenMathReasoning for SFT and OpenThoughts for OPSD.

Table 34: Ten-step SFT followed by 100-step no-privilege OPSD. OT/OMR denote the SFT corpus. OPSD uses OpenThoughts throughout. Bold marks the best Avg@8 / Pass@8 within each model–corpus block. Per-dataset results are in Table 80.

Model	SFT data	Stage	Avg@8	Pass@8	Length
Qwen3-1.7B	OT	Base	**34.0**	**59.2**	18.6k
		SFT-10	5.1	17.5	6.9k
		+ OPSD-100	5.2	16.7	6.6k
Qwen3-1.7B	OMR	Base	**34.0**	**59.2**	18.6k
		SFT-10	28.0	54.2	16.9k
		+ OPSD-100	27.1	49.2	16.9k
Qwen3-4B-Thinking	OT	Base	**75.4**	**86.7**	22.8k
		SFT-10	4.9	20.8	2.0k
		+ OPSD-100	7.8	28.3	3.1k
OLMo-3-7B-Think	OT	Base	**63.5**	**81.7**	18.7k
		SFT-10	60.1	80.0	18.4k
		+ OPSD-100	60.3	**81.7**	18.3k

OPSD remains sensitive to the post-SFT state, as in subsection 5.6: it does not reliably recover degraded checkpoints and changes stable checkpoints little.

E.4 LoRA Diagnostics and Qwen3.5 Collapse

We compare matched-mode full-parameter and LoRA OPSD across six model settings, then analyze Qwen3.5-4B collapse under matched modes (Table 35, Table 36).

Table 35: Matched-mode full-parameter and LoRA OPSD after 100 updates. Each cell reports Avg@8 / Pass@8 / mean output length. Bold marks the best Avg@8 and Pass@8 within each model. Ties are all bolded.

Model	Base	Full OPSD	LoRA OPSD
Qwen3-1.7B	34.0 / 59.2 / 18.6k	35.3 / 60.8 / 22.4k	**35.8** / **61.7** / 23.3k
Qwen3-4B	**62.3** / 79.2 / 17.8k	57.9 / 75.8 / 19.9k	59.9 / **80.0** / 20.2k
Qwen3-4B-Instruct	48.5 / **71.7** / 9.1k	**50.0** / 70.0 / 10.6k	49.2 / 70.8 / 11.2k
Qwen3-4B-Thinking	**75.4** / **86.7** / 22.8k	65.8 / 83.3 / 43.4k	68.0 / 82.5 / 40.5k
OLMo-3-7B-Think	**63.5** / **81.7** / 18.7k	58.8 / 77.5 / 21.7k	62.5 / 77.5 / 21.3k
OLMo-3-7B-Instruct	39.5 / 60.8 / 7.6k	40.8 / **62.5** / 7.4k	**41.3** / **62.5** / 7.9k

OPSD gains remain model-dependent, as in subsection 5.5. LoRA reduces drift and keeps most checkpoints closer to Base, but does not yield consistent gains.

Table 36: Qwen3.5-4B full-parameter and LoRA results after 100 matched-mode OPSD updates.

Eval mode	Evaluated state	Avg@8	Pass@8	Format	Length
NTh	Base	52.71	77.50	97.71	10.2k
	Full OPSD (NTh/NTh)	8.02	35.00	14.90	70.7k
	LoRA OPSD (NTh/NTh)	45.73	68.33	99.48	8.4k
Th	Base	80.62	91.67	90.52	38.1k
	Full OPSD (Th/Th)	0.00	0.00	5.42	81.9k
	LoRA OPSD (Th/Th)	8.54	20.83	10.62	77.4k

E.4.1 The NTh Mode Is Not Strictly Isolated

Reasoning mode remains a confound rather than a clean semantic control, as in subsection 5.1. Specifically, Qwen3.5-4B sometimes emits an additional `</think>` delimiter after the chat template closes the empty thinking block and produces longer outputs than Qwen3-4B. The delimiter persists after full-parameter NTh/NTh training, so the two modes are not operationally isolated.

E.4.2 The Distinguishing Signal Is Dynamic, Not Initial

Consistent with subsection 6.2, initial KL does not predict training stability. Matched modes start with lower forward KL than their corresponding mismatches, yet both later collapse; the subsequent entropy and forward-KL rise separates Qwen3.5-4B from the bounded Qwen3 controls (Table 37, Figure 9).

E.4.3 LoRA Attenuates the Drift

Parameter restriction attenuates but does not remove the collapse behavior. The LoRA controls use rank 64 with scale 128 on `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`; all other parameters are frozen. LoRA removes length growth but leaves an accuracy drop for NTh/NTh, while Th/Th still collapses (Figure 9).

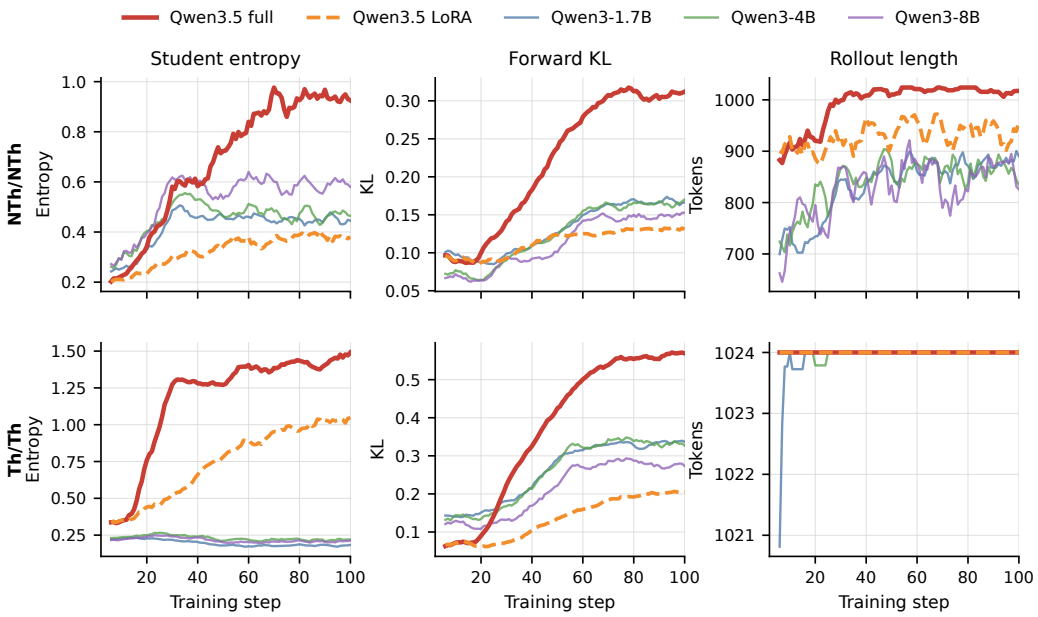


Figure 9: Matched-mode training dynamics for full-parameter and LoRA Qwen3.5-4B versus three Qwen3 full-parameter controls. Curves are five-step moving averages from step 2 onward. Step 1 contains a one-off startup batch and is omitted.

Table 37: Initial Qwen3.5-4B token-level signals. Definitions follow Table 7. Per-setting and per-dataset results for the downstream mode grid are in Table 49.

S	T	Fwd. KL	$D > .05$	Agree.	Enc.	Disc.
NTh	NTh	0.087	17.7	94.7	34.1	63.4
NTh	Th	0.208	31.2	92.7	14.9	84.6
Th	NTh	0.140	27.2	92.0	56.0	43.4
Th	Th	0.066	16.5	95.0	40.2	58.7

E.5 Replications of Recent OPSD Variants

We reproduce two recent OPSD variants: Purified OPSD replaces the raw privileged-teacher target with a PMI-corrected target, while β-OPSD uses an interpolated teacher–reference target with return-to-go credit assignment. We test whether vanilla OPSD's model and configuration sensitivity persists under these targets.

Purified OPSD uses the authors' main PMI setting: LoRA rank 64 (scaling 128), learning rate 5×10^{-6}, 200 updates, global batch 32, 1,024-token prompt and rollout caps, symmetric JSD, PMI strength 1, and PMI soft-clip threshold 10. We evaluate answer-only and full-solution fields. β-OPSD uses the same rank, learning rate, step count, batch size, and caps; its teacher weight increases from 0.5 to 0.8, its reference is the stop-gradient current student, and its return-to-go discount is 0.99. All rows use the four-benchmark, eight-sample protocol.

Table 38: Paper-configuration replication results, macro-averaged over AIME 2024–2026 and HMMT25. Each $\Delta_{\mathrm{B/O}}$ entry gives the change relative to Base / standard OPSD-1024 in the same native matched reasoning mode. Avg@8 and Pass@8 changes are absolute percentage points; Length changes are relative percentages. Per-dataset results are in Table 75 and Table 76.

Variant	Model	Privileged field	Avg@8	$\Delta_{\mathrm{B/O}}$	Pass@8	$\Delta_{\mathrm{B/O}}$	Length	$\Delta_{\mathrm{B/O}}$
Purified OPSD	Qwen3-1.7B	Answer	39.9	$+5.9/+4.6$	63.3	$+4.1/+2.5$	24.2k	$+29.7/+7.9$
	Qwen3-1.7B	Solution	37.5	$+3.5/+2.2$	65.0	$+5.8/+4.2$	20.8k	$+11.5/-7.2$
	Qwen3-4B	Answer	63.3	$+1.0/+5.4$	79.2	$+0.0/+3.3$	19.6k	$+10.1/-2.0$
	Qwen3-4B	Solution	61.3	$-1.0/+3.4$	77.5	$-1.7/+1.7$	18.5k	$+4.3/-7.1$
	Qwen3-4B-Thinking	Answer	75.1	$-0.3/+9.3$	84.2	$-2.5/+0.8$	24.8k	$+8.8/-42.9$
	Qwen3-4B-Thinking	Solution	75.3	$-0.1/+9.5$	86.7	$+0.0/+3.3$	24.2k	$+6.2/-44.2$
	OLMo-3-7B-Think	Answer	64.5	$+1.0/+5.7$	80.8	$-0.9/+3.3$	19.2k	$+2.4/-11.7$
	OLMo-3-7B-Think	Solution	63.8	$+0.3/+5.0$	80.0	$-1.7/+2.5$	18.9k	$+0.6/-13.2$
β-OPSD	Qwen3-1.7B	Solution	25.1	$-8.9/-10.2$	49.2	$-10.0/-11.7$	13.4k	$-28.0/-40.0$
	Qwen3-4B-Thinking	Solution	68.6	$-6.8/+2.8$	82.5	$-4.2/-0.8$	18.2k	$-20.3/-58.1$
	Qwen3-4B-Instruct	Solution	47.2	$-1.3/-2.8$	65.8	$-5.9/-4.2$	11.0k	$+20.3/+4.5$
	OLMo-3-7B-Think	Solution	64.1	$+0.6/+5.3$	81.7	$+0.0/+4.2$	18.8k	$+0.2/-13.5$
	OLMo-3-7B-Instruct	Solution	42.4	$+2.9/+1.6$	65.0	$+4.2/+2.5$	8.3k	$+9.1/+11.9$

The replications retain the model dependence reported in subsection 5.5: each variant improves standard OPSD-1024 in some settings, but neither improves over Base consistently (Table 38).

F Robustness Checks

F.1 Seed Sensitivity

For matched Qwen3-1.7B and OLMo-3-7B-Think, Table 39 lists the configurations repeated over training or evaluation seeds 42, 1024, and 65536. Training-seed tables contain independent training runs evaluated with seed 42. Evaluation-seed tables reuse the seed-42-trained checkpoints and vary the decoding seed.

Table 39: Seed repetitions for Qwen3-1.7B and OLMo-3-7B-Think. "Train" denotes independent training runs. "Eval" denotes decoding seeds applied to the seed-42-trained checkpoint.

Experiment family	Conditions	Training seeds	Evaluation seeds
Base checkpoint	Untrained Base	–	42 / 1024 / 65536
Rollout length	256 / 1024	42 / 1024 / 65536	42 / 1024 / 65536
Loss-token position	First / Uniform / Final-256	42 / 1024 / 65536	42 / 1024 / 65536
Teacher context	Solution / Answer / Unrelated	42 / 1024 / 65536	42 / 1024 / 65536

Table 40 reports the unannotated Base references under evaluation seeds 42/1024/65536. Tables 41–46 mark Avg@8 and Pass@8 with $^{\uparrow}/^{\downarrow}$ relative to Base: training-seed tables use Base@eval-42, and evaluation-seed tables use the matched evaluation seed. Ties receive no symbol. Length is unannotated.

Table 40: Base checkpoints under evaluation seeds 42/1024/65536. Training-seed tables use Base@eval-42. Evaluation-seed tables use Base at the matched evaluation seed. Per-dataset results are in Table 65.

Model	Evaluation seed								
	42			1024			65536		
	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length
Qwen3-1.7B	33.96	59.17	18.6k	35.10	55.00	18.6k	35.31	61.67	18.6k
OLMo-3-7B-Think	63.54	81.67	18.7k	62.19	80.83	19.0k	63.12	79.17	18.9k

F.1.1 Maximum Student Rollout Length

The seed repetitions support the rollout-length finding in subsection 5.2: OPSD-256 has higher Avg@8 than OPSD-1024 across all training and evaluation seeds (Table 41, Table 42).

F.1.2 Loss-Token Position Selection

The seed repetitions also support the position finding in subsection 6.3: with a fixed 256-position budget, First-256 has the highest mean Avg@8 on Qwen3-1.7B, while later-position variants remain below Base on OLMo-3-7B-Think (Table 43, Table 44).

F.1.3 Teacher-Side Context (Prefix) Interventions

The seed repetitions support the context finding in subsection 5.3: Answer only does not consistently lead, and Unrelated solution remains comparable to Full solution (Table 45, Table 46).

Table 41: Rollout-length results across training seeds 42/1024/65536, evaluated with seed 42. OPSD-256/1024 use full-solution context and matched Th/Th modes (subsection 5.2). Per-dataset results are in Table 66.

Model	Condition	Training seed								
		42			1024			65536		
		Avg@8	Pass@8	Length	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length
Qwen3-1.7B	OPSD-256	38.12$^{\uparrow}$	62.50$^{\uparrow}$	21.4k	39.17$^{\uparrow}$	65.00$^{\uparrow}$	21.1k	37.08$^{\uparrow}$	65.00$^{\uparrow}$	21.3k
	OPSD-1024	35.31$^{\uparrow}$	60.83$^{\uparrow}$	22.4k	35.94$^{\uparrow}$	56.67$^{\downarrow}$	22.0k	34.58$^{\uparrow}$	62.50$^{\uparrow}$	22.5k
OLMo-3-7B-Think	OPSD-256	60.83$^{\downarrow}$	80.00$^{\downarrow}$	21.4k	61.46$^{\downarrow}$	79.17$^{\downarrow}$	21.5k	61.98$^{\downarrow}$	81.67	21.5k
	OPSD-1024	58.75$^{\downarrow}$	77.50$^{\downarrow}$	21.7k	59.17$^{\downarrow}$	78.33$^{\downarrow}$	21.8k	59.69$^{\downarrow}$	81.67	21.4k

Table 42: Rollout-length results across evaluation seeds 42/1024/65536 using the seed-42-trained checkpoints. Per-dataset results are in Table 67.

Model	Condition	Evaluation seed								
		42			1024			65536		
		Avg@8	Pass@8	Length	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length
Qwen3-1.7B	OPSD-256	38.12$^{\uparrow}$	62.50$^{\uparrow}$	21.4k	36.98$^{\uparrow}$	59.17$^{\uparrow}$	21.0k	36.67$^{\uparrow}$	62.50$^{\uparrow}$	21.3k
	OPSD-1024	35.31$^{\uparrow}$	60.83$^{\uparrow}$	22.4k	36.77$^{\uparrow}$	65.00$^{\uparrow}$	21.9k	36.56$^{\uparrow}$	61.67	22.2k
OLMo-3-7B-Think	OPSD-256	60.83$^{\downarrow}$	80.00$^{\downarrow}$	21.4k	61.98$^{\downarrow}$	80.83	21.7k	61.35$^{\downarrow}$	77.50$^{\downarrow}$	21.6k
	OPSD-1024	58.75$^{\downarrow}$	77.50$^{\downarrow}$	21.7k	58.85$^{\downarrow}$	78.33$^{\downarrow}$	21.5k	60.31$^{\downarrow}$	79.17	21.6k

Table 43: Loss-token position results across training seeds. First-256 reuses the corresponding OPSD-256 run; Uni-256 and Final-256 use 1,024-token rollouts and retain at most 256 uniformly sampled or final loss positions. Per-dataset results are in Table 68.

Model	Condition	Training seed								
		42			1024			65536		
		Avg@8	Pass@8	Length	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length
Qwen3-1.7B	First-256 / OPSD-256	38.12$^{\uparrow}$	62.50$^{\uparrow}$	21.4k	39.17$^{\uparrow}$	65.00$^{\uparrow}$	21.1k	37.08$^{\uparrow}$	65.00$^{\uparrow}$	21.3k
	Uni-256	35.21$^{\uparrow}$	61.67$^{\uparrow}$	22.6k	34.90$^{\uparrow}$	59.17	22.4k	36.56$^{\uparrow}$	62.50$^{\uparrow}$	22.8k
	Final-256	34.27$^{\uparrow}$	59.17	23.7k	33.23$^{\downarrow}$	55.00$^{\downarrow}$	23.8k	34.17$^{\uparrow}$	58.33$^{\downarrow}$	23.7k
OLMo-3-7B-Think	First-256 / OPSD-256	60.83$^{\downarrow}$	80.00$^{\downarrow}$	21.4k	61.46$^{\downarrow}$	79.17$^{\downarrow}$	21.5k	61.98$^{\downarrow}$	81.67	21.5k
	Uni-256	59.06$^{\downarrow}$	82.50$^{\uparrow}$	21.5k	58.54$^{\downarrow}$	75.00$^{\downarrow}$	21.6k	60.00$^{\downarrow}$	80.83$^{\downarrow}$	21.1k
	Final-256	61.25$^{\downarrow}$	79.17$^{\downarrow}$	21.1k	57.39$^{\downarrow}$	78.33$^{\downarrow}$	21.7k	57.29$^{\downarrow}$	77.50$^{\downarrow}$	21.8k

Table 44: Loss-token position results across evaluation seeds using the seed-42-trained checkpoints. First-256 reuses OPSD-256. Per-dataset results are in Table 69.

Model	Condition	Evaluation seed								
		42			1024			65536		
		Avg@8	Pass@8	Length	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length
Qwen3-1.7B	First-256 / OPSD-256	38.12$^{\uparrow}$	62.50$^{\uparrow}$	21.4k	36.98$^{\uparrow}$	59.17$^{\uparrow}$	21.0k	36.67$^{\uparrow}$	62.50$^{\uparrow}$	21.3k
	Uni-256	35.21$^{\uparrow}$	61.67$^{\uparrow}$	22.6k	35.52$^{\uparrow}$	59.17$^{\uparrow}$	22.3k	34.27$^{\downarrow}$	58.33$^{\downarrow}$	22.8k
	Final-256	34.27$^{\uparrow}$	59.17	23.7k	35.00$^{\downarrow}$	61.67$^{\uparrow}$	23.5k	33.85$^{\downarrow}$	58.33$^{\downarrow}$	23.3k
OLMo-3-7B-Think	First-256 / OPSD-256	60.83$^{\downarrow}$	80.00$^{\downarrow}$	21.4k	61.98$^{\downarrow}$	80.83	21.7k	61.35$^{\downarrow}$	77.50$^{\downarrow}$	21.6k
	Uni-256	59.06$^{\downarrow}$	82.50$^{\uparrow}$	21.5k	59.69$^{\downarrow}$	78.33$^{\downarrow}$	21.2k	58.54$^{\downarrow}$	79.17	21.5k
	Final-256	61.25$^{\downarrow}$	79.17$^{\downarrow}$	21.1k	59.27$^{\downarrow}$	79.17$^{\downarrow}$	21.1k	58.23$^{\downarrow}$	80.00$^{\uparrow}$	21.3k

Table 45: Teacher-context results across training seeds at a 1,024-token rollout cap (Full solution / Answer only / Unrelated solution, subsection 5.3). Per-dataset results are in Table 70.

Model	Condition	Training seed								
		42			1024			65536		
		Avg@8	Pass@8	Length	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length
Qwen3-1.7B	Solution	35.31$^{\uparrow}$	60.83$^{\uparrow}$	22.4k	35.94$^{\uparrow}$	56.67$^{\downarrow}$	22.0k	34.58$^{\uparrow}$	62.50$^{\uparrow}$	22.5k
	Answer	35.62$^{\uparrow}$	59.17	25.8k	36.67$^{\uparrow}$	64.17$^{\uparrow}$	25.6k	34.69$^{\uparrow}$	59.17	25.5k
	Unrelated	37.08$^{\uparrow}$	60.00$^{\uparrow}$	22.0k	38.96$^{\uparrow}$	62.50$^{\uparrow}$	21.9k	37.29$^{\uparrow}$	62.50$^{\uparrow}$	21.6k
OLMo-3-7B-Think	Solution	58.75$^{\downarrow}$	77.50$^{\downarrow}$	21.7k	59.17$^{\downarrow}$	78.33$^{\downarrow}$	21.8k	59.69$^{\downarrow}$	81.67	21.4k
	Answer	56.15$^{\downarrow}$	75.83$^{\downarrow}$	24.1k	56.88$^{\downarrow}$	74.17$^{\downarrow}$	24.4k	56.98$^{\downarrow}$	74.17$^{\downarrow}$	23.9k
	Unrelated	57.50$^{\downarrow}$	76.67$^{\downarrow}$	23.4k	56.46$^{\downarrow}$	75.00$^{\downarrow}$	23.4k	56.46$^{\downarrow}$	76.67$^{\downarrow}$	23.8k

Table 46: Teacher-context results across evaluation seeds using the seed-42-trained checkpoints and a 1,024-token rollout cap. Per-dataset results are in Table 71.

Model	Condition	Evaluation seed								
		42			1024			65536		
		Avg@8	Pass@8	Length	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length
Qwen3-1.7B	Solution	35.31$^{\uparrow}$	60.83$^{\uparrow}$	22.4k	36.77$^{\uparrow}$	65.00$^{\uparrow}$	21.9k	36.56$^{\uparrow}$	61.67	22.2k
	Answer	35.62$^{\uparrow}$	59.17	25.8k	35.83$^{\uparrow}$	55.83$^{\uparrow}$	25.8k	35.94$^{\uparrow}$	60.00$^{\downarrow}$	25.9k
	Unrelated	37.08$^{\uparrow}$	60.00$^{\uparrow}$	22.0k	35.31$^{\uparrow}$	61.67$^{\uparrow}$	22.2k	36.67$^{\uparrow}$	59.17$^{\downarrow}$	22.0k
OLMo-3-7B-Think	Solution	58.75$^{\downarrow}$	77.50$^{\downarrow}$	21.7k	58.85$^{\downarrow}$	78.33$^{\downarrow}$	21.5k	60.31$^{\downarrow}$	79.17	21.6k
	Answer	56.15$^{\downarrow}$	75.83$^{\downarrow}$	24.1k	58.85$^{\downarrow}$	79.17$^{\downarrow}$	23.9k	58.12$^{\downarrow}$	80.00$^{\uparrow}$	24.1k
	Unrelated	57.50$^{\downarrow}$	76.67$^{\downarrow}$	23.4k	58.02$^{\downarrow}$	80.00$^{\downarrow}$	23.4k	58.65$^{\downarrow}$	75.00$^{\downarrow}$	23.3k

F.2 Temperature Robustness Checks

We vary the evaluation decoding temperature, student rollout-sampling temperature, and teacher scoring temperature while holding all other settings fixed.

F.2.1 Decoding Temperature

The decoding-temperature check leaves the model-dependent conclusion in subsection 5.5 unchanged: small Instruct-model gains can reverse, while Qwen3-4B-Thinking degrades at both temperatures with substantial length growth (Table 47).

Table 47: Decoding-temperature robustness check, grouped by identical decode temperature within each model. The Base and OPSD values within each row are directly comparable. Δ is OPSD minus Base. Each OPSD row uses the checkpoint trained under the default configuration. Results are unweighted averages over four benchmarks. Per-dataset results are in Table 73.

Model	Decode τ	Base			100-step OPSD			Δ (OPSD – Base)		
		Avg@8	Pass@8	Length	Avg@8	Pass@8	Length	Avg@8	Pass@8	Length
Qwen3-1.7B	0.6	33.96	59.17	18.6k	35.31	60.83	22.4k	+1.35	+1.66	+3.8k
	1.0	38.12	62.50	18.0k	36.67	65.83	21.4k	−1.45	+3.33	+3.4k
Qwen3-4B-Instruct	0.7	48.54	71.67	9.2k	50.00	70.00	10.6k	+1.46	−1.67	+1.4k
	1.0	49.17	70.83	9.2k	48.75	66.67	10.7k	−0.42	−4.16	+1.5k
Qwen3-4B-Thinking	0.6	75.42	86.67	22.8k	65.83	83.33	43.4k	−9.59	−3.34	+20.6k
	1.0	76.04	86.67	23.1k	66.77	82.50	43.7k	−9.27	−4.17	+20.6k
OLMo-3-7B-Instruct	0.6	39.48	60.83	7.6k	40.83	62.50	7.4k	+1.35	+1.67	−0.2k
	1.0	41.88	68.33	7.6k	38.02	61.67	7.2k	−3.86	−6.66	−0.4k

F.2.2 Training Sampling and Scoring Temperatures

Table 48: Training temperature check for Qwen3-1.7B OPSD. τ_S is the student rollout-sampling temperature and τ_T is the teacher scoring temperature. All other settings are fixed. Values are unweighted averages over four benchmarks. Bold marks the best Avg@8 and Pass@8. Ties are all bolded. Per-dataset results are in Table 74.

τ_S	τ_T	Eval. τ	Avg@8	Pass@8	Length
1.1	1.1	0.6	35.31	**60.83**	22.4k
0.6	0.6	0.6	**35.42**	**60.83**	20.0k
0.6	1.1	0.6	35.21	56.67	22.3k
1.1	0.6	0.6	32.29	57.50	18.3k

Training-temperature changes likewise provide no consistent remedy: lowering both temperatures preserves accuracy while shortening outputs, whereas asymmetric settings do not improve accuracy consistently (Table 48).

G PER-DATASET EVALUATION RESULTS

The tables are grouped into main-text, seed-sensitivity, and additional-experiment results. Dataset averages use unrounded benchmark-level metrics, so averages of displayed values can differ by 0.1.

G.1 MAIN-TEXT EXPERIMENT BREAKDOWNS

G.1.1 STUDENT–TEACHER REASONING-MODE ALIGNMENT

S/T denote student/teacher modes (Th/NTh), and Eval denotes the inference mode. Each model lists its Th and NTh Base rows once because the Base checkpoint does not depend on teacher mode. NO-PRIV. retains the mismatched NTh/Th modes and problem context but removes the privileged solution.

Table 49: Per-dataset reasoning-mode results for Table 1 and Table 14, including the no-privilege control. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean. Bold marks the best Dataset Avg. for Avg@8 and Pass@8 within each model. Ties are all bolded.

Model	S	T	Eval	Checkpoint	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
					Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	Th	–	Th	Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	Th	Th	Th	OPSD	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
	Th	NTh	Th	OPSD	28.8	53.3	12469	19.2	40.0	9756	17.5	36.7	11343	11.3	26.7	10867	19.2	39.2	11109
	NTh	–	NTh	Base	14.2	30.0	4958	10.4	23.3	2808	8.8	20.0	4977	5.8	10.0	2948	9.8	20.8	3923
	NTh	Th	NTh	OPSD	20.0	36.7	10349	15.0	33.3	8806	11.7	23.3	11657	9.6	20.0	8112	14.1	28.3	9731
	NTh	NTh	NTh	OPSD	10.8	23.3	10015	10.8	30.0	7387	9.6	20.0	10503	5.4	10.0	5946	9.2	20.8	8463
	NTh	Th	Th	OPSD	50.4	76.7	19576	42.1	60.0	19946	42.5	66.7	19793	28.3	50.0	20668	40.8	**63.3**	19996
	NTh	Th	Th	NO-PRIV.	53.3	70.0	19748	42.1	63.3	20385	42.9	70.0	20423	25.8	46.7	21387	**41.0**	62.5	20486
	Th	NTh	NTh	OPSD	8.8	23.3	3512	11.3	26.7	2458	6.3	13.3	4045	3.8	16.7	1999	7.5	20.0	3003
Qwen3-4B	Th	–	Th	Base	74.6	90.0	15495	65.4	80.0	18696	65.8	80.0	17249	43.3	66.7	19636	62.3	**79.2**	17769
	Th	Th	Th	OPSD	66.3	80.0	18395	62.9	80.0	19655	62.9	83.3	18119	39.6	60.0	23625	57.9	75.8	19949
	Th	NTh	Th	OPSD	57.5	83.3	14327	46.3	70.0	15505	52.5	73.3	14392	35.4	50.0	16211	47.9	69.2	15108
	NTh	–	NTh	Base	25.0	50.0	8239	21.7	36.7	5727	19.6	43.3	7309	12.1	23.3	5500	19.6	38.3	6694
	NTh	Th	NTh	OPSD	26.2	53.3	17508	25.8	53.3	15433	20.4	50.0	16489	13.3	26.7	19388	21.5	45.8	17204
	NTh	NTh	NTh	OPSD	22.9	40.0	15764	23.3	36.7	15308	14.6	30.0	15289	7.9	23.3	17266	17.2	32.5	15907
	NTh	Th	Th	OPSD	72.5	86.7	16943	68.3	80.0	20399	67.1	86.7	18510	44.2	63.3	21324	**63.0**	**79.2**	19294
	NTh	Th	Th	NO-PRIV.	72.9	83.3	17976	65.0	76.7	19795	65.4	83.3	19041	40.4	56.7	22048	60.9	75.0	19715
	Th	NTh	NTh	OPSD	21.3	40.0	7058	19.6	30.0	4629	14.6	30.0	6524	8.8	20.0	3940	16.0	30.0	5538
Qwen3-8B	Th	–	Th	Base	76.3	86.7	14860	69.6	80.0	18059	67.1	86.7	16744	43.3	60.0	20924	64.1	78.3	17647
	Th	Th	Th	OPSD	67.5	76.7	19649	62.5	83.3	21001	65.0	86.7	19844	40.4	60.0	25193	58.9	76.7	21422
	Th	NTh	Th	OPSD	55.8	83.3	12931	45.0	70.0	14801	49.2	80.0	13775	21.3	46.7	13113	42.8	70.0	13655
	NTh	–	NTh	Base	24.2	43.3	7391	19.2	40.0	4697	16.2	40.0	6901	12.5	26.7	4667	18.0	37.5	5914
	NTh	Th	NTh	OPSD	28.8	56.7	15015	27.9	43.3	14525	23.3	53.3	14755	12.9	26.7	17077	23.2	45.0	15343
	NTh	NTh	NTh	OPSD	25.0	43.3	17311	23.8	36.7	14711	18.3	33.3	17127	9.2	16.7	16985	19.1	32.5	16533
	NTh	Th	Th	OPSD	76.3	80.0	16687	67.1	80.0	19952	70.0	83.3	18159	45.8	63.3	22370	**64.8**	76.7	19292
	NTh	Th	Th	NO-PRIV.	71.3	86.7	17480	72.1	83.3	20107	70.8	90.0	18597	42.9	63.3	22612	64.3	**80.8**	19699
	Th	NTh	NTh	OPSD	25.0	43.3	8132	23.8	36.7	5237	15.8	33.3	8612	9.6	20.0	5103	18.5	33.3	6771
Qwen3.5-4B	Th	–	Th	Base	87.5	96.7	33160	80.8	90.0	38872	84.6	93.3	35106	69.6	86.7	45246	**80.6**	**91.7**	38096
	Th	Th	Th	OPSD	0.0	0.0	81920	0.0	0.0	81920	0.0	0.0	81920	0.0	0.0	81920	0.0	0.0	81920
	Th	NTh	Th	OPSD	10.8	33.3	71442	8.8	30.0	73966	6.7	20.0	75485	8.3	26.7	71964	8.6	27.5	73214
	NTh	–	NTh	Base	67.1	93.3	8634	47.5	66.7	9523	57.1	80.0	9623	39.2	70.0	12991	52.7	77.5	10193
	NTh	Th	NTh	OPSD	0.0	0.0	81920	0.0	0.0	81920	0.0	0.0	81920	0.0	0.0	81920	0.0	0.0	81920
	NTh	NTh	NTh	OPSD	9.2	43.3	71749	7.5	30.0	70294	6.7	40.0	69655	8.8	26.7	71144	8.0	35.0	70710
	NTh	Th	Th	OPSD	0.0	0.0	81920	0.0	0.0	81920	0.0	0.0	81920	0.0	0.0	81920	0.0	0.0	81920
	Th	NTh	NTh	OPSD	45.8	73.3	12692	38.8	70.0	12564	42.1	70.0	13516	18.3	43.3	15825	36.3	64.2	13649

G.1.2 MAXIMUM STUDENT ROLLOUT LENGTH

Dataset Avg. gives the solid-curve points and dashed Base levels in Figure 3. Th/Th and NTh/NTh denote matched student–teacher modes. The NTh/NTh Qwen3 length sweeps are a diagnostic exception to the native Th/Th setting used for subsequent hybrid-model conclusions: they test whether the length trend also appears when the same hybrid checkpoints are operated in NTh mode. The 1024 / Uni-256 and 1024 / Final-256 rows keep the rollout cap at 1,024 tokens and apply the loss to at most 256 positions. Uniform samples without replacement, whereas Final retains the last positions. Both retain every valid position in shorter completions. The sweep curves exclude these position controls. Table 2 averages the Qwen3-1.7B and OLMo-3-7B-Think results.

Table 50: Per-dataset rollout-length results for the matched NTh/NThQwen3 sweeps in Figure 3. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. gives the plotted mean.

Model	Training rollout	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B NTh/NTh	Base	14.2	30.0	4958	10.4	23.3	2808	8.8	20.0	4977	5.8	10.0	2948	9.8	20.8	3923
	128	12.1	26.7	5258	10.0	23.3	3116	6.7	16.7	5283	4.2	13.3	3335	8.2	20.0	4248
	256	14.2	33.3	6162	12.9	30.0	5492	9.6	20.0	7519	5.0	16.7	3751	10.4	25.0	5731
	512	13.3	36.7	9160	10.4	30.0	8869	13.3	26.7	9046	4.2	13.3	5936	10.3	26.7	8253
	1024	10.8	23.3	10015	10.8	30.0	7387	9.6	20.0	10503	5.4	10.0	5946	9.2	20.8	8463
	2048	15.4	33.3	7255	11.7	23.3	6224	11.2	23.3	8313	5.4	20.0	4573	10.9	25.0	6591
	4096	13.8	30.0	9799	12.9	36.7	6252	6.2	20.0	10046	7.1	16.7	5354	10.0	25.8	7863
	6144	12.1	26.7	11950	14.2	33.3	7612	11.7	16.7	9960	4.6	13.3	5514	10.6	22.5	8759
Qwen3-4B NTh/NTh	Base	25.0	50.0	8239	21.7	36.7	5727	19.6	43.3	7309	12.1	23.3	5500	19.6	38.3	6694
	128	20.0	40.0	11529	22.1	40.0	8779	15.0	23.3	10465	7.9	16.7	9715	16.2	30.0	10122
	256	19.6	36.7	16505	17.5	36.7	15102	14.6	33.3	16479	8.8	20.0	17498	15.1	31.7	16396
	512	20.8	36.7	16896	19.2	33.3	16487	13.3	30.0	17239	7.1	16.7	21014	15.1	29.2	17909
	1024	22.9	40.0	15764	23.3	36.7	15308	14.6	30.0	15289	7.9	23.3	17266	17.2	32.5	15907
	2048	18.8	40.0	17382	20.0	33.3	15974	17.1	36.7	16931	6.7	16.7	20779	15.6	31.7	17767
	4096	21.2	36.7	17786	20.4	36.7	16162	14.2	36.7	18063	6.7	16.7	20520	15.6	31.7	18133
	6144	18.3	33.3	15444	23.8	40.0	15505	15.4	36.7	15878	9.6	20.0	13964	16.8	32.5	15198
Qwen3-8B NTh/NTh	Base	24.2	43.3	7391	19.2	40.0	4697	16.2	40.0	6901	12.5	26.7	4667	18.0	37.5	5914
	128	21.2	36.7	13958	17.9	33.3	11013	12.1	26.7	14225	7.5	20.0	10226	14.7	29.2	12356
	256	22.1	33.3	16930	18.3	30.0	13522	14.2	30.0	17638	7.9	20.0	15279	15.6	28.3	15842
	512	25.4	36.7	15883	25.8	40.0	14532	18.8	46.7	16116	10.0	20.0	17036	20.0	35.8	15891
	1024	25.0	43.3	17311	23.8	36.7	14710	18.3	33.3	17127	9.2	16.7	16985	19.1	32.5	16533
	2048	25.4	53.3	17870	21.7	33.3	14804	18.3	36.7	16582	8.3	20.0	18155	18.4	35.8	16853
	4096	24.6	46.7	17374	21.7	36.7	14721	15.8	36.7	16801	9.6	23.3	20046	17.9	35.8	17235
	6144	25.4	50.0	16549	20.4	43.3	14972	18.3	33.3	17073	8.3	16.7	18578	18.1	35.8	16793

Table 51: Per-dataset rollout-length results for the remaining settings in Figure 3 and 256-position controls for Qwen3-1.7B and OLMo-3-7B-Think. First-256 reuses the 256-token rollout row; Uni-256 and Final-256 use a 1,024-token rollout. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted mean.

Model	Training rollout	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B Th/Th	Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	128	47.5	70.0	19086	37.1	56.7	20269	31.7	60.0	20086	21.7	46.7	21820	34.5	58.3	20315
	256	46.3	70.0	20783	36.7	66.7	21697	45.0	73.3	20995	24.6	40.0	22306	38.1	62.5	21445
	512	50.0	76.7	20560	31.7	56.7	21037	38.3	63.3	20711	22.9	50.0	22648	35.7	61.7	21239
	1024	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
	1024 / Uni-256	48.8	76.7	21875	31.7	60.0	21873	36.7	63.3	22032	23.8	46.7	24462	35.2	61.7	22560
	1024 / Final-256	42.5	73.3	23510	34.6	53.3	22599	37.1	63.3	22987	22.9	46.7	25800	34.3	59.2	23724
	2048	42.5	73.3	22598	32.5	63.3	22135	36.3	63.3	21822	19.6	40.0	25466	32.7	60.0	23005
	4096	41.3	70.0	23134	32.5	53.3	22858	32.9	56.7	23500	21.3	36.7	26796	32.0	54.2	24072
	6144	36.7	63.3	23970	32.1	56.7	22655	33.8	53.3	22585	20.8	36.7	27047	30.8	52.5	24064
Qwen3-4B Th/Th	Base	74.6	90.0	15495	65.4	80.0	18696	65.8	80.0	17249	43.3	66.7	19636	62.3	79.2	17769
	128	70.8	86.7	16054	67.5	76.7	18469	67.9	83.3	17432	47.1	66.7	19740	63.3	78.3	17924
	256	72.9	83.3	16485	63.8	83.3	19818	63.3	83.3	18075	43.8	53.3	21951	60.9	75.8	19082
	512	71.7	86.7	16861	60.0	76.7	20223	62.9	83.3	18215	41.7	66.7	22330	59.1	78.3	19407
	1024	66.3	80.0	18395	62.9	80.0	19655	62.9	83.3	18119	39.6	60.0	23625	57.9	75.8	19949
	2048	63.3	83.3	17684	56.7	80.0	18741	57.9	76.7	17374	35.0	53.3	21690	53.2	73.3	18872
	4096	60.0	83.3	18233	60.8	80.0	17913	55.0	83.3	17629	28.3	56.7	22830	51.0	75.8	19151
	6144	59.2	83.3	18954	57.5	83.3	18694	57.5	73.3	17645	32.1	60.0	22426	51.6	75.0	19430

Table 51 (continued).

Model	Training rollout	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-4B-Instruct NTh/NTh	Base	62.5	86.7	9412	45.8	70.0	8558	55.0	80.0	8434	30.8	50.0	10254	48.5	71.7	9165
	128	60.0	83.3	9790	47.1	70.0	8749	58.8	80.0	9095	32.1	50.0	11587	49.5	70.8	9805
	256	63.8	86.7	10378	47.5	73.3	10414	55.8	73.3	10143	31.2	50.0	12315	49.6	70.8	10813
	512	62.5	86.7	10748	47.1	73.3	10873	55.0	73.3	11005	30.4	53.3	11977	48.8	71.7	11151
	1024	65.0	83.3	10141	45.8	73.3	10539	58.3	76.7	10130	30.8	46.7	11409	50.0	70.0	10555
	2048	62.5	83.3	10870	45.8	63.3	10014	53.8	73.3	9558	32.9	56.7	11763	48.8	69.2	10551
	4096	60.4	76.7	10090	42.9	66.7	11031	60.0	80.0	9245	30.8	43.3	11911	48.5	66.7	10569
	6144	58.8	80.0	10459	44.2	63.3	11440	58.8	80.0	9921	31.7	46.7	11483	48.3	67.5	10826
OLMo-3-7B-Think Th/Th	Base	71.7	86.7	17296	66.7	80.0	19137	71.2	90.0	17560	44.6	70.0	20950	63.5	81.7	18736
	128	71.2	83.3	19310	68.8	83.3	20503	66.2	86.7	19815	45.4	70.0	23308	62.9	80.8	20734
	256	67.9	86.7	20282	62.5	80.0	21406	67.5	83.3	20212	45.4	70.0	23691	60.8	80.0	21398
	512	67.1	80.0	20896	64.2	83.3	22492	67.9	86.7	20801	43.8	63.3	24783	60.7	78.3	22243
	1024	68.8	76.7	20502	55.4	80.0	22027	67.1	86.7	20285	43.8	66.7	24083	58.8	77.5	21724
	1024 / Uni-256	67.1	90.0	19995	60.4	83.3	21781	66.3	86.7	20468	42.5	70.0	23707	59.1	82.5	21488
	1024 / Final-256	70.0	83.3	19567	64.6	83.3	21531	67.1	86.7	20365	43.3	63.3	23137	61.3	79.2	21150
	2048	64.6	83.3	20833	54.6	76.7	23336	66.2	86.7	20502	38.8	60.0	25313	56.0	76.7	22496
	4096	57.5	76.7	21608	47.5	76.7	24474	56.7	83.3	22804	34.6	56.7	25445	49.1	73.3	23583
	6144	52.5	73.3	23037	44.2	70.0	25369	49.2	86.7	24015	32.1	56.7	26872	44.5	71.7	24823
OLMo-3-7B-Instruct NTh/NTh	Base	48.8	76.7	7459	39.6	60.0	7182	42.5	66.7	7277	27.1	40.0	8528	39.5	60.8	7612
	128	55.0	80.0	8329	41.7	63.3	7560	48.8	73.3	7592	25.0	43.3	9661	42.6	65.0	8286
	256	52.1	83.3	8329	41.2	73.3	7928	50.4	70.0	7865	27.9	46.7	9323	42.9	68.3	8361
	512	54.6	80.0	7619	39.6	66.7	7554	49.2	66.7	7329	24.6	46.7	9160	42.0	65.0	7915
	1024	53.8	76.7	7201	39.6	63.3	6818	46.2	66.7	6983	23.8	43.3	8698	40.8	62.5	7425
	2048	52.1	76.7	7034	39.2	60.0	6763	45.0	76.7	6945	24.6	43.3	8600	40.2	64.2	7335
	4096	50.0	80.0	7374	37.9	63.3	6591	44.2	63.3	6828	25.0	40.0	8047	39.3	61.7	7210
	6144	48.8	73.3	7589	39.6	70.0	7030	44.2	70.0	7063	20.4	36.7	8200	38.2	62.5	7471

G.1.3 Extended-Training OPSD Checkpoints

Step 0 is Base, step 100 is the main-text checkpoint, and later rows continue the same matched Th/Th OpenThoughts run to 300 updates. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean.

Table 52: Per-dataset extended-training OPSD results for Figure 2.

Steps	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len
Qwen3-1.7B															
Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
50	47.9	80.0	23206	35.4	60.0	23071	38.3	63.3	23010	22.1	43.3	25737	35.9	61.7	23756
100	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
150	43.3	73.3	19383	31.7	56.7	19359	36.3	60.0	18536	20.4	43.3	21521	32.9	58.3	19700
200	41.7	73.3	19724	32.5	46.7	18206	30.4	60.0	18620	21.3	40.0	20508	31.5	55.0	19264
250	40.8	70.0	19622	27.5	43.3	19806	34.2	63.3	17825	21.3	33.3	20405	30.9	52.5	19414
300	42.1	73.3	19243	29.6	60.0	18902	33.8	56.7	18687	23.8	40.0	20578	32.3	57.5	19353
OLMo-3-7B-Think															
Base	71.7	86.7	17296	66.7	80.0	19137	71.2	90.0	17560	44.6	70.0	20950	63.5	81.7	18736
50	66.7	80.0	20029	66.3	83.3	20831	65.8	86.7	19779	43.3	70.0	24332	60.5	80.0	21243
100	68.8	76.7	20502	55.4	80.0	22027	67.1	86.7	20285	43.8	66.7	24083	58.8	77.5	21724
150	66.7	80.0	19975	61.7	86.7	22177	62.1	86.7	21138	39.2	60.0	24007	57.4	78.3	21824
200	59.6	83.3	20854	49.2	76.7	23271	58.8	86.7	21628	38.8	56.7	24472	51.6	75.8	22556
250	62.9	83.3	20572	49.6	76.7	23718	52.1	83.3	22067	34.6	53.3	25399	49.8	74.2	22939
300	61.3	80.0	20792	50.0	73.3	22752	55.0	83.3	21970	38.3	56.7	25127	51.1	73.3	22660

G.1.4 Teacher-Side Context

All rows use OpenThoughts Math 30k. None denotes Base without privileged context. For Qwen3-1.7B and OLMo-3-7B-Think, CoT + solution uses the same examples and order as Full solution. The teacher context and prompt cap (8,192 instead of 1,024 tokens) change.

Table 53: Per-dataset teacher-context results for Table 3 and Table 4. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean.

Model	Rollout	Privileged context	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
			Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	–	None	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	256	Full solution	46.3	70.0	20783	36.7	66.7	21697	45.0	73.3	20995	24.6	40.0	22306	38.1	62.5	21445
	256	Answer only	49.6	70.0	23223	38.8	60.0	24084	38.3	66.7	22631	25.0	40.0	25702	37.9	59.2	23910
	256	Unrelated	47.9	73.3	19074	40.0	63.3	19943	36.3	66.7	20098	23.3	46.7	20165	36.9	62.5	19820
	1024	Full solution	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
	1024	Answer only	47.1	76.7	25291	34.2	56.7	25710	39.6	70.0	24298	21.7	33.3	27985	35.6	59.2	25821
	1024	Unrelated	44.2	70.0	21601	36.7	66.7	22245	42.5	60.0	21496	25.0	43.3	22792	37.1	60.0	22034
	1024	CoT + solution	35.4	70.0	18947	30.0	53.3	17435	32.9	56.7	18752	25.0	43.3	19299	30.8	55.8	18608
Qwen3-4B	–	None	74.6	90.0	15495	65.4	80.0	18696	65.8	80.0	17249	43.3	66.7	19636	62.3	79.2	17769
	256	Full solution	72.9	83.3	16485	63.8	83.3	19818	63.3	83.3	18075	43.8	53.3	21951	60.9	75.8	19082
	256	Answer only	68.3	86.7	19266	62.9	73.3	21919	65.8	83.3	20063	40.8	56.7	23971	59.5	75.0	21305
	256	Unrelated	70.8	83.3	16952	66.3	80.0	20057	67.9	83.3	18752	42.1	60.0	21451	61.8	76.7	19303
	1024	Full solution	66.3	80.0	18395	62.9	80.0	19655	62.9	83.3	18119	39.6	60.0	23625	57.9	75.8	19949
	1024	Answer only	64.6	80.0	20633	58.8	73.3	22577	62.5	83.3	20981	40.0	60.0	25504	56.5	74.2	22424
	1024	Unrelated	67.1	83.3	22335	62.1	86.7	25174	65.4	83.3	23074	43.8	60.0	26029	59.6	78.3	24153
Qwen3-4B-Instruct	–	None	62.5	86.7	9412	45.8	70.0	8558	55.0	80.0	8434	30.8	50.0	10254	48.5	71.7	9165
	256	Full solution	63.8	86.7	10378	47.5	73.3	10414	55.8	73.3	10143	31.2	50.0	12315	49.6	70.8	10813
	256	Answer only	59.6	83.3	9867	44.2	70.0	9542	55.4	80.0	9266	31.7	46.7	11185	47.7	70.0	9965
	256	Unrelated	63.3	83.3	9095	47.9	73.3	8854	53.8	76.7	8264	27.9	43.3	9605	48.2	69.2	8954
	1024	Full solution	65.0	83.3	10141	45.8	73.3	10539	58.3	76.7	10130	30.8	46.7	11409	50.0	70.0	10555
	1024	Answer only	62.1	83.3	9913	45.0	70.0	10016	52.1	83.3	9309	23.8	46.7	11450	45.7	70.8	10172
	1024	Unrelated	53.8	73.3	11368	41.2	60.0	11797	52.9	66.7	9520	26.2	36.7	11686	43.5	59.2	11093
OLMo-3-7B-Think	–	None	71.7	86.7	17296	66.7	80.0	19137	71.2	90.0	17560	44.6	70.0	20950	63.5	81.7	18736
	256	Full solution	67.9	86.7	20282	62.5	80.0	21406	67.5	83.3	20212	45.4	70.0	23691	60.8	80.0	21398
	256	Answer only	69.2	80.0	20708	63.8	80.0	22298	66.2	86.7	21301	42.1	60.0	23954	60.3	76.7	22065
	256	Unrelated	66.7	83.3	19971	63.3	83.3	21395	63.3	83.3	20782	44.6	63.3	23084	59.5	78.3	21308
	1024	Full solution	68.8	76.7	20502	55.4	80.0	22027	67.1	86.7	20285	43.8	66.7	24083	58.8	77.5	21724
	1024	Answer only	64.6	80.0	22921	59.2	76.7	24151	61.2	86.7	23492	39.6	60.0	25903	56.1	75.8	24117
	1024	Unrelated	62.5	80.0	22662	58.3	76.7	23626	64.6	83.3	22398	44.6	66.7	24912	57.5	76.7	23400
	1024	CoT + solution	64.6	80.0	19337	64.6	80.0	21006	62.1	90.0	20053	43.3	60.0	22843	58.6	77.5	20810
OLMo-3-7B-Instruct	–	None	48.8	76.7	7459	39.6	60.0	7182	42.5	66.7	7277	27.1	40.0	8528	39.5	60.8	7612
	256	Full solution	52.1	83.3	8329	41.2	73.3	7928	50.4	70.0	7865	27.9	46.7	9323	42.9	68.3	8361
	256	Answer only	32.9	60.0	4603	22.1	40.0	3889	26.7	56.7	4482	14.2	33.3	4500	24.0	47.5	4368
	256	Unrelated	50.4	80.0	8152	37.1	60.0	7741	44.6	70.0	7590	20.8	36.7	8035	38.2	61.7	7880
	1024	Full solution	53.8	76.7	7201	39.6	63.3	6818	46.2	66.7	6983	23.8	43.3	8698	40.8	62.5	7425
	1024	Answer only	39.6	66.7	4595	27.9	46.7	4104	28.8	63.3	4594	14.2	30.0	5018	27.6	51.7	4578
	1024	Unrelated	55.8	80.0	9579	40.4	60.0	8707	47.9	70.0	9305	21.2	40.0	10835	41.4	62.5	9606

G.1.4.1 OpenMathReasoning Replication. We repeat the context comparison on Qwen3-1.7B using two disjoint OpenMathReasoning subsets defined by post-thinking length. Solution and CoT + solution use identical examples and ordering within each subset. Adding CoT lowers Avg@8 from 34.4 to 27.8 on the at-most-1k subset and from 34.7 to 29.0 on the 2–4k subset (Table 54).

Table 54: Qwen3-1.7B teacher-context replication on the OpenMathReasoning `cot` split with a 1,024-token student rollout cap. Each pair uses identical examples. Subset labels give the rendered post-thinking solution length before CoT is added. CoT + solution uses a 12,288-token teacher cap.

Post-thinking subset	Privileged context	Teacher cap	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
			Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
$\leq$1k	Solution	1,024	45.8	73.3	22865	35.0	60.0	22109	34.6	63.3	21651	22.1	40.0	24262	34.4	59.2	22722
	CoT + solution	12,288	37.9	76.7	16997	27.9	53.3	16101	28.3	56.7	16444	17.1	36.7	18137	27.8	55.8	16920
2–4k	Solution	4,096	45.4	66.7	21721	37.1	66.7	22442	32.5	56.7	22517	23.8	46.7	24077	34.7	59.2	22689
	CoT + solution	12,288	38.3	80.0	17884	27.9	43.3	16633	27.1	53.3	16678	22.5	40.0	17620	29.0	54.2	17204

G.1.5 Divergence-Objective Ablation

Forward KL uses $\beta = 0$, symmetric JSD uses $\beta = 0.5$, and reverse KL uses $\beta = 1$.

Table 55: Per-dataset divergence-ablation results for Table 5 and Table 16. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean.

Model	Objective	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	Fwd. KL ($\beta = 0$)	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
	JSD ($\beta = 0.5$)	40.0	70.0	14950	32.9	53.3	15222	33.3	60.0	14764	18.8	36.7	15931	31.3	55.0	15217
	Rev. KL ($\beta = 1$)	37.5	63.3	23381	31.7	46.7	22391	23.8	46.7	23906	12.9	33.3	27968	26.5	47.5	24411
Qwen3-4B	Fwd. KL ($\beta = 0$)	66.3	80.0	18395	62.9	80.0	19655	62.9	83.3	18119	39.6	60.0	23625	57.9	75.8	19949
	JSD ($\beta = 0.5$)	67.1	83.3	12485	59.6	76.7	14881	60.8	83.3	13550	38.8	60.0	15634	56.6	75.8	14137
	Rev. KL ($\beta = 1$)	37.9	73.3	27816	36.3	56.7	27897	33.8	60.0	26749	20.0	36.7	29080	32.0	56.7	27886
Qwen3-4B-Thinking-2507	Fwd. KL ($\beta = 0$)	73.3	93.3	40191	70.4	83.3	44970	73.3	90.0	41319	46.3	66.7	46992	65.8	83.3	43368
	JSD ($\beta = 0.5$)	76.7	86.7	15840	73.3	86.7	17668	74.2	83.3	17130	47.9	73.3	20247	68.0	82.5	17721
	Rev. KL ($\beta = 1$)	83.8	93.3	30054	79.6	90.0	34265	74.2	83.3	33081	59.2	73.3	39100	74.2	85.0	34125
OLMo-3-7B-Think	Fwd. KL ($\beta = 0$)	68.8	76.7	20502	55.4	80.0	22027	67.1	86.7	20285	43.8	66.7	24083	58.8	77.5	21724
	JSD ($\beta = 0.5$)	70.0	83.3	15345	64.6	83.3	17685	70.0	90.0	16173	43.8	70.0	20154	62.1	81.7	17339
	Rev. KL ($\beta = 1$)	72.9	93.3	18338	65.0	83.3	20245	66.3	86.7	19107	45.4	66.7	22097	62.4	82.5	19947
OLMo-3-7B-Instruct	Fwd. KL ($\beta = 0$)	53.8	76.7	7201	39.6	63.3	6818	46.3	66.7	6983	23.8	43.3	8698	40.8	62.5	7425
	JSD ($\beta = 0.5$)	50.4	80.0	7414	38.8	60.0	6771	39.6	60.0	7270	20.0	33.3	8551	37.2	58.3	7501
	Rev. KL ($\beta = 1$)	57.5	76.7	8439	40.0	63.3	7775	47.9	66.7	8368	24.6	40.0	9583	42.5	61.7	8541
Falcon-H1R-7B	Fwd. KL ($\beta = 0$)	91.3	96.7	19401	87.9	93.3	21056	93.3	96.7	20100	83.8	100.0	27668	89.1	96.7	22056
	JSD ($\beta = 0.5$)	92.1	96.7	17352	87.5	93.3	20365	91.7	93.3	17862	82.5	100.0	25862	88.4	95.8	20360
	Rev. KL ($\beta = 1$)	89.6	96.7	18012	87.5	96.7	20396	90.8	96.7	18069	84.6	96.7	25214	88.1	96.7	20423
MiMo-7B-RL	Fwd. KL ($\beta = 0$)	54.6	76.7	15269	48.8	73.3	15996	50.0	66.7	15811	23.8	40.0	19524	44.3	64.2	16650
	JSD ($\beta = 0.5$)	65.8	80.0	12495	54.6	76.7	14069	56.7	80.0	13370	33.3	53.3	16558	52.6	72.5	14123
	Rev. KL ($\beta = 1$)	60.8	83.3	18295	48.3	70.0	19721	57.5	76.7	18197	34.2	60.0	22142	50.2	72.5	19589

G.1.6 STUDENT-ENTROPY LOSS MASKING

All runs use a 256-token rollout. Each dataset reports Avg@8, Pass@8, and mean output length.

Table 56: Per-dataset student-entropy loss-masking results for Table 17. Full uses all response tokens. HE20/HE80 and LE20/LE80 retain the highest- or lowest-entropy fraction of each completion.

Model	Subset	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	Full	46.3	70.0	20783	36.7	66.7	21697	45.0	73.3	20995	24.6	40.0	22306	38.1	62.5	21445
	HE20	44.6	76.7	21009	36.7	63.3	22166	39.6	73.3	21696	26.7	50.0	22074	36.9	65.8	21736
	HE80	52.5	76.7	20083	34.6	56.7	21365	40.0	70.0	20013	24.2	43.3	21676	37.8	61.7	20784
	LE20	49.6	73.3	17001	37.9	63.3	17766	39.2	73.3	18538	23.8	46.7	18619	37.6	64.2	17981
	LE80	45.8	73.3	17275	37.5	60.0	16975	30.0	70.0	17773	19.6	33.3	18633	33.2	59.2	17664
Qwen3-4B	Base	74.6	90.0	15495	65.4	80.0	18696	65.8	80.0	17249	43.3	66.7	19636	62.3	79.2	17769
	Full	72.9	83.3	16485	63.8	83.3	19818	63.3	83.3	18075	43.8	53.3	21951	60.9	75.8	19082
	HE20	72.1	83.3	17707	67.1	83.3	20862	65.0	83.3	18957	44.6	66.7	21570	62.2	79.2	19774
	HE80	74.2	86.7	15964	67.1	76.7	19567	65.4	83.3	18045	42.9	60.0	21436	62.4	76.7	18753
	LE20	71.3	86.7	14544	65.4	80.0	18244	63.8	83.3	16468	39.2	60.0	19038	59.9	77.5	17073
	LE80	71.7	83.3	14497	61.7	83.3	17405	62.5	80.0	16126	40.4	63.3	18115	59.1	77.5	16536
Qwen3-4B-Thinking-2507	Base	82.1	93.3	20319	81.7	86.7	22385	81.2	86.7	21835	56.7	80.0	26570	75.4	86.7	22777
	Full	80.0	90.0	28157	75.4	83.3	30535	75.8	90.0	29163	50.0	70.0	34249	70.3	83.3	30526
	HE20	78.3	90.0	27507	77.9	86.7	30304	80.4	93.3	28946	50.4	70.0	33683	71.8	85.0	30110
	HE80	81.7	93.3	27949	75.8	86.7	30466	80.0	90.0	28820	50.8	63.3	33618	72.1	83.3	30213
	LE20	82.9	90.0	20451	80.4	86.7	22139	81.2	93.3	21194	50.8	70.0	26172	73.9	85.0	22489
	LE80	80.8	86.7	23551	81.2	86.7	25817	80.0	90.0	24942	55.8	70.0	29674	74.5	83.3	25996
Qwen3-4B-Instruct	Base	62.5	86.7	9412	45.8	70.0	8558	55.0	80.0	8434	30.8	50.0	10254	48.5	71.7	9165
	Full	63.8	86.7	10378	47.5	73.3	10414	55.8	73.3	10143	31.2	50.0	12315	49.6	70.8	10813
	HE20	63.8	83.3	9759	45.4	73.3	10840	58.8	80.0	9289	31.2	50.0	11281	49.8	71.7	10292
	HE80	60.0	83.3	10720	45.4	73.3	11324	57.1	76.7	9767	30.8	50.0	13143	48.3	70.8	11238
	LE20	63.8	76.7	9913	49.6	73.3	9075	60.8	76.7	8062	33.3	60.0	11582	51.9	71.7	9658
	LE80	65.4	80.0	9603	46.2	66.7	8721	55.8	80.0	8899	32.5	50.0	10999	50.0	69.2	9556
OLMo-3-7B-Think	Base	71.7	86.7	17296	66.7	80.0	19137	71.2	90.0	17560	44.6	70.0	20950	63.5	81.7	18736
	Full	67.9	86.7	20282	62.5	80.0	21406	67.5	83.3	20212	45.4	70.0	23691	60.8	80.0	21398
	HE20	69.6	86.7	19609	64.2	80.0	21386	69.6	86.7	19989	45.4	73.3	23829	62.2	81.7	21203
	HE80	69.6	80.0	19730	63.8	80.0	22001	62.5	83.3	20629	43.8	66.7	24235	59.9	77.5	21649
	LE20	72.5	86.7	17819	62.9	86.7	19437	69.2	90.0	17872	45.0	66.7	21031	62.4	82.5	19040
	LE80	73.3	93.3	17884	65.0	83.3	19981	62.9	86.7	19255	43.8	66.7	22259	61.2	82.5	19845
OLMo-3-7B-Instruct	Base	48.8	76.7	7459	39.6	60.0	7182	42.5	66.7	7277	27.1	40.0	8528	39.5	60.8	7612
	Full	52.1	83.3	8329	41.2	73.3	7928	50.4	70.0	7865	27.9	46.7	9323	42.9	68.3	8361
	HE20	52.9	80.0	7967	41.2	66.7	7464	50.4	76.7	7763	24.2	53.3	9374	42.2	69.2	8142
	HE80	51.2	80.0	7930	42.1	73.3	7497	52.9	76.7	7440	22.9	43.3	9530	42.3	68.3	8099
	LE20	51.2	80.0	7652	40.0	66.7	7468	46.2	73.3	7465	23.3	46.7	8976	40.2	66.7	7890
	LE80	56.7	76.7	8288	43.8	66.7	7999	47.9	76.7	8025	27.9	56.7	9573	44.1	69.2	8471

G.1.7 VOCABULARY TOP-k DISTILLATION LOSS

Full-vocab rows are matched 100-step controls: forward KL for Qwen3-1.7B and Qwen3-4B-Instruct, and reverse KL for OLMo-3-7B-Think.

Table 57: Per-dataset vocabulary top-k results for Table 19. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean.

Model	Support	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	Full vocab	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
	$k = 16$	42.5	80.0	22299	31.7	56.7	22440	34.6	63.3	21396	23.3	50.0	24293	33.0	62.5	22607
	$k = 4$	28.3	60.0	28891	28.8	46.7	28108	31.3	53.3	26761	15.0	30.0	31383	25.8	47.5	28786
Qwen3-4B-Instruct	Full vocab	65.0	83.3	10141	45.8	73.3	10539	58.3	76.7	10130	30.8	46.7	11409	50.0	70.0	10555
	$k = 16$	63.8	80.0	10156	50.8	70.0	10317	59.6	80.0	9726	33.8	53.3	12112	52.0	70.8	10578
	$k = 4$	57.5	80.0	8275	45.4	66.7	7677	49.6	76.7	6874	30.4	53.3	9850	45.7	69.2	8169
OLMo-3-7B-Think	Full vocab	72.9	93.3	18338	65.0	83.3	20245	66.3	86.7	19107	45.4	66.7	22097	62.4	82.5	19947
	$k = 16$	70.0	86.7	18976	62.9	86.7	19850	64.6	86.7	18940	42.1	66.7	23117	59.9	81.7	20221
	$k = 4$	4.2	23.3	32285	2.9	23.3	32015	4.6	26.7	32082	2.5	20.0	32482	3.5	23.3	32216

G.1.8 INITIAL MODEL ABILITY

Checkpoint denotes Base or the rollout cap used for full-solution OPSD. Rows follow the Base Pass@8 ordering in Figure 4. Each dataset reports Avg@8, Pass@8, and mean output length.

Table 58: Per-dataset results for Figure 4. 256 and 1024 are full-solution OPSD conditions. Dataset Avg. is the unweighted four-benchmark mean.

Model	Checkpoint	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-0.6B	Base	8.8	30.0	16815	17.5	40.0	15188	12.9	26.7	16334	8.8	23.3	15532	12.0	30.0	15967
	256	10.8	30.0	21812	18.3	36.7	19027	10.4	30.0	21177	9.2	26.7	19708	12.2	30.8	20431
	1024	9.2	30.0	20924	12.5	30.0	18855	8.3	16.7	19339	7.5	16.7	20224	9.4	23.3	19835
DeepSeek-R1-Distill-Qwen-1.5B	Base	28.3	53.3	17211	24.6	46.7	15944	19.6	43.3	17906	14.2	40.0	18742	21.7	45.8	17451
	256	28.8	66.7	17814	23.3	43.3	17795	22.5	50.0	18387	15.0	26.7	18842	22.4	46.7	18209
	1024	30.4	63.3	17405	21.7	40.0	16524	20.4	46.7	18662	16.7	30.0	18357	22.3	45.0	17737
Qwen3-1.7B	Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	256	46.3	70.0	20783	36.7	66.7	21697	45.0	73.3	20995	24.6	40.0	22306	38.1	62.5	21445
	1024	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
OLMo-3-7B-Instruct	Base	48.8	76.7	7459	39.6	60.0	7182	42.5	66.7	7277	27.1	40.0	8528	39.5	60.8	7612
	256	52.1	83.3	8329	41.2	73.3	7928	50.4	70.0	7865	27.9	46.7	9323	42.9	68.3	8361
	1024	53.8	76.7	7201	39.6	63.3	6818	46.2	66.7	6983	23.8	43.3	8698	40.8	62.5	7425
Qwen3-4B-Instruct	Base	62.5	86.7	9412	45.8	70.0	8558	55.0	80.0	8434	30.8	50.0	10254	48.5	71.7	9165
	256	63.8	86.7	10378	47.5	73.3	10414	55.8	73.3	10143	31.2	50.0	12315	49.6	70.8	10813
	1024	65.0	83.3	10141	45.8	73.3	10539	58.3	76.7	10130	30.8	46.7	11409	50.0	70.0	10555
MiMo-7B-RL	Base	66.2	83.3	14824	61.2	80.0	16769	59.6	83.3	15674	35.4	50.0	19549	55.6	74.2	16704
	256	67.9	83.3	15634	55.8	80.0	17449	62.5	83.3	16580	35.0	53.3	20466	55.3	75.0	17532
	1024	54.6	76.7	15269	48.8	73.3	15996	50.0	66.7	15811	23.8	40.0	19524	44.3	64.2	16650
Qwen3-8B	Base	76.3	86.7	14860	69.6	80.0	18059	67.1	86.7	16744	43.3	60.0	20924	64.1	78.3	17647
	256	71.7	80.0	17976	67.5	83.3	20351	66.7	86.7	18616	42.9	63.3	22795	62.2	78.3	19934
	1024	67.5	76.7	19649	62.5	83.3	21001	65.0	86.7	19844	40.4	60.0	25193	58.9	76.7	21422
Qwen3-4B	Base	74.6	90.0	15495	65.4	80.0	18696	65.8	80.0	17249	43.3	66.7	19636	62.3	79.2	17769
	256	72.9	83.3	16485	63.8	83.3	19818	63.3	83.3	18075	43.8	53.3	21951	60.9	75.8	19082
	1024	66.3	80.0	18395	62.9	80.0	19655	62.9	83.3	18119	39.6	60.0	23625	57.9	75.8	19949
OLMo-3-7B-Think	Base	71.7	86.7	17296	66.7	80.0	19137	71.2	90.0	17560	44.6	70.0	20950	63.5	81.7	18736
	256	67.9	86.7	20282	62.5	80.0	21406	67.5	83.3	20212	45.4	70.0	23691	60.8	80.0	21398
	1024	68.8	76.7	20502	55.4	80.0	22027	67.1	86.7	20285	43.8	66.7	24083	58.8	77.5	21724
Qwen3-4B-Thinking-2507	Base	82.1	93.3	20319	81.7	86.7	22385	81.2	86.7	21835	56.7	80.0	26570	75.4	86.7	22777
	256	80.0	90.0	28157	75.4	83.3	30535	75.8	90.0	29164	50.0	70.0	34249	70.3	83.3	30526
	1024	73.3	93.3	40191	70.4	83.3	44970	73.3	90.0	41319	46.2	66.7	46992	65.8	83.3	43368
Falcon-H1R-7B	Base	91.7	96.7	17498	88.8	96.7	20960	91.7	93.3	18905	85.8	100.0	25429	89.5	96.7	20698
	256	92.1	100.0	18825	87.5	96.7	20933	91.2	96.7	19202	84.6	93.3	26952	88.9	96.7	21478
	1024	91.2	96.7	19401	87.9	93.3	21056	93.3	96.7	20100	83.8	100.0	27668	89.1	96.7	22056

G.1.9 SFT INITIALIZATION SWEEP

Direct evaluates the SFT checkpoint. After OPSD evaluates the matched 100-step frozen-self-teacher checkpoint. Base denotes zero SFT updates. Each dataset reports Avg@8, Pass@8, and

mean output length. Dataset Avg. is the unweighted four-benchmark mean. The 7.5k and 12.5k rows report direct SFT.

Table 59: Per-dataset Qwen3-1.7B-Base SFT-initialization results for Table 22. Thinking-mode evaluation uses a 38,912-token maximum output length.

SFT steps	Evaluation	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Base	Direct	3.8	13.3	6629	1.2	10.0	5214	3.3	10.0	5654	0.0	0.0	4805	2.1	8.3	5576
5k	Direct	8.8	26.7	24373	9.6	23.3	23444	7.5	23.3	23689	2.5	13.3	23591	7.1	21.7	23774
7.5k	Direct	9.2	26.7	24317	10.0	26.7	23141	9.6	26.7	25677	4.2	20.0	23953	8.2	25.0	24272
10k	Direct	10.4	20.0	22854	12.9	23.3	18654	10.4	23.3	20890	4.6	16.7	20950	9.6	20.8	20837
12.5k	Direct	12.9	33.3	20910	16.2	36.7	19194	10.8	30.0	21869	4.2	13.3	20361	11.0	28.3	20584
15k	Direct	12.9	36.7	21830	15.8	36.7	20261	12.1	30.0	22976	5.0	13.3	20620	11.5	29.2	21422
Base	After OPSD	2.9	16.7	8339	0.4	3.3	7601	1.2	10.0	8881	0.0	0.0	6987	1.1	7.5	7952
5k	After OPSD	0.8	6.7	35018	0.4	3.3	34134	0.0	0.0	34849	0.0	0.0	34599	0.3	2.5	34650
10k	After OPSD	5.4	16.7	34934	3.8	13.3	34525	2.5	10.0	35201	0.8	6.7	36296	3.1	11.7	35239
15k	After OPSD	2.5	13.3	35896	1.7	10.0	36323	2.1	6.7	35962	0.0	0.0	36550	1.6	7.5	36183

Table 60: Per-dataset Qwen3-4B-Base SFT-initialization results for Table 22.

SFT steps	Evaluation	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Base	Direct	12.1	26.7	3692	7.5	26.7	2707	8.3	20.0	2889	1.2	6.7	2863	7.3	20.0	3038
5k	Direct	30.4	66.7	16251	30.0	53.3	16427	25.4	50.0	16250	16.7	33.3	18485	25.6	50.8	16853
7.5k	Direct	35.0	63.3	14817	33.8	56.7	14477	31.2	56.7	15043	20.0	33.3	15893	30.0	52.5	15057
10k	Direct	37.1	70.0	13999	34.6	56.7	13907	36.7	66.7	14128	20.8	50.0	14470	32.3	60.8	14126
12.5k	Direct	42.5	73.3	13632	32.5	56.7	14781	33.3	66.7	14865	20.0	33.3	15521	32.1	57.5	14700
15k	Direct	45.8	73.3	13843	34.2	63.3	13968	36.2	60.0	13625	22.5	43.3	15708	34.7	60.0	14286
Base	After OPSD	7.9	26.7	9927	4.6	23.3	10265	4.2	10.0	10319	2.5	10.0	10718	4.8	17.5	10308
5k	After OPSD	0.4	3.3	33234	1.7	6.7	32905	1.2	6.7	33055	0.0	0.0	33277	0.8	4.2	33118
10k	After OPSD	1.2	6.7	33153	2.9	13.3	32525	1.7	10.0	32934	0.8	3.3	32766	1.7	8.3	32844
15k	After OPSD	2.1	13.3	33095	3.3	10.0	32466	2.5	10.0	32770	0.0	0.0	33220	2.0	8.3	32888

The table expands the first six rows of Table 5, from Qwen3-4B-Base through SFT+GRPO+OPSD. Each dataset reports Avg@8, Pass@8, boxed-answer rate, and mean output length.

Table 61: Per-dataset post-training-pipeline results for the first six rows of Table 5. Fmt is the boxed-answer rate. Table 64 compares GRPO-then-OPSD with matched Qwen3-4B Th/Th.

Init. / method	AIME 2024				AIME 2025				AIME 2026				HMMT25			
	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len
Qwen3-4B-Base (direct)	12.1	26.7	86.2	3692	7.5	26.7	89.2	2707	8.3	20.0	86.7	2889	1.2	6.7	86.7	2863
+ 100-step OPSD	7.9	26.7	74.2	9927	4.6	23.3	76.2	10265	4.2	10.0	73.8	10319	2.5	10.0	74.2	10718
SFT-15k (direct)	45.8	73.3	89.6	13843	34.2	63.3	90.4	13968	36.2	60.0	88.8	13625	22.5	43.3	88.3	15708
+ 100-step OPSD	2.1	13.3	3.3	33095	3.3	10.0	5.0	32466	2.5	10.0	3.3	32770	0.0	0.0	2.1	33220
+ GRPO	43.8	73.3	93.3	14430	34.6	56.7	96.2	13451	35.8	70.0	93.3	13860	20.4	43.3	96.2	14990
+ GRPO + OPSD	3.3	10.0	7.1	37973	5.0	13.3	5.4	37043	3.3	20.0	5.4	37485	2.5	13.3	4.6	37812

The 50-step OPSD and GRPO runs start from released Qwen3-1.7B and Qwen3-4B Th checkpoints and consume exactly the same 3,200 OpenMathReasoning problems in the same order. Each dataset reports Avg@8, Pass@8, boxed-answer rate, and mean output length. Table 49 gives the Base rows.

Table 62: Per-dataset 50-step GRPO results on released Qwen3 thinking checkpoints for Table 6. Fmt is the boxed-answer rate.

Checkpoint	AIME 2024				AIME 2025				AIME 2026				HMMT25			
	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len
Qwen3-1.7B + GRPO-50	47.1	76.7	95.8	19761	37.1	63.3	97.5	19985	38.8	66.7	97.1	20334	25.4	46.7	95.8	21751
Qwen3-4B + GRPO-50	72.5	83.3	97.5	16342	67.9	86.7	92.9	19731	67.1	83.3	92.5	18507	43.8	60.0	91.3	22044

Table 63: Per-dataset 50-step OPSD results on the same OpenMathReasoning examples and in the same order as GRPO, corresponding to Table 6. OPSD uses the post-thinking solution as privileged teacher context. Fmt is the boxed-answer rate.

Checkpoint	AIME 2024				AIME 2025				AIME 2026				HMMT25			
	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len
Qwen3-1.7B + OPSD-50	48.3	76.7	87.9	21217	40.4	66.7	93.3	21169	40.0	63.3	88.3	21876	23.3	46.7	91.3	23860
Qwen3-4B + OPSD-50	70.4	83.3	91.7	17829	63.8	86.7	88.3	20209	61.3	83.3	88.8	18498	45.0	66.7	89.2	22175

G.1.10 OPSD AFTER SFT THEN GRPO

Qwen3-4B uses the matched Th/Th 100-step run. 4B SFT-15k + GRPO is the GRPO step-100 actor, and + OPSD applies the same matched-mode OPSD protocol. Fmt is the boxed-answer rate. Length is mean output tokens.

Table 64: Per-dataset results for Table 24. Fmt is the boxed-answer rate.

Checkpoint	Stage	AIME 2024				AIME 2025				AIME 2026				HMMT25			
		Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len	Avg	Pass	Fmt	Len
Qwen3-4B (Th/Th)	Base	74.6	90.0	98.8	15495	65.4	80.0	96.2	18696	65.8	80.0	95.0	17249	43.3	66.7	98.3	19636
	+ OPSD	66.3	80.0	86.7	18395	62.9	80.0	86.2	19655	62.9	83.3	89.6	18119	39.6	60.0	82.9	23625
4B SFT-15k + GRPO	GRPO-100	43.8	73.3	93.3	14430	34.6	56.7	96.2	13451	35.8	70.0	93.3	13860	20.4	43.3	96.2	14990
	+ OPSD	3.3	10.0	7.1	37973	5.0	13.3	5.4	37043	3.3	20.0	5.4	37485	2.5	13.3	4.6	37812

G.2 SEED-SENSITIVITY BREAKDOWNS

G.2.1 SEED SENSITIVITY: BASE CHECKPOINTS (EVALUATION SEED)

Table 65: Per-dataset Base results for Table 40. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean.

Model	Eval seed	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	42	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	1024	47.5	76.7	17907	36.7	53.3	18615	35.8	56.7	17919	20.4	33.3	19812	35.1	55.0	18563
	65536	42.9	73.3	18102	37.9	56.7	18542	33.3	66.7	18614	27.1	50.0	19196	35.3	61.7	18613
OLMo-3-7B-Think	42	71.7	86.7	17296	66.7	80.0	19137	71.2	90.0	17560	44.6	70.0	20950	63.5	81.7	18736
	1024	67.9	83.3	17617	64.2	86.7	19148	72.1	86.7	17955	44.6	66.7	21317	62.2	80.8	19009
	65536	70.8	86.7	17101	65.8	76.7	19099	70.0	86.7	17775	45.8	66.7	21497	63.1	79.2	18868

G.2.2 SEED SENSITIVITY: ROLLOUT LENGTH (TRAINING SEED)

Table 66: Per-dataset training-seed results for Table 41. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean. $^{\uparrow}/^{\downarrow}$ mark accuracy changes relative to Base@eval-42. Length is unannotated.

Model	Condition	Train seed	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
			Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	OPSD-256	42	46.2$^{\uparrow}$	70.0$^{\downarrow}$	20783	36.7$^{\uparrow}$	66.7$^{\uparrow}$	21697	45.0$^{\uparrow}$	73.3$^{\uparrow}$	20995	24.6$^{\uparrow}$	40.0$^{\downarrow}$	22306	38.1$^{\uparrow}$	62.5$^{\uparrow}$	21445
		1024	51.2$^{\uparrow}$	83.3$^{\uparrow}$	20253	36.2$^{\uparrow}$	63.3$^{\uparrow}$	21118	42.1$^{\uparrow}$	66.7$^{\uparrow}$	21295	27.1$^{\uparrow}$	46.7	21828	39.2$^{\uparrow}$	65.0$^{\uparrow}$	21124
		65536	50.4$^{\uparrow}$	76.7	20253	37.1$^{\uparrow}$	63.3$^{\uparrow}$	20962	38.3$^{\uparrow}$	70.0$^{\uparrow}$	21160	22.5$^{\uparrow}$	50.0$^{\uparrow}$	22975	37.1$^{\uparrow}$	65.0$^{\uparrow}$	21338
Qwen3-1.7B	OPSD-1024	42	45.8$^{\uparrow}$	73.3$^{\downarrow}$	22175	36.7$^{\uparrow}$	66.7$^{\uparrow}$	22143	36.7$^{\uparrow}$	63.3$^{\uparrow}$	21288	22.1$^{\uparrow}$	40.0$^{\downarrow}$	23913	35.3$^{\uparrow}$	60.8$^{\uparrow}$	22380
		1024	49.2$^{\uparrow}$	73.3$^{\downarrow}$	21091	36.7$^{\uparrow}$	63.3$^{\uparrow}$	21896	37.5$^{\uparrow}$	60.0$^{\uparrow}$	21149	20.4$^{\downarrow}$	30.0$^{\downarrow}$	23892	35.9$^{\uparrow}$	56.7$^{\downarrow}$	22007
		65536	41.7$^{\downarrow}$	76.7	22332	35.8$^{\uparrow}$	53.3$^{\downarrow}$	21324	37.1$^{\uparrow}$	70.0$^{\uparrow}$	21537	23.8$^{\uparrow}$	50.0$^{\uparrow}$	24818	34.6$^{\uparrow}$	62.5$^{\uparrow}$	22503
OLMo-3-7B-Think	OPSD-256	42	67.9$^{\downarrow}$	86.7	20282	62.5$^{\downarrow}$	80.0	21406	67.5$^{\downarrow}$	83.3$^{\downarrow}$	20212	45.4$^{\uparrow}$	70.0	23691	60.8$^{\downarrow}$	80.0$^{\downarrow}$	21398
		1024	70.0$^{\downarrow}$	90.0$^{\uparrow}$	20030	64.2$^{\downarrow}$	80.0	21893	67.5$^{\downarrow}$	86.7$^{\downarrow}$	20009	44.2$^{\downarrow}$	60.0$^{\downarrow}$	24256	61.5$^{\downarrow}$	79.2$^{\downarrow}$	21547
		65536	70.8$^{\downarrow}$	80.0$^{\downarrow}$	20091	66.7	86.7$^{\uparrow}$	21376	65.8$^{\downarrow}$	90.0	20739	44.6	70.0	23802	62.0$^{\downarrow}$	81.7	21502
OLMo-3-7B-Think	OPSD-1024	42	68.8$^{\downarrow}$	76.7$^{\downarrow}$	20502	55.4$^{\downarrow}$	80.0	22027	67.1$^{\downarrow}$	86.7$^{\downarrow}$	20285	43.8$^{\downarrow}$	66.7$^{\downarrow}$	24083	58.8$^{\downarrow}$	77.5$^{\downarrow}$	21724
		1024	67.9$^{\downarrow}$	80.0$^{\downarrow}$	19871	62.5$^{\downarrow}$	80.0	21995	64.6$^{\downarrow}$	90.0	21048	41.7$^{\downarrow}$	63.3$^{\downarrow}$	24189	59.2$^{\downarrow}$	78.3$^{\downarrow}$	21776
		65536	69.2$^{\downarrow}$	86.7	19846	61.2$^{\downarrow}$	83.3$^{\uparrow}$	21643	67.1$^{\downarrow}$	90.0	20151	41.2$^{\downarrow}$	66.7$^{\downarrow}$	24033	59.7$^{\downarrow}$	81.7	21418

G.2.3 SEED SENSITIVITY: ROLLOUT LENGTH (EVALUATION SEED)

Table 67: Per-dataset evaluation-seed results for Table 42. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean. $^{\uparrow}/^{\downarrow}$ mark accuracy changes relative to Base at the matched evaluation seed. Length is unannotated.

Model	Condition	Eval seed	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
			Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	OPSD-256	42	46.2$^{\uparrow}$	70.0$^{\downarrow}$	20783	36.7$^{\uparrow}$	66.7$^{\uparrow}$	21697	45.0$^{\uparrow}$	73.3$^{\uparrow}$	20995	24.6$^{\uparrow}$	40.0$^{\downarrow}$	22306	38.1$^{\uparrow}$	62.5$^{\uparrow}$	21445
		1024	47.9$^{\uparrow}$	70.0$^{\downarrow}$	20152	36.7	60.0$^{\uparrow}$	21390	37.9$^{\uparrow}$	56.7	20678	25.4$^{\uparrow}$	50.0$^{\uparrow}$	21863	37.0$^{\uparrow}$	59.2$^{\uparrow}$	21021
		65536	47.5$^{\uparrow}$	73.3	20297	34.2$^{\downarrow}$	63.3$^{\uparrow}$	21394	41.7$^{\uparrow}$	66.7	21027	23.3$^{\downarrow}$	46.7$^{\downarrow}$	22365	36.7$^{\uparrow}$	62.5$^{\uparrow}$	21271
Qwen3-1.7B	OPSD-1024	42	45.8$^{\uparrow}$	73.3$^{\downarrow}$	22175	36.7$^{\uparrow}$	66.7$^{\uparrow}$	22143	36.7$^{\uparrow}$	63.3$^{\uparrow}$	21288	22.1$^{\uparrow}$	40.0$^{\downarrow}$	23913	35.3$^{\uparrow}$	60.8$^{\uparrow}$	22380
		1024	49.2$^{\uparrow}$	76.7	21279	34.2$^{\downarrow}$	63.3$^{\uparrow}$	21384	38.8$^{\uparrow}$	66.7$^{\uparrow}$	21373	25.0$^{\uparrow}$	53.3$^{\uparrow}$	23458	36.8$^{\uparrow}$	65.0$^{\uparrow}$	21874
		65536	49.6$^{\uparrow}$	73.3	21563	35.0$^{\downarrow}$	63.3$^{\uparrow}$	21611	38.8$^{\uparrow}$	66.7	21850	22.9$^{\downarrow}$	43.3$^{\downarrow}$	23902	36.6$^{\uparrow}$	61.7	22231
OLMo-3-7B-Think	OPSD-256	42	67.9$^{\downarrow}$	86.7	20282	62.5$^{\downarrow}$	80.0	21406	67.5$^{\downarrow}$	83.3$^{\downarrow}$	20212	45.4$^{\uparrow}$	70.0	23691	60.8$^{\downarrow}$	80.0$^{\downarrow}$	21398
		1024	70.0$^{\uparrow}$	90.0$^{\uparrow}$	20461	65.4$^{\uparrow}$	83.3$^{\downarrow}$	21875	67.9$^{\downarrow}$	86.7	20858	44.6	63.3$^{\downarrow}$	23498	62.0$^{\downarrow}$	80.8	21673
		65536	69.6$^{\downarrow}$	86.7	20365	65.0$^{\downarrow}$	80.0$^{\uparrow}$	22070	67.5$^{\downarrow}$	86.7	20694	43.3$^{\downarrow}$	56.7$^{\downarrow}$	23373	61.4$^{\downarrow}$	77.5$^{\downarrow}$	21626
OLMo-3-7B-Think	OPSD-1024	42	68.8$^{\downarrow}$	76.7$^{\downarrow}$	20502	55.4$^{\downarrow}$	80.0	22027	67.1$^{\downarrow}$	86.7$^{\downarrow}$	20285	43.8$^{\downarrow}$	66.7$^{\downarrow}$	24083	58.8$^{\downarrow}$	77.5$^{\downarrow}$	21724
		1024	68.8$^{\uparrow}$	86.7$^{\uparrow}$	19964	62.1$^{\downarrow}$	83.3$^{\downarrow}$	21831	62.5$^{\downarrow}$	83.3$^{\downarrow}$	20286	42.1$^{\downarrow}$	60.0$^{\downarrow}$	24108	58.9$^{\downarrow}$	78.3$^{\downarrow}$	21547
		65536	68.3$^{\downarrow}$	83.3$^{\downarrow}$	19832	62.5$^{\downarrow}$	83.3$^{\uparrow}$	21611	68.8$^{\downarrow}$	86.7	20612	41.7$^{\downarrow}$	63.3$^{\downarrow}$	24196	60.3$^{\downarrow}$	79.2	21563

G.2.4 SEED SENSITIVITY: LOSS-TOKEN POSITION (TRAINING SEED)

Table 68: Per-dataset training-seed results for Table 43. First-256 reuses the corresponding OPSD-256 run. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean. $^{\uparrow}/^{\downarrow}$ mark accuracy changes relative to Base@eval-42. Length is unannotated.

Model	Condition	Train seed	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
			Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	First-256 / OPSD-256	42	46.2$^{\uparrow}$	70.0$^{\downarrow}$	20783	36.7$^{\uparrow}$	66.7$^{\uparrow}$	21697	45.0$^{\uparrow}$	73.3$^{\uparrow}$	20995	24.6$^{\uparrow}$	40.0$^{\downarrow}$	22306	38.1$^{\uparrow}$	62.5$^{\uparrow}$	21445
		1024	51.2$^{\uparrow}$	83.3$^{\uparrow}$	20253	36.2$^{\uparrow}$	63.3$^{\uparrow}$	21118	42.1$^{\uparrow}$	66.7$^{\uparrow}$	21295	27.1$^{\uparrow}$	46.7	21828	39.2$^{\uparrow}$	65.0$^{\uparrow}$	21124
		65536	50.4$^{\uparrow}$	76.7	20253	37.1$^{\uparrow}$	63.3$^{\uparrow}$	20962	38.3$^{\uparrow}$	70.0$^{\uparrow}$	21160	22.5$^{\uparrow}$	50.0$^{\uparrow}$	22975	37.1$^{\uparrow}$	65.0$^{\uparrow}$	21338
Qwen3-1.7B	Uni-256	42	48.8$^{\uparrow}$	76.7	21874	31.7$^{\downarrow}$	60.0	21873	36.7$^{\uparrow}$	63.3$^{\uparrow}$	22032	23.8$^{\uparrow}$	46.7	24462	35.2$^{\uparrow}$	61.7$^{\uparrow}$	22560
		1024	47.5$^{\uparrow}$	73.3$^{\downarrow}$	21809	34.2$^{\downarrow}$	63.3$^{\uparrow}$	22245	34.6$^{\downarrow}$	60.0$^{\uparrow}$	22053	23.3$^{\uparrow}$	40.0$^{\downarrow}$	23502	34.9$^{\uparrow}$	59.2	22402
		65536	47.9$^{\uparrow}$	73.3$^{\downarrow}$	21954	32.9$^{\downarrow}$	63.3$^{\uparrow}$	22906	39.6$^{\uparrow}$	66.7$^{\uparrow}$	22102	25.8$^{\uparrow}$	46.7	24416	36.6$^{\uparrow}$	62.5$^{\uparrow}$	22845
Qwen3-1.7B	Final-256	42	42.5$^{\downarrow}$	73.3$^{\downarrow}$	23510	34.6$^{\downarrow}$	53.3$^{\downarrow}$	22599	37.1$^{\uparrow}$	63.3$^{\uparrow}$	22987	22.9$^{\uparrow}$	46.7	25800	34.3$^{\uparrow}$	59.2	23724
		1024	44.6$^{\uparrow}$	73.3$^{\downarrow}$	23011	32.5$^{\downarrow}$	56.7$^{\downarrow}$	23427	34.2$^{\downarrow}$	53.3	22334	21.7	36.7$^{\downarrow}$	26471	33.2$^{\downarrow}$	55.0$^{\downarrow}$	23811
		65536	43.8	73.3$^{\downarrow}$	23228	33.8$^{\downarrow}$	56.7$^{\downarrow}$	22923	35.4$^{\uparrow}$	60.0$^{\uparrow}$	22885	23.8$^{\uparrow}$	43.3$^{\downarrow}$	25698	34.2$^{\uparrow}$	58.3$^{\downarrow}$	23683
OLMo-3-7B-Think	First-256 / OPSD-256	42	67.9$^{\downarrow}$	86.7	20282	62.5$^{\downarrow}$	80.0	21406	67.5$^{\downarrow}$	83.3$^{\downarrow}$	20212	45.4$^{\uparrow}$	70.0	23691	60.8$^{\downarrow}$	80.0$^{\downarrow}$	21398
		1024	70.0$^{\downarrow}$	90.0$^{\uparrow}$	20030	64.2$^{\downarrow}$	80.0	21893	67.5$^{\downarrow}$	86.7$^{\downarrow}$	20009	44.2$^{\downarrow}$	60.0$^{\downarrow}$	24256	61.5$^{\downarrow}$	79.2$^{\downarrow}$	21547
		65536	70.8$^{\downarrow}$	80.0$^{\downarrow}$	20091	66.7	86.7$^{\uparrow}$	21376	65.8$^{\downarrow}$	90.0	20739	44.6	70.0	23802	62.0$^{\downarrow}$	81.7	21502
OLMo-3-7B-Think	Uni-256	42	67.1$^{\downarrow}$	90.0$^{\uparrow}$	19995	60.4$^{\downarrow}$	83.3$^{\uparrow}$	21781	66.2$^{\downarrow}$	86.7$^{\downarrow}$	20468	42.5$^{\downarrow}$	70.0	23707	59.1$^{\downarrow}$	82.5$^{\uparrow}$	21488
		1024	66.2$^{\downarrow}$	80.0$^{\downarrow}$	20540	62.1$^{\downarrow}$	76.7$^{\downarrow}$	21536	62.9$^{\downarrow}$	86.7$^{\downarrow}$	20011	42.9$^{\downarrow}$	56.7$^{\downarrow}$	24225	58.5$^{\downarrow}$	75.0$^{\downarrow}$	21578
		65536	69.2$^{\downarrow}$	83.3$^{\downarrow}$	19260	62.1$^{\downarrow}$	83.3$^{\uparrow}$	21575	64.6$^{\downarrow}$	86.7$^{\downarrow}$	19867	44.2$^{\downarrow}$	70.0	23562	60.0$^{\downarrow}$	80.8$^{\downarrow}$	21066
OLMo-3-7B-Think	Final-256	42	70.0$^{\downarrow}$	83.3$^{\downarrow}$	19567	64.6$^{\downarrow}$	83.3$^{\uparrow}$	21531	67.1$^{\downarrow}$	86.7$^{\downarrow}$	20365	43.3$^{\downarrow}$	63.3$^{\downarrow}$	23137	61.2$^{\downarrow}$	79.2$^{\downarrow}$	21150
		1024	64.6$^{\downarrow}$	86.7	20191	61.2$^{\downarrow}$	86.7$^{\uparrow}$	22046	65.8$^{\downarrow}$	86.7$^{\downarrow}$	20157	37.9$^{\downarrow}$	53.3$^{\downarrow}$	24494	57.4$^{\downarrow}$	78.3$^{\downarrow}$	21722
		65536	67.1$^{\downarrow}$	83.3$^{\downarrow}$	20310	57.9$^{\downarrow}$	80.0	21842	62.5$^{\downarrow}$	86.7$^{\downarrow}$	20943	41.7$^{\downarrow}$	60.0$^{\downarrow}$	23971	57.3$^{\downarrow}$	77.5$^{\downarrow}$	21766

G.2.5 SEED SENSITIVITY: LOSS-TOKEN POSITION (EVALUATION SEED)

Table 69: Per-dataset evaluation-seed results for Table 44. First-256 reuses OPSD-256. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean. $^{\uparrow}$/$^{\downarrow}$ mark accuracy changes relative to Base at the matched evaluation seed. Length is unannotated.

Model	Condition	Eval seed	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
			Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	First-256 / OPSD-256	42	46.2$^{\uparrow}$	70.0$^{\downarrow}$	20783	36.7$^{\uparrow}$	66.7$^{\uparrow}$	21697	45.0$^{\uparrow}$	73.3$^{\uparrow}$	20995	24.6$^{\uparrow}$	40.0$^{\downarrow}$	22306	38.1$^{\uparrow}$	62.5$^{\uparrow}$	21445
		1024	47.9$^{\uparrow}$	70.0$^{\downarrow}$	20152	36.7	60.0$^{\uparrow}$	21390	37.9$^{\uparrow}$	56.7	20678	25.4$^{\uparrow}$	50.0$^{\uparrow}$	21863	37.0$^{\uparrow}$	59.2$^{\uparrow}$	21021
		65536	47.5$^{\uparrow}$	73.3	20297	34.2$^{\downarrow}$	63.3$^{\uparrow}$	21394	41.7$^{\uparrow}$	66.7	21027	23.3$^{\downarrow}$	46.7$^{\downarrow}$	22365	36.7$^{\uparrow}$	62.5$^{\uparrow}$	21271
Qwen3-1.7B	Uni-256	42	48.8$^{\uparrow}$	76.7	21874	31.7$^{\downarrow}$	60.0	21873	36.7$^{\uparrow}$	63.3$^{\uparrow}$	22032	23.8$^{\uparrow}$	46.7	24462	35.2$^{\uparrow}$	61.7$^{\uparrow}$	22560
		1024	47.1$^{\downarrow}$	73.3$^{\downarrow}$	22073	34.6$^{\downarrow}$	63.3$^{\uparrow}$	21959	36.7$^{\uparrow}$	60.0$^{\uparrow}$	21197	23.8$^{\uparrow}$	40.0$^{\uparrow}$	24054	35.5$^{\uparrow}$	59.2$^{\uparrow}$	22321
		65536	45.0$^{\uparrow}$	73.3	22731	33.8$^{\downarrow}$	60.0$^{\uparrow}$	22116	36.2$^{\uparrow}$	60.0$^{\downarrow}$	21268	22.1$^{\downarrow}$	40.0$^{\downarrow}$	25005	34.3$^{\downarrow}$	58.3$^{\downarrow}$	22780
Qwen3-1.7B	Final-256	42	42.5$^{\downarrow}$	73.3$^{\downarrow}$	23510	34.6$^{\downarrow}$	53.3$^{\downarrow}$	22599	37.1$^{\uparrow}$	63.3$^{\uparrow}$	22987	22.9$^{\uparrow}$	46.7	25800	34.3$^{\uparrow}$	59.2	23724
		1024	47.1$^{\downarrow}$	73.3$^{\downarrow}$	22673	35.8$^{\downarrow}$	70.0$^{\uparrow}$	22386	32.9$^{\downarrow}$	53.3$^{\downarrow}$	22467	24.2$^{\uparrow}$	50.0$^{\uparrow}$	26276	35.0$^{\downarrow}$	61.7$^{\uparrow}$	23450
		65536	43.3$^{\uparrow}$	70.0$^{\downarrow}$	22396	34.2$^{\downarrow}$	53.3$^{\downarrow}$	22415	36.7$^{\uparrow}$	66.7	22742	21.2$^{\downarrow}$	43.3$^{\downarrow}$	25626	33.9$^{\downarrow}$	58.3$^{\downarrow}$	23295
OLMo-3-7B-Think	First-256 / OPSD-256	42	67.9$^{\downarrow}$	86.7	20282	62.5$^{\downarrow}$	80.0	21406	67.5$^{\downarrow}$	83.3$^{\downarrow}$	20212	45.4$^{\uparrow}$	70.0	23691	60.8$^{\downarrow}$	80.0$^{\downarrow}$	21398
		1024	70.0$^{\uparrow}$	90.0$^{\uparrow}$	20461	65.4$^{\uparrow}$	83.3$^{\downarrow}$	21875	67.9$^{\downarrow}$	86.7	20858	44.6	63.3$^{\downarrow}$	23498	62.0$^{\downarrow}$	80.8	21673
		65536	69.6$^{\downarrow}$	86.7	20365	65.0$^{\downarrow}$	80.0$^{\uparrow}$	22070	67.5$^{\downarrow}$	86.7	20694	43.3$^{\downarrow}$	56.7$^{\downarrow}$	23373	61.4$^{\downarrow}$	77.5$^{\downarrow}$	21626
OLMo-3-7B-Think	Uni-256	42	67.1$^{\downarrow}$	90.0$^{\uparrow}$	19995	60.4$^{\downarrow}$	83.3$^{\uparrow}$	21781	66.2$^{\downarrow}$	86.7$^{\downarrow}$	20468	42.5$^{\downarrow}$	70.0	23707	59.1$^{\downarrow}$	82.5$^{\uparrow}$	21488
		1024	70.4$^{\uparrow}$	86.7$^{\uparrow}$	19379	62.5$^{\downarrow}$	80.0$^{\downarrow}$	20906	64.6$^{\downarrow}$	86.7	20448	41.2$^{\downarrow}$	60.0$^{\downarrow}$	23909	59.7$^{\downarrow}$	78.3$^{\downarrow}$	21160
		65536	66.2$^{\downarrow}$	90.0$^{\uparrow}$	19865	58.3$^{\downarrow}$	83.3$^{\uparrow}$	21861	66.7$^{\downarrow}$	86.7	20597	42.9$^{\downarrow}$	56.7$^{\downarrow}$	23683	58.5$^{\downarrow}$	79.2	21502
OLMo-3-7B-Think	Final-256	42	70.0$^{\downarrow}$	83.3$^{\downarrow}$	19567	64.6$^{\downarrow}$	83.3$^{\uparrow}$	21531	67.1$^{\downarrow}$	86.7$^{\downarrow}$	20365	43.3$^{\downarrow}$	63.3$^{\downarrow}$	23137	61.2$^{\downarrow}$	79.2$^{\downarrow}$	21150
		1024	66.7$^{\downarrow}$	83.3	19477	62.9$^{\downarrow}$	86.7	21028	65.0$^{\downarrow}$	83.3$^{\downarrow}$	20347	42.5$^{\downarrow}$	63.3$^{\downarrow}$	23681	59.3$^{\downarrow}$	79.2$^{\downarrow}$	21133
		65536	69.6$^{\downarrow}$	80.0$^{\downarrow}$	19135	60.8$^{\downarrow}$	86.7$^{\uparrow}$	21740	62.9$^{\downarrow}$	86.7	20802	39.6$^{\downarrow}$	66.7	23577	58.2$^{\downarrow}$	80.0$^{\uparrow}$	21313

G.2.6 SEED SENSITIVITY: TEACHER-SIDE CONTEXT (TRAINING SEED)

Table 70: Per-dataset training-seed results for Table 45. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean. $^{\uparrow}$/$^{\downarrow}$ mark accuracy changes relative to Base@eval-42. Length is unannotated.

Model	Condition	Train seed	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
			Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	Solution	42	45.8$^{\uparrow}$	73.3$^{\downarrow}$	22175	36.7$^{\uparrow}$	66.7$^{\uparrow}$	22143	36.7$^{\uparrow}$	63.3$^{\uparrow}$	21288	22.1$^{\uparrow}$	40.0$^{\downarrow}$	23913	35.3$^{\uparrow}$	60.8$^{\uparrow}$	22380
		1024	49.2$^{\uparrow}$	73.3$^{\downarrow}$	21091	36.7$^{\uparrow}$	63.3$^{\uparrow}$	21896	37.5$^{\uparrow}$	60.0$^{\uparrow}$	21149	20.4$^{\downarrow}$	30.0$^{\downarrow}$	23892	35.9$^{\uparrow}$	56.7$^{\downarrow}$	22007
		65536	41.7$^{\downarrow}$	76.7	22332	35.8$^{\uparrow}$	53.3$^{\downarrow}$	21324	37.1$^{\uparrow}$	70.0$^{\uparrow}$	21537	23.8$^{\uparrow}$	50.0$^{\uparrow}$	24818	34.6$^{\uparrow}$	62.5$^{\uparrow}$	22503
Qwen3-1.7B	Answer	42	47.1$^{\uparrow}$	76.7	25291	34.2$^{\downarrow}$	56.7$^{\downarrow}$	25710	39.6$^{\uparrow}$	70.0$^{\uparrow}$	24298	21.7	33.3$^{\downarrow}$	27985	35.6$^{\uparrow}$	59.2	25821
		1024	44.2$^{\uparrow}$	73.3$^{\downarrow}$	25539	37.1$^{\uparrow}$	63.3$^{\uparrow}$	25758	40.8$^{\uparrow}$	70.0$^{\uparrow}$	23793	24.6$^{\uparrow}$	50.0$^{\uparrow}$	27365	36.7$^{\uparrow}$	64.2$^{\uparrow}$	25614
		65536	41.7$^{\downarrow}$	70.0$^{\downarrow}$	26299	32.9$^{\downarrow}$	60.0	25511	39.6$^{\uparrow}$	66.7$^{\uparrow}$	23025	24.6$^{\uparrow}$	40.0$^{\downarrow}$	27336	34.7$^{\uparrow}$	59.2	25543
Qwen3-1.7B	Unrelated	42	44.2$^{\uparrow}$	70.0$^{\downarrow}$	21601	36.7$^{\uparrow}$	66.7$^{\uparrow}$	22245	42.5$^{\uparrow}$	60.0$^{\uparrow}$	21496	25.0$^{\uparrow}$	43.3$^{\downarrow}$	22792	37.1$^{\uparrow}$	60.0$^{\uparrow}$	22034
		1024	48.8$^{\uparrow}$	76.7	21931	39.2$^{\uparrow}$	66.7$^{\uparrow}$	22286	44.2$^{\uparrow}$	70.0$^{\uparrow}$	20750	23.8$^{\uparrow}$	36.7$^{\downarrow}$	22481	39.0$^{\uparrow}$	62.5$^{\uparrow}$	21862
		65536	49.2$^{\uparrow}$	73.3$^{\downarrow}$	21084	33.8$^{\downarrow}$	63.3$^{\uparrow}$	22712	41.7$^{\uparrow}$	73.3$^{\uparrow}$	20774	24.6$^{\uparrow}$	40.0$^{\downarrow}$	21909	37.3$^{\uparrow}$	62.5$^{\uparrow}$	21620
OLMo-3-7B-Think	Solution	42	68.8$^{\downarrow}$	76.7$^{\downarrow}$	20502	55.4$^{\downarrow}$	80.0	22027	67.1$^{\downarrow}$	86.7$^{\downarrow}$	20285	43.8$^{\downarrow}$	66.7$^{\downarrow}$	24083	58.8$^{\downarrow}$	77.5$^{\downarrow}$	21724
		1024	67.9$^{\downarrow}$	80.0$^{\downarrow}$	19871	62.5$^{\downarrow}$	80.0	21995	64.6$^{\downarrow}$	90.0	21048	41.7$^{\downarrow}$	63.3$^{\downarrow}$	24189	59.2$^{\downarrow}$	78.3$^{\downarrow}$	21776
		65536	69.2$^{\downarrow}$	86.7	19846	61.2$^{\downarrow}$	83.3$^{\uparrow}$	21643	67.1$^{\downarrow}$	90.0	20151	41.2$^{\downarrow}$	66.7$^{\downarrow}$	24033	59.7$^{\downarrow}$	81.7	21418
OLMo-3-7B-Think	Answer	42	64.6$^{\downarrow}$	80.0$^{\downarrow}$	22921	59.2$^{\downarrow}$	76.7$^{\downarrow}$	24151	61.2$^{\downarrow}$	86.7$^{\downarrow}$	23492	39.6$^{\downarrow}$	60.0$^{\downarrow}$	25903	56.1$^{\downarrow}$	75.8$^{\downarrow}$	24117
		1024	62.1$^{\downarrow}$	80.0$^{\downarrow}$	23093	63.3$^{\downarrow}$	76.7$^{\downarrow}$	24413	63.8$^{\downarrow}$	83.3$^{\downarrow}$	23875	38.3$^{\downarrow}$	56.7$^{\downarrow}$	26355	56.9$^{\downarrow}$	74.2$^{\downarrow}$	24434
		65536	60.8$^{\downarrow}$	80.0$^{\downarrow}$	23399	61.7$^{\downarrow}$	80.0	23703	63.8$^{\downarrow}$	80.0$^{\downarrow}$	22945	41.7$^{\downarrow}$	56.7$^{\downarrow}$	25495	57.0$^{\downarrow}$	74.2$^{\downarrow}$	23886
OLMo-3-7B-Think	Unrelated	42	62.5$^{\downarrow}$	80.0$^{\downarrow}$	22662	58.3$^{\downarrow}$	76.7$^{\downarrow}$	23626	64.6$^{\downarrow}$	83.3$^{\downarrow}$	22398	44.6	66.7$^{\downarrow}$	24912	57.5$^{\downarrow}$	76.7$^{\downarrow}$	23400
		1024	62.5$^{\downarrow}$	83.3$^{\downarrow}$	22345	58.8$^{\downarrow}$	76.7$^{\downarrow}$	23273	64.6$^{\downarrow}$	86.7$^{\downarrow}$	22615	40.0$^{\downarrow}$	53.3$^{\downarrow}$	25376	56.5$^{\downarrow}$	75.0$^{\downarrow}$	23402
		65536	62.9$^{\downarrow}$	83.3$^{\downarrow}$	22696	61.2$^{\downarrow}$	83.3$^{\uparrow}$	23954	62.9$^{\downarrow}$	83.3$^{\downarrow}$	22577	38.8$^{\downarrow}$	56.7$^{\downarrow}$	25881	56.5$^{\downarrow}$	76.7$^{\downarrow}$	23777

G.2.7 SEED SENSITIVITY: TEACHER-SIDE CONTEXT (EVALUATION SEED)

Table 71: Per-dataset evaluation-seed results for Table 46. Each dataset reports Avg@8, Pass@8, and mean output length. Dataset Avg. is the unweighted four-benchmark mean. $^{\uparrow}$/$^{\downarrow}$ mark accuracy changes relative to Base at the matched evaluation seed. Length is unannotated.

Model	Condition	Eval seed	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
			Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	Solution	42	45.8$^{\uparrow}$	73.3$^{\downarrow}$	22175	36.7$^{\uparrow}$	66.7$^{\uparrow}$	22143	36.7$^{\uparrow}$	63.3$^{\uparrow}$	21288	22.1$^{\uparrow}$	40.0$^{\downarrow}$	23913	35.3$^{\uparrow}$	60.8$^{\uparrow}$	22380
		1024	49.2$^{\uparrow}$	76.7	21279	34.2$^{\downarrow}$	63.3$^{\uparrow}$	21384	38.8$^{\uparrow}$	66.7$^{\uparrow}$	21373	25.0$^{\uparrow}$	53.3$^{\uparrow}$	23458	36.8$^{\uparrow}$	65.0$^{\uparrow}$	21874
		65536	49.6$^{\uparrow}$	73.3	21563	35.0$^{\downarrow}$	63.3$^{\uparrow}$	21611	38.8$^{\uparrow}$	66.7	21850	22.9$^{\downarrow}$	43.3$^{\downarrow}$	23902	36.6$^{\uparrow}$	61.7	22231
Qwen3-1.7B	Answer	42	47.1$^{\uparrow}$	76.7	25291	34.2$^{\downarrow}$	56.7$^{\downarrow}$	25710	39.6$^{\uparrow}$	70.0$^{\uparrow}$	24298	21.7	33.3$^{\downarrow}$	27985	35.6$^{\uparrow}$	59.2	25821
		1024	46.7$^{\downarrow}$	70.0$^{\downarrow}$	25361	39.2$^{\uparrow}$	66.7$^{\uparrow}$	25576	37.1$^{\uparrow}$	53.3$^{\downarrow}$	24341	20.4	33.3	27802	35.8$^{\uparrow}$	55.8$^{\uparrow}$	25770
		65536	44.2$^{\uparrow}$	66.7$^{\downarrow}$	25516	36.2$^{\downarrow}$	60.0$^{\uparrow}$	26059	38.3$^{\uparrow}$	63.3$^{\downarrow}$	23938	25.0$^{\downarrow}$	50.0	28063	35.9$^{\uparrow}$	60.0$^{\downarrow}$	25894
Qwen3-1.7B	Unrelated	42	44.2$^{\uparrow}$	70.0$^{\downarrow}$	21601	36.7$^{\uparrow}$	66.7$^{\uparrow}$	22245	42.5$^{\uparrow}$	60.0$^{\uparrow}$	21496	25.0$^{\uparrow}$	43.3$^{\downarrow}$	22792	37.1$^{\uparrow}$	60.0$^{\uparrow}$	22034
		1024	46.7$^{\downarrow}$	73.3$^{\downarrow}$	21415	30.8$^{\downarrow}$	63.3$^{\uparrow}$	23628	40.4$^{\uparrow}$	63.3$^{\uparrow}$	21183	23.3$^{\uparrow}$	46.7$^{\uparrow}$	22724	35.3$^{\uparrow}$	61.7$^{\uparrow}$	22237
		65536	46.7$^{\uparrow}$	73.3	21583	32.9$^{\downarrow}$	53.3$^{\downarrow}$	22608	40.4$^{\uparrow}$	63.3$^{\downarrow}$	21519	26.7$^{\downarrow}$	46.7$^{\downarrow}$	22116	36.7$^{\uparrow}$	59.2$^{\downarrow}$	21956
OLMo-3-7B-Think	Solution	42	68.8$^{\downarrow}$	76.7$^{\downarrow}$	20502	55.4$^{\downarrow}$	80.0	22027	67.1$^{\downarrow}$	86.7$^{\downarrow}$	20285	43.8$^{\downarrow}$	66.7$^{\downarrow}$	24083	58.8$^{\downarrow}$	77.5$^{\downarrow}$	21724
		1024	68.8$^{\uparrow}$	86.7$^{\uparrow}$	19964	62.1$^{\downarrow}$	83.3$^{\downarrow}$	21831	62.5$^{\downarrow}$	83.3$^{\downarrow}$	20286	42.1$^{\downarrow}$	60.0$^{\downarrow}$	24108	58.9$^{\downarrow}$	78.3$^{\downarrow}$	21547
		65536	68.3$^{\downarrow}$	83.3$^{\downarrow}$	19832	62.5$^{\downarrow}$	83.3$^{\uparrow}$	21611	68.8$^{\downarrow}$	86.7	20612	41.7$^{\downarrow}$	63.3$^{\downarrow}$	24196	60.3$^{\downarrow}$	79.2	21563
OLMo-3-7B-Think	Answer	42	64.6$^{\downarrow}$	80.0$^{\downarrow}$	22921	59.2$^{\downarrow}$	76.7$^{\downarrow}$	24151	61.2$^{\downarrow}$	86.7$^{\downarrow}$	23492	39.6$^{\downarrow}$	60.0$^{\downarrow}$	25903	56.1$^{\downarrow}$	75.8$^{\downarrow}$	24117
		1024	64.6$^{\downarrow}$	86.7$^{\uparrow}$	22818	64.6$^{\uparrow}$	83.3$^{\downarrow}$	23461	63.8$^{\downarrow}$	83.3$^{\downarrow}$	23104	42.5$^{\downarrow}$	63.3$^{\downarrow}$	26202	58.9$^{\downarrow}$	79.2$^{\downarrow}$	23896
		65536	64.6$^{\downarrow}$	90.0$^{\uparrow}$	23039	61.2$^{\downarrow}$	80.0$^{\uparrow}$	24471	63.8$^{\downarrow}$	83.3$^{\downarrow}$	23203	42.9$^{\downarrow}$	66.7	25817	58.1$^{\downarrow}$	80.0$^{\uparrow}$	24132
OLMo-3-7B-Think	Unrelated	42	62.5$^{\downarrow}$	80.0$^{\downarrow}$	22662	58.3$^{\downarrow}$	76.7$^{\downarrow}$	23626	64.6$^{\downarrow}$	83.3$^{\downarrow}$	22398	44.6	66.7$^{\downarrow}$	24912	57.5$^{\downarrow}$	76.7$^{\downarrow}$	23400
		1024	66.7$^{\downarrow}$	86.7$^{\uparrow}$	21998	61.7$^{\downarrow}$	83.3$^{\downarrow}$	23733	62.5$^{\downarrow}$	83.3$^{\downarrow}$	22389	41.2$^{\downarrow}$	66.7	25369	58.0$^{\downarrow}$	80.0$^{\downarrow}$	23372
		65536	63.3$^{\downarrow}$	76.7$^{\downarrow}$	22340	62.9$^{\downarrow}$	80.0$^{\uparrow}$	23792	66.2$^{\downarrow}$	83.3$^{\downarrow}$	22036	42.1$^{\downarrow}$	60.0$^{\downarrow}$	24961	58.6$^{\downarrow}$	75.0$^{\downarrow}$	23282

G.3 ADDITIONAL-EXPERIMENT BREAKDOWNS

G.3.1 TEACHER-PREFIX LENGTH STRATIFICATION

Table 72: Per-dataset teacher-prefix-length results for Table 15. Full uses the complete training set and a 256-token rollout. Short-prefix and Long-prefix are defined by gold-trace length and contain different problems. Each dataset reports Avg@8, Pass@8, and mean output length.

Model	Subset	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	Full	46.3	70.0	20783	36.7	66.7	21697	45.0	73.3	20995	24.6	40.0	22306	38.1	62.5	21445
	Short-prefix	49.2	80.0	20157	38.8	66.7	21312	36.3	60.0	20889	24.2	43.3	21264	37.1	62.5	20905
	Long-prefix	48.8	73.3	19618	35.4	63.3	21613	37.5	66.7	20203	25.0	46.7	21638	36.7	62.5	20768
Qwen3-4B	Base	74.6	90.0	15495	65.4	80.0	18696	65.8	80.0	17249	43.3	66.7	19636	62.3	79.2	17769
	Full	72.9	83.3	16485	63.8	83.3	19818	63.3	83.3	18075	43.8	53.3	21951	60.9	75.8	19082
	Short-prefix	69.6	80.0	16890	69.2	86.7	20068	65.0	83.3	17880	40.4	66.7	21160	61.0	79.2	18999
	Long-prefix	73.8	83.3	16666	63.3	80.0	20078	65.4	83.3	18218	40.0	50.0	21381	60.6	74.2	19086
Qwen3-4B-Instruct	Base	62.5	86.7	9412	45.8	70.0	8558	55.0	80.0	8434	30.8	50.0	10254	48.5	71.7	9165
	Full	63.8	86.7	10378	47.5	73.3	10414	55.8	73.3	10143	31.2	50.0	12315	49.6	70.8	10813
	Short-prefix	62.1	80.0	10540	47.1	70.0	10069	58.3	83.3	10383	27.9	40.0	13198	48.9	68.3	11048
	Long-prefix	61.2	80.0	10775	49.6	70.0	10335	53.8	76.7	9665	30.8	43.3	11672	48.9	67.5	10612
OLMo-3-7B-Instruct	Base	48.8	76.7	7459	39.6	60.0	7182	42.5	66.7	7277	27.1	40.0	8528	39.5	60.8	7612
	Full	52.1	83.3	8329	41.2	73.3	7928	50.4	70.0	7865	27.9	46.7	9323	42.9	68.3	8361
	Short-prefix	53.8	83.3	7950	45.4	70.0	7884	49.2	66.7	7440	29.2	43.3	9411	44.4	65.8	8171
	Long-prefix	51.2	80.0	8137	40.8	56.7	7320	50.0	76.7	7975	24.6	43.3	9499	41.7	64.2	8232

G.3.2 TEMPERATURE ROBUSTNESS CHECKS

Table 73: Per-dataset decoding-temperature results for Table 47. Base and OPSD use the listed temperature.

Model	Decode τ / ckpt	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
Qwen3-1.7B	0.6 / Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	0.6 / OPSD	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
	1.0 / Base	49.2	73.3	17458	39.2	63.3	17941	40.4	66.7	17523	23.8	46.7	19009	38.1	62.5	17983
	1.0 / OPSD	51.2	80.0	20477	35.4	73.3	21752	34.6	60.0	20409	25.4	50.0	22901	36.7	65.8	21385
Qwen3-4B-Instruct	0.7 / Base	62.5	86.7	9412	45.8	70.0	8558	55.0	80.0	8434	30.8	50.0	10254	48.5	71.7	9165
	0.7 / OPSD	65.0	83.3	10141	45.8	73.3	10539	58.3	76.7	10130	30.8	46.7	11409	50.0	70.0	10555
	1.0 / Base	65.0	90.0	9529	48.3	73.3	8174	54.6	80.0	8459	28.8	40.0	10532	49.2	70.8	9174
	1.0 / OPSD	62.5	76.7	10912	46.2	70.0	10588	55.8	73.3	9258	30.4	46.7	12089	48.8	66.7	10712
Qwen3-4B-Thinking-2507	0.6 / Base	82.1	93.3	20319	81.7	86.7	22385	81.2	86.7	21835	56.7	80.0	26570	75.4	86.7	22777
	0.6 / OPSD	73.3	93.3	40191	70.4	83.3	44970	73.3	90.0	41319	46.2	66.7	46992	65.8	83.3	43368
	1.0 / Base	83.3	93.3	20402	79.2	90.0	22682	83.3	90.0	22180	58.3	73.3	26979	76.0	86.7	23061
	1.0 / OPSD	74.6	90.0	40965	72.9	86.7	44799	72.9	90.0	42020	46.7	63.3	47080	66.8	82.5	43716
OLMo-3-7B-Instruct	0.6 / Base	48.8	76.7	7459	39.6	60.0	7182	42.5	66.7	7277	27.1	40.0	8528	39.5	60.8	7612
	0.6 / OPSD	53.8	76.7	7201	39.6	63.3	6818	46.2	66.7	6983	23.8	43.3	8698	40.8	62.5	7425
	1.0 / Base	52.9	80.0	7455	41.7	63.3	7198	48.3	76.7	7332	24.6	53.3	8518	41.9	68.3	7626
	1.0 / OPSD	46.7	80.0	7175	41.7	63.3	6815	42.5	66.7	6921	21.2	36.7	8088	38.0	61.7	7250

Table 74: Per-dataset Qwen3-1.7B training-temperature results for Table 48. All rows use decoding temperature 0.6.

τ_S	τ_T	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length	Avg	Pass	Length
1.1	1.1	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
0.6	0.6	46.7	76.7	19889	37.1	63.3	19212	35.4	63.3	19985	22.5	40.0	21010	35.4	60.8	20024
0.6	1.1	46.2	76.7	21424	34.2	53.3	22346	38.3	56.7	21255	22.1	40.0	24204	35.2	56.7	22307
1.1	0.6	43.8	76.7	17488	34.2	60.0	18115	30.4	53.3	18741	20.8	40.0	19019	32.3	57.5	18341

G.3.3 REPLICATIONS OF RECENT OPSD VARIANTS

These replications use the recent papers' LoRA(Hu et al., 2021)-based 200-step configurations. Each benchmark cell reports Avg@8 / Pass@8. The final columns give the unweighted mean and mean output length.

Table 75: Per-dataset Purified OPSD results for Table 38. Priv. is the privileged teacher field used to construct the PMI target.

Model	Priv.	AIME24	AIME25	AIME26	HMMT25	Macro Avg	Macro Pass	Length
Qwen3-1.7B	Answer	52.9 / 76.7	41.7 / 63.3	41.7 / 70.0	23.3 / 43.3	39.9	63.3	24.2k
Qwen3-1.7B	Solution	46.3 / 76.7	37.9 / 60.0	36.7 / 63.3	29.2 / 60.0	37.5	65.0	20.8k
Qwen3-4B	Answer	75.0 / 83.3	70.8 / 80.0	66.7 / 86.7	40.8 / 66.7	63.3	79.2	19.6k
Qwen3-4B	Solution	72.5 / 80.0	62.1 / 80.0	64.6 / 83.3	46.3 / 66.7	61.3	77.5	18.5k
Qwen3-4B-Thinking-2507	Answer	83.8 / 90.0	82.5 / 86.7	78.3 / 86.7	55.8 / 73.3	75.1	84.2	24.8k
Qwen3-4B-Thinking-2507	Solution	84.2 / 93.3	79.6 / 86.7	81.3 / 90.0	56.3 / 76.7	75.3	86.7	24.2k
OLMo-3-7B-Think	Answer	73.8 / 86.7	69.2 / 80.0	70.4 / 90.0	44.6 / 66.7	64.5	80.8	19.2k
OLMo-3-7B-Think	Solution	73.3 / 86.7	66.3 / 86.7	71.3 / 86.7	44.2 / 60.0	63.8	80.0	18.9k

Table 76: Per-dataset β-OPSD results for Table 38. Each benchmark cell reports Avg@8 / Pass@8.

Model	AIME24	AIME25	AIME26	HMMT25	Macro Avg	Macro Pass	Length
Qwen3-1.7B	32.5 / 63.3	27.9 / 56.7	25.4 / 50.0	14.6 / 26.7	25.1	49.2	13.4k
Qwen3-4B-Thinking-2507	76.7 / 86.7	74.6 / 86.7	77.1 / 90.0	46.3 / 66.7	68.6	82.5	18.2k
Qwen3-4B-Instruct	59.2 / 76.7	43.3 / 60.0	57.5 / 83.3	28.8 / 43.3	47.2	65.8	11.0k
OLMo-3-7B-Think	69.2 / 80.0	69.2 / 86.7	72.1 / 90.0	45.8 / 70.0	64.1	81.7	18.8k
OLMo-3-7B-Instruct	55.0 / 83.3	45.0 / 60.0	45.0 / 66.7	24.6 / 50.0	42.4	65.0	8.3k

G.3.4 Additional Training Controls

Table 77: Per-dataset English instruction-prefix controls for Table 32. C/D denote concise/detailed student-to-teacher instructions. Both instruction-only rows contain no privileged teacher context.

Model	Condition	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len
Qwen3-1.7B	Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	Full solution	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
	No priv.: C→D	30.0	53.3	12632	25.0	43.3	12005	22.5	53.3	12903	15.0	36.7	12791	23.1	46.7	12583
	No priv.: D→C	43.3	73.3	25307	37.5	66.7	25255	42.1	70.0	23577	22.1	40.0	25965	36.3	62.5	25026
OLMo-3-7B-Think	Base	71.7	86.7	17296	66.7	80.0	19137	71.2	90.0	17560	44.6	70.0	20950	63.5	81.7	18736
	Full solution	68.8	76.7	20502	55.4	80.0	22027	67.1	86.7	20285	43.8	66.7	24083	58.8	77.5	21724
	No priv.: C→D	72.9	86.7	14621	63.3	83.3	17110	65.0	90.0	15766	41.7	63.3	18436	60.7	80.8	16483
	No priv.: D→C	62.1	80.0	23097	62.5	80.0	23728	61.7	86.7	23608	39.2	53.3	25808	56.4	75.0	24060

Table 78: Per-dataset AdvT4 / AdvT16 results for Table 20.

Model	Stage	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len
Qwen3-1.7B	Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	c256	46.2	70.0	20783	36.7	66.7	21697	45.0	73.3	20995	24.6	40.0	22306	38.1	62.5	21445
	AdvT4	47.1	80.0	19206	34.6	50.0	20145	35.8	56.7	18732	21.2	50.0	20992	34.7	59.2	19769
	AdvT16	51.2	80.0	18716	37.1	60.0	19128	36.2	60.0	19273	24.2	46.7	20824	37.2	61.7	19485
Qwen3-4B	Base	74.6	90.0	15495	65.4	80.0	18696	65.8	80.0	17249	43.3	66.7	19636	62.3	79.2	17769
	c256	72.9	83.3	16485	63.8	83.3	19818	63.3	83.3	18075	43.8	53.3	21951	60.9	75.8	19082
	AdvT4	71.2	86.7	15568	65.0	80.0	19779	64.6	86.7	17354	41.2	53.3	20862	60.5	76.7	18391
	AdvT16	75.4	86.7	14956	67.1	80.0	18989	64.6	86.7	17268	41.7	60.0	20689	62.2	78.3	17976
Qwen3-4B-Thinking-2507	Base	82.1	93.3	20319	81.7	86.7	22385	81.2	86.7	21835	56.7	80.0	26570	75.4	86.7	22777
	c256	80.0	90.0	28157	75.4	83.3	30535	75.8	90.0	29164	50.0	70.0	34249	70.3	83.3	30526
	AdvT4	83.8	93.3	23379	79.2	86.7	27465	80.8	93.3	25947	53.8	70.0	30788	74.4	85.8	26895
	AdvT16	81.7	86.7	23679	82.1	86.7	26947	78.8	90.0	25529	58.3	76.7	31045	75.2	85.0	26800
OLMo-3-7B-Think	Base	71.7	86.7	17296	66.7	80.0	19137	71.2	90.0	17560	44.6	70.0	20950	63.5	81.7	18736
	c256	67.9	86.7	20282	62.5	80.0	21406	67.5	83.3	20212	45.4	70.0	23691	60.8	80.0	21398
	AdvT4	67.5	83.3	19714	67.1	83.3	21114	67.5	90.0	20116	47.1	73.3	23766	62.3	82.5	21178
	AdvT16	71.7	93.3	19715	67.5	83.3	21636	65.8	83.3	20329	42.5	63.3	23956	61.9	80.8	21409

Table 79: Per-dataset pointwise vocabulary-contribution-cap results for Table 33.

Model	τ_{clip}	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
		Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len
Qwen3-1.7B	Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
	0.01	40.4	66.7	25792	30.8	56.7	25552	32.1	63.3	24715	22.1	40.0	28918	31.4	56.7	26244
	0.05	45.8	73.3	22175	36.7	66.7	22143	36.7	63.3	21288	22.1	40.0	23913	35.3	60.8	22380
	0.1	44.2	76.7	16599	32.1	53.3	16320	32.9	60.0	16266	22.5	50.0	17404	32.9	60.0	16647
	0.2	34.6	63.3	14852	32.5	53.3	15336	31.7	53.3	15600	23.8	43.3	16392	30.6	53.3	15545
OLMo-3-7B-Think	Base	71.7	86.7	17296	66.7	80.0	19137	71.2	90.0	17560	44.6	70.0	20950	63.5	81.7	18736
	0.01	67.9	80.0	19125	59.6	86.7	20633	63.8	86.7	19651	41.7	63.3	22835	58.2	79.2	20561
	0.05	68.8	76.7	20502	55.4	80.0	22027	67.1	86.7	20285	43.8	66.7	24083	58.8	77.5	21724
	0.1	71.2	86.7	17730	66.7	86.7	19442	69.2	90.0	17978	45.4	60.0	22416	63.1	80.8	19391
	0.2	72.1	90.0	15481	65.8	80.0	17428	70.8	93.3	15372	43.8	63.3	19545	63.1	81.7	16956

Table 80: Per-dataset 10-step SFT followed by 100-step no-privilege OPSD for Table 34. OT/OMR denote the SFT corpus. OPSD uses OpenThoughts.

Model	SFT data	Stage	AIME 2024			AIME 2025			AIME 2026			HMMT25			Dataset Avg.		
			Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len
Qwen3-1.7B	OT	Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
		SFT-10	7.1	23.3	9991	7.5	20.0	5382	5.8	26.7	6527	0.0	0.0	5897	5.1	17.5	6949
		+ OPSD-100	7.9	23.3	8030	5.8	20.0	6372	5.8	16.7	5708	1.3	6.7	6183	5.2	16.7	6573
Qwen3-1.7B	OMR	Base	43.8	76.7	18647	35.4	60.0	18266	35.0	53.3	18161	21.7	46.7	19421	34.0	59.2	18624
		SFT-10	35.4	66.7	17957	30.0	46.7	15927	29.6	63.3	17051	17.1	40.0	16562	28.0	54.2	16874
		+ OPSD-100	36.7	63.3	17613	29.2	53.3	16041	26.3	53.3	17416	16.3	26.7	16416	27.1	49.2	16872
Qwen3-4B-Thinking-2507	OT	Base	82.1	93.3	20319	81.7	86.7	22385	81.2	86.7	21835	56.7	80.0	26570	75.4	86.7	22777
		SFT-10	6.2	20.0	2270	6.2	30.0	1485	6.2	26.7	2841	0.8	6.7	1276	4.9	20.8	1968
		+ OPSD-100	9.6	36.7	4579	10.4	40.0	1731	7.5	23.3	3478	3.8	13.3	2414	7.8	28.3	3050
OLMo-3-7B-Think	OT	Base	71.7	86.7	17296	66.7	80.0	19137	71.2	90.0	17560	44.6	70.0	20950	63.5	81.7	18736
		SFT-10	71.7	90.0	16595	57.9	80.0	18892	70.0	90.0	17459	40.8	60.0	20552	60.1	80.0	18374
		+ OPSD-100	68.3	83.3	16684	63.8	86.7	18251	65.4	90.0	17863	43.8	66.7	20342	60.3	81.7	18285